AN INVESTIGATION OF DIMENSIONALITY ACROSS GRADE LEVELS AND EFFECTS ON VERTICAL LINKING FOR ELEMENTARY GRADE MATHEMATICS ACHIEVEMENT TESTS

Samantha S. Burg

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctorate of Education in the School of Education

Chapel Hill
2007

Approved by:

Advisor: Gregory J. Cizek

Reader: Donald Burdick

Reader: Carol Malloy

Reader: David Thissen

Reader: William Ware

ABSTRACT

SAMANTHA S. BURG: An Investigation of Dimensionality across Grade Levels and Effects on Vertical Linking for Elementary Grade Mathematics Achievement Tests (Under the direction of Gregory J. Cizek)

It is a widely held belief that mathematical content strands reflect different constructs which produce multidimensionality in mathematical achievement tests for Grade 3-8. This study analyzes the dimensional structure of mathematical achievement tests aligned to NCTM content strands using four different methods for assessing dimensionality. The effect of including off-grade linking items as a potential source of dimensionality was also considered. The results indicate that although mathematical achievement tests for Grades 3-8 are complex and exhibit some multidimensionality, the sources of dimensionality are not related to the content strands or the inclusion of several off-grade linking items. The complexity of the data structure along with the known overlap of mathematical skills suggest that mathematical achievement tests could represent a fundamentally unidimensional construct. Refining the definition of dimensionality to include "detectable dimensionality" is discussed.

DEDICATION

To the people who have walked this road before and with me, I express my many thanks and much appreciation. I am especially grateful to my family of educational enthusiasts who have been sources of unending support and love, Drs. William and Je Anne Burg, and Drs. Timothy and Karen Burg. I would also like to express my thanks to Jesus Christ, my Prince of Peace, my Counselor, my Wisdom who was there through the highs and through the lows and never let go of me.

ACKNOWLEDGMENTS

Completing a dissertation is about extending the work of others or perhaps better said as standing on the shoulders of giants. I would like to thank the staff at MetaMetrics, Inc. for all their support, knowledge, data and encouragement. In particular, I would like to thank my "giants" Ellie Sanford, Don Burdick, Jack Stenner, and Robin Baker. Your help and wisdom was extremely invaluable. I would also like to express my gratitude to another set of giants, my advisor and his wife, Greg and Rita Cizek, for their team efforts and support over the last several years.

Completing a dissertation is also part of a much larger community of people who walk the road with you such as my water aerobics class and in particular Earl Holman, Larry and Wanda Smith and the editorial assistance of Diane Johnson. I would also like to thank my fellow sojourners along the PhD path: Heather Koons, Sharyn Rosenberg, Pat Sylvester, Bev Faircloth, Bev Glienke, Sarah Hurwitz, and Elaine Rodeck.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Measurement is the process of assigning a number to represent the relationship between the item or characteristic under study and a unit of measurement; a scale weighs a person in pounds, a protractor measures an angle in degrees. Measuring instruments are the means by which this translation is made and all measuring instruments are subject to varying degrees of instrument error. In some ways, quantifying or measuring student achievement is not very different from measuring physical qualities—the researcher should use the most accurate, precise, appropriate instrument available to minimize instrument error. While no single instrument or assessment approach can perfectly measure student achievement, one of the most prevalent measures of achievement in grades K-12 is the standardized multiple choice test. To ensure that test scores are meaningful and provide accurate data and information, developing a high-quality, suitable measurement instrument is important.

In any specific area of K-12 instruction, developing valid tests that consistently and fairly assess the domain the test is intended to measure requires many steps and decisions throughout the entire test development, administration, and scoring process. In broad terms, these steps include: clearly defining the construct; preparing test specifications; conducting item development and analyzes; gathering validity evidence; and scaling and reporting test results. A unifying concept that underlies these central issues in test development is test dimensionality. Because it affects so much of the test development process--and thereby

affects the meaning of test scores-- further understanding of test dimensionality, sources of multidimensionality, assessment of dimensional structure, and consequences of violations of dimensionality assumptions is warranted.

What is Test Dimensionality?

In the context of measuring student achievement, test dimensionality is defined as the number of examinee characteristics or abilities measured by the items comprising an achievement test. The term *achievement* is used hereafter as an all-purpose expression for what the student knows and is able to do with respect to a specific domain. However, the terms *ability, latent ability, construct*, *dimension*, and *factor* are also used interchangeably to refer to the concept or characteristic that a test is designed to measure. A *construct* is a theoretical representation of the underlying trait, concept, attribute, process and/or structure that the test is designed to measure (Messick, 1989).

A test can be considered to measure one latent trait, construct or ability (in which case it is called *unidimensional*) or a combination of abilities (in which it is referred to as *multidimensional*). The dimensional structure of a test is intricately tied into the purpose and definition of the construct to be measured. Some tests are designed to be unidimensional while other tests are developed to measure several factors. However, it is sometimes the case that a test that is intended to be unidimensional may unintentionally be measuring more than one latent variable. Wainer and Thissen (1996) distinguished between two types of multidimensional tests: those with fixed multidimensionality and those with random multidimensionality. Tate (2002) similarly distinguished between planned and unintentional sources of dimensionality. Fixed or planned multidimensionality refers to the inclusion of several content areas or process levels in the test specifications and development. Random or

unintentional multidimensionality can be caused by many different sources which are described next.

Sources of Dimensionality and IRT Models

Many of the models used to analyze test data and develop test scores assume unidimensionality but this assumption cannot be strictly met because there are always other cognitive, personality, and test-taking factors that have an impact on test performance to some extent (Hambleton & Swaminathan, 1985). It is important to note that the dimensionality of a test also depends on the interaction of a set of items with a particular sample of examinees from its underlying population (Ackerman, 1994; Hattie, 1985; Reckase, 1990). Examinees differ in many ways, such as level of test anxiety, mathematics anxiety, motivation, out-of-class learning experiences that are relevant to in-school learning experiences, test taking skills and strategies, and other physical, cognitive, emotional, and personality characteristics that can influence test performance. These factors are in addition to and/or influence the dominant ability intended to be measured by a set of test items. Furthermore, it is possible for a test to be unidimensional within one population of examinees but multidimensional in another. Even if a test model is used that does not assume unidimensionality, the presence of multidimensionality can still be problematic and demand the attention of the test developers. According to Embretson and Reise (2000):

> Most commonly employed IRT models assume that a single latent-trait dimension underlies the probability of an item response...even in the application of multidimensional IRT models, the correct number of latent factors must be identified a priori, and hence, determining dimensionality is a critical issue in IRT modeling regardless of whether unidimensional or multidimensional models are being considered (p. 227).

However, while traditional IRT requires that the test be unidimensional, multidimensional item response theory (MIRT) allows the data to reflect more than one construct. Extensive research and the availability of greater computing power have opened up the availability and possibilities of MIRT. However, when MIRT models are used, reporting a single score is problematic. Score interpretation becomes increasing difficult when the items measure more than one construct or ability (Hattie, 1984). Determining which model is appropriate is an important decision and uncritical use of IRT models can result in serious statistical errors which then affect accuracy of individual examinee scores and inferences (Nandakumar, 1991). It is easy to conclude from this work that it is important for researchers to investigate dimensionality before applying IRT procedures for test development or scoring (Drasgow & Parsons, 1983; Stout, 2002).

An Example of Dimensionality in Mathematics Assessment

A specific example of the interaction of intended test characteristics and examinee characteristics can be seen in the measurement of mathematics achievement. Mathematics achievement tests have become more applied and contextual in part due to curriculum reforms fostered by the National Council of Teachers of Mathematics (NCTM). The NCTM *Principles and Standards for School Mathematics* (National Council of Teachers of Mathematics, 2000, hereafter NCTM *Standards*) provide educators with a vision of what it means to understand and know mathematics and outline the mathematics content and processes that students should be able to know and use as they progress through school. To encourage learning and conceptual understanding, an emphasis is placed on application and

context as evident in an NCTM statement of belief: "Learning mathematics is enhanced when content is placed in context and is connected to other subject areas and when students are given multiple opportunities to apply mathematics in meaningful ways as part of the learning process" (National Council of Teachers of Mathematics, 2006). However, from a psychometric perspective adding emphasis on problem solving, mathematics reasoning and mathematical communication could result in a multidimensional test compared to a test intended only to measure computational knowledge and skill (Walker & Beretvas, 2000). That is, other unintended factors such as reading profiency and/or cultural knowledge have been added to the measurement target. If there is variability in the examinee population in terms of reading ability, this would introduce a construct-irrelevant factor and could weaken validity evidence. It is also possible that test scores of different subpopulations may be differentially affected by different sources of other variation induced by the context itself (Crocker & Algina, 1986). The context may be difficult to translate across subpopulations such as those associated with different areas of the country, differences in social-economic status, various immigrate groups, students for whom English is a second language, and other cultural dissimilarities.

Dimensionality, Curriculum, and Large-Scale Assessment

The complicated nature of mathematics and the curriculum standards most states have adopted also contribute to other possible sources of dimensionality. In addition to application and process skills mentioned previously, the NCTM *Standards* highlight the growth of expectations in five content areas (called "strands"): Number Sense and Operations, Algebra, Geometry, Measurement, Data Analysis and Probability. It is not

expected that every topic would be addressed to the same extent instructionally each year; rather, students would develop a certain depth of understanding of concepts and acquire certain levels of fluency in a curriculum so that subsequent instruction can build on this understanding. For example, the curriculum for students in earlier elementary school would have a heavier focus on Number Sense and would introduce the simple ideas of Algebra. As the students progress through elementary school toward middle school, the curricular emphasis changes; instructional time spent on Number Sense and Operations would decrease while the focus on Algebra would increase.

The NCTM *Standards* provide guidelines for curriculums that many states have adopted or follow closely. While the instructional emphasis of the different mathematics strands changes over a typical mathematics curriculum, standardized tests report a single mathematics achievement or proficiency score at each grade. Because "achievement tests that are constructed with an emphasis on content specifications are likely not to be unidimensional" (Reckase, Davey, & Ackerman, 1989, p.2), further research is needed to explore the unintentional sources of multidimensionality that may arise due to mathematics test construction traditions that follow the NCTM *Standards* and explore whether test dimensionality changes with the grade appropriate curriculum.

Dimensionality, Linking, and Score Interpretation

In addition to test development, one of the most important activities of a testing program is the reporting and interpretation of test scores. Test scores are usually reported on scales designed to assist score interpretation. According to the *Standards for Educational and Psychological Testing* (American Educational Research Association, American

Psychological Association, & National Council on Measurement in Education), "scale scores are often created to enhance comparability across different forms of the same test, across different test formats or administration conditions or even across test designed to measure different constructs" (1999, p. 49).

There are many situations in which different examinees are measured with different instruments that are supposed to measure the same construct. For example, due to test security concerns, many testing programs develop *alternate* (sometimes called *parallel* or *equivalent*) forms for each grade and use the scores from these forms interchangeably. Alternate forms are constructed to the same content and test specifications but might differ somewhat in difficulty. The process of placing scores from alternate forms on a common scale and adjusting for possible differences in difficulty is done using various *equating* methods (Kolen & Brennan, 2004). Equating is a statistical process that is used to adjust scores on test forms so that scores on the forms are on the same scale. Equating done for the purpose of establishing comparability of scores from alternate forms is sometimes referred to as *horizontal* equating.

Other situations in which different examinees are measured with different, purposefully non-equivalent instruments involve the creation of vertical or developmental scales. Similar to equating, there are processes more properly referred to as *scaling to achieve comparability* (American Educational Research Association et al., 1999) or *linking* (Linn, 1993). *Vertical scaling* (frequently called *vertical "equating"*) is one of these processes and is often used to create developmental scores for achievement tests (Kolen & Brennan, 2004). For some applications such as value-added modeling and growth modeling

used for tracking student progress and achievement over time, a valid vertical scale is needed.

In the context of educational achievement testing, vertical scaling places scores from tests intended for different educational levels on the same scale; the tests differ in difficulty but are intended to measure the same construct. As part of the vertical scale development, tests for a given grade level are often constructed to include items that are below-grade level, above-grade level, or both, in addition to the appropriate on-grade-level items. These below- and above-grade items represent potential sources of unintentional dimensionality due to the out-of-grade content of the items or their likely differences in difficulty, readability, and so on.

According to Crocker and Algina, "an issue that must be considered in a vertical equating project is the possibility that tests differ substantially in difficulty also differ in the traits they measure despite having similar content" (1986, p. 474). In the case of vertically-scaled mathematics achievement tests designed to measure a mathematics curriculum that changes over grades to reflect the emphasis of the different strands, a test that also changed to reflect these content differences could introduce different traits being measured across the grades. Reckase (2004) has given an illustration of this problem:

> For math, tests at 3rd grade measure predominantly arithmetic skills. By 8th grade, the test shifts to problem solving, pre-algebra and algebra skills. Yet, the way the results are reported on the vertical scales seem to imply that the tests are measuring the same thing…more complicated is that within test at a grade, the difficult items may be measuring different combinations of skills than easy items … growth in student performance may take a circuitous path through many domains of test content (pp. 118-119).

In summary, whether developing and analyzing items, equating forms or establishing a vertical scale, the assessment the dimensional structure of tests is an important and ongoing activity.

Assessment of Dimensional Structure

There is no consensus in the measurement community on what constitutes best professional practice for assessing dimensionality, although a number of sound approaches exist. Historically, linear factor analytic methods have been used to investigate the internal structure of tests, although there are problems with this approach. Namely, the relationship between item performance and the underlying latent ability is often nonlinear (Hattie, 1984) and there is no standard criterion for determining the number of meaningful factors. The use of different decision rules such as Cattell's Scree test (plotting the eigenvalues), the Kaiser rule, and the minimum average partial (MAP) method are recommended when attempting to determine the number of dimensions to retain in an exploratory factor analysis (Preacher & MacCallum, 2003). However, although these approaches are widely used, it has been suggested that "researchers should now be starting to move away from reporting heuristic indices such as 'variance accounted for by the first factor' or 'ratio of the first to second eigenvalue' and start implementing the new procedures that tackle these issues" (Embretson & Reise, 2000, p. 245).

The dimensionality assessment and methods/programs currently available are a dramatic improvement over the often *ad hoc* methods that were common 15 years ago (Tate, 2002). These newer procedures include a family of item factor analytic procedures which are extensions of linear factor analysis modified to better model dichotomous item responses.

Another family of newer procedures test the assumption of unidimensionality by considering local independence and examining the conditional item associations. Both families have strengths and limitations. Further details of dimensionality assessment methods will be discussed in the next chapter.

Consequences of Violations of Dimensionality Assumptions

It is reasonable to ask the question, "If a test is intended to be unidimensional, but it is determined analytically to be measuring more than one dimension, what are the consequences of such a situation?" In practical test development for many large-scale achievement tests, unidimensionality is assumed by the chosen measurement models (often a Rasch model). Thus, the consequences of violations of unidimensionality assumption must be considered. Violations are typically associated with three areas: item analysis, validity, and linking.

The first area of concern is item analysis. Items are typically analyzed using item response theory (IRT); IRT models are widely used to develop and score K-12 achievement tests and many of the IRT models typically used in these contexts require the assumptions of unidimensionality and local independence. *Local independence* is related to unidimensionality. Local independence asserts that, after taking an examinee's ability into account, no relationship exists between the examinee's responses to different items on the test. However, "many educational and psychological tests are inherently multidimensional, meaning these test measure two or more constructs" (Ackerman, Gierl, & Walker, 2003, p. 198); that is, test item responses may not always be locally independent. Traub and Lam purport "the assumption of unidimensionality seems inappropriate for many kinds of test

data, especially those pertaining to tests of educational achievement" (1985, p. 22). There is also increasing recognition of the multidimensional nature of educational and psychological instruments (Embretson & Reise, 2000; Roussos, Stout, & Marden, 1998) as well as the findings that real test data often cannot be well modeled by locally independent unidimensional models (Ansley & Forsyth, 1985; Nandakumar, 1991; Reckase, 1979; Reckase, Carlson, & Ackerman, 1985).

Previous research has shown that using unidimensional models with multidimensional data can be problematic (Ansley & Forsyth, 1985; Drasgow & Parsons, 1983; Reckase, 1979). Walker and Beretvas (2000) have addressed concerns about dimensionality specifically in the context of mathematics achievement tests. According to these authors, while there is always some degree of measurement error involved, "by continuing to model mathematical proficiency using a model that assumes the construct is unidimensional, when we have substantive and empirical reasons to believe mathematical proficiency is a multidimensional construct, we are, perhaps unwittingly, increasing our error of measurement" (p. 24).

Item analysis often involves the estimation of IRT item and ability parameters. An oft-cited advantage of using an IRT approach is the property of invariance of item and ability parameters that are generated by the IRT models. This property states that the parameters that characterize an item do not depend on the ability distribution of the examinees who responded to the items and the parameter estimate that characterizes an examinee does not depend on the particular set of test items the examinee was administered. However, "parameter invariance properties... can no longer be completely trusted when the assumption of unidimensionality is violated" (Tate, 2002, p. 188). Because the invariance property is a

cornerstone of IRT and makes possible important applications such as equating, item banking, investigation of item bias, and adaptive testing (Hambleton, Swaminathan, & Rogers, 1991), the violation of the unidimensionality assumption can have serious consequences for these applications.

In addition to affecting model fit and item parameter estimation, violations of dimensionality can make gathering validity evidence difficult. Assessment of statistical test structure can provide empirical support of the content and cognitive process aspects of test validity (AERA/APA/NCME, 1999). Indications of multidimensionality could weaken the validity evidence. Unintended factors introduce construct-irrelevant abilities which pose threats to test fairness due to item bias and differential item functioning (DIF). On the other hand, the finding that more than one construct is measured by a test may support the test framework if multidimensionality was intended.

Finally, although dimensionality is related to nearly every other aspect of measurement, it is of particular concern if it is necessary to link tests. Kolen and Brennan (2004) note that there are several factors that might affect linking: the design of the data collection, the complexity (dimensionality) of the subject matter area and the curriculum dependence of the subject matter. And, as noted by Chin, Kim, and Nering, "linking scores from several measurements can only be sensible when all measurements involved share a single underlying construct. This unidimensionality assumption is often questionable for a vertical scaling operation" (Chin, Kim, & Nering, 2006, p. 2). Dimensionality violations can affect the construction of vertical scales via the IRT parameter estimations that are used in the developmental methods.

When using IRT scaling methods, item parameters are typically estimated by concurrent estimation which requires only one computer run, or by separate estimation which involves estimating parameters for each grade. Violation of the unidimensionality assumption might be most severe when concurrent estimation is used since the assumption "requires that a single ability be measured across all grades, which seems unlikely with achievement tests" (Kolen & Brennan, 2004, p. 391). With separate estimation, violations of the IRT unidimensionality assumption may have less impact on the parameter estimates because the parameters are estimated for only one grade level at a time (Young, 2006).

Summary and Purpose of the Study

The extent and importance of educational testing has increased in recent years. Children entering kindergarten are tested several times each year as they progress from grade to grade until they reach high school where they encounter a graduation test and most likely admission tests to college and career placement. The high stakes attached to these test results affect not only students but also parents, teachers and administrators as well. Evidence of student learning and accountability are important issues. Therefore, the ability to quantify student achievement and learning has become a critical and constructive tool.

In addition to the increased prevalence of educational testing, standardized testing has also received an increased amount of criticism and controversy. Parents and educators are concerned about the effects of test anxiety, the narrowing of the curriculum and the amount of time spent preparing for tests to name a few typical critiques. These points are well taken and monitoring the use and consequences of testing is warranted. This involves policy and decisions that many school boards confront and debate quite regularly. However there is

another aspect to consider—the technical quality of the tests themselves. If such important consequences and rewards are being attached to the scores, growth, annual yearly progress (AYP), then the technical quality of the tests that are being administered should be monitored as well.

One of the fundamental responsibilities of test developers is to ensure that these tests are high quality, fair, meaningful and valid instruments of student achievement. Assessment of test dimensionality is an important part of the development, evaluation, and maintenance of large-scale tests and scales. Test dimensionality is the minimum number of abilities that accounts for student performance on a set of items. It is a key concept that underlies most of the central issues in the development and use of large-scale tests. Many of the commonly used test models assume unidimensionality. These test models are the basis for developing student scores and it is these student scores to which are attached high stakes for both students and schools. Unidimensionality is also assumed in the development of vertical scales which are used to monitor student growth over time. However, multidimensionality may be present (either intentionally or unintentionally).

Unintentional sources of multidimensionality may exist particularly in a subject like mathematics where mathematics achievement tests typically measure a combination of several areas such as algebra and geometry. Unintentional sources of multidimensionality may also be introduced when developing a vertical scale and the inclusion of off-grade-level items on a test. Therefore the purpose of this study is twofold; 1) to examine the stability of the dimensional structure across elementary grades mathematics achievement tests; and 2) to investigate the dimensional structure of these mathematics achievement tests in situations

where vertical linking items (below and above grade level) are included in on-grade level tests.

CHAPTER 2

REVIEW OF LITERATURE

As previously noted, test dimensionality is defined as the minimum number of abilities or constructs measured by a set of test items. Dimensionality assessment and the implications of dimensionality are important and evolving areas in psychometric research. Historically, methods for calculating many of the proposed dimensionality indices were ad hoc with little or no rationale and no empirical support (Hattie, 1985). However, recent theoretical and empirical work and greater computing resources have yielded promising new perspectives and methods.

This chapter provides a review of the research on dimensionality. The first section discusses conceptual and mathematical definitions of dimensionality as well as related issues such as factor structure and local independence. The second section explores possible sources of dimensionality including differences in examinees and unintended sources of dimensionality that are relevant to the specific context of the proposed study: mathematics achievement testing. The third section reviews methods of investigating dimensional structure. The methods are categorized into two families: parametric and nonparametric methods. The final section presents consequences of violating dimensionality assumptions and is divided into the three test development areas affected by dimensional assumptions: item analysis, validity and linking.

Defining Test Dimensionality and Related Topics

When defining what constitutes the dimensional structure of a test, several perspectives and related topics must be considered. For example, there are two frequent uses of the term test dimensionality, one referring to the psychological dimensionality of a test and the other to the statistical dimensionality of a test. It is also important to consider the related concept of conditional independence of item scores. This concept is the basis of many dimensionality-related topics (e.g. local item dependence) and applications (e.g., factor analysis). In the following section, the various ways in which test dimensionality can be defined are described. The general approaches to dimensionality include: a contrast between psychological and statistical dimensionality, applications of conditional independence, the relationship between dimensionality and local independence, evaluating assumptions of local independence, and other issues related to dimensionality.

Psychological and Statistical Dimensionality

A distinction is often made to the meaning of the term dimensionality. One common application of the term dimensionality refers to the number of hypothesized psychological constructs believed to be account for performance on a test (psychological dimensionality). Another use refers to the minimum number of variables that are needed to summarize a matrix of item response data (statistical dimensionality) (Reckase, 1990). The psychological definition emphasizes the actual test content and cognitive processes required by examinees to respond to items on the test. It could also be considered as the substantive hypotheses and interpretations. For example, a mathematics word problem could be hypothesized to require two dimensions (perhaps numerical computation and verbal reasoning) to respond to the

item. Statistical dimensionality uses quantitative analytic methods to assess the interrelationships of the item responses. The meaning of these two uses may or may not be the same. According to Reckase:

> Differences in level on the mathematical variables may not translate directly into differences on the psychological constructs. Whether or not they have the same meaning is a question of the validity of the measures obtained using the particular mathematical model of the interactions of persons and test items. (1990, p. 2-3)

Gathering all available validity evidence to support inferences based on test scores is a critical component of a testing program and therefore the psychological and statistical dimensions must be considered simultaneously. Because "the nature and dimensionality of the interitem structure should reflect the nature and dimensionality of the construct domain" (Messick, 1993, pp. 43-44), Camilli, Wang and Fesq (1995) believed that judgments regarding the content of a test should also affect those regarding the dimensionality of the test. They argue that statistical procedures alone (such as factor analysis) provide an incomplete conceptualization of dimensionality because dimensionality is dependent not only on the set of items and a particular set of examinees, but also on the test use. Similarly, Tate (2002) recommended that determination of dimensionality should be guided by substantive considerations based on the content and purpose of the test. Therefore, the final assessment of dimensionality should incorporate both judgments about test content and statistical analyses evidence. While substantive considerations may involve qualitative expert review, test dimensionality based on statistical criteria requires an empirical model. The foundation for these statistical models is conditional independence of the item scores.

Conditional Independence

Two random variables (i.e., examinee responses to two items), x_1 and x_2, are conditionally independent given θ if, once θ is known, the value of x_2 does not add any additional information about x_1. In other words, the trait value (θ) provides all the relevant information about an examinee's performance. Slightly different, but conceptually similar, forms of conditional independence of item scores are used in factor analysis and in item response theory (Lord & Novick, 1968). In IRT, conditional independence is described in terms of local item independence. When a pair of items is locally independent, the conditional probability given an examinee's ability level, θ, is the product of the probabilities for separate items. That is, once θ known, then the performance on one item is independent from another item. Similarly, for factor analysis, once the first factor (θ) is defined and removed (i.e., conditioned on) from the data, the residual correlation matrix describes any unaccounted factors or other abilities. Researchers have investigated dimensionality using two approaches to conditional independence: (1) evaluating the assumptions of local independence and (2) a factor analytic perspective.

Dimensionality and Local Independence

One of the forms used to express conditional independence is local independence. Test dimensionality is closely related to the concept of local independence. According to Lord and Novick (1968), local independence means that "within any group of examinees all characterized by the same values $\theta_1, \theta_2, \ldots, \theta_k$, the (conditional) distributions of the item scores are all independent of each other" (p. 361). In other words, once the *k*-common traits

are controlled for, the responses to any item are unrelated to the responses to any other item. A more formal mathematical definition is as follows:

$$P(U = u \mid \boldsymbol{\theta}) = \prod_{i=1}^{n} P_i(U_i = u_i \mid \boldsymbol{\theta}) \tag{1}$$

where U represents a vector of binary variables taking the values of 0 or 1, $\boldsymbol{\theta}$ is a *k*-dimensional vector of latent traits and $P\{\ \}$ represents probability and $P_i\{\ \}$ represents an item response function. If a population of examinees is characterized by *k* latent traits which completely span (i.e., define) the latent space, then the responses of a subpopulation of examinees with fixed values for $\boldsymbol{\theta}$ are mutually independent. If, however, a model specifies a number of latent traits less than *k*, which do not completely span the latent space, then there will still remain mutual dependencies among the items for fixed values of θ (Berger & Knol, 1990). In summary, according to Lord and Novick (1968), "the assumption of local independence is equivalent to the assumption that the $\theta_1, \theta_2, \ldots, \theta_{k'}$ under consideration span the complete latent space" (p. 361).

It follows that "in order to determine the dimensionality of a set of items it is necessary and sufficient to identify the minimal set of traits such that at all fixed levels of these traits the item responses are independent" (Hattie, Krakowski, Rogers, & Swaminathan, 1996, p. 1). McDonald (1981) noted that local independence is the principle on which the notion of dimensionality is founded. Furthermore, Traub (1983) concluded that there is no meaningful definition of unidimensionality and no basis for a test of dimensionality without local independence.

An item might display *local dependence* because responses to the item are related to (i.e., not independent of) responses to one or more other items, again controlling for ability

level. For example, local dependencies might be seen in the items that follow a passage on a reading comprehension test. Local dependencies might also appear when speediness of test completion is a factor; some of the items at the end of the test might be omitted and thus be locally dependent (Chen & Thissen, 1997). Item dependencies can be positive or negative (Habing & Roussos, 2003; Yen, 1984). Locally dependent items are redundant in the sense that they contain less information than the IRT model would predict. Dependencies can also have an effect on item parameter estimates. Clusters of locally dependent items make a test multidimensional (Wainer & Thissen, 1996). There are several indices to measure local dependence (Chen & Thissen, 1997) and also several methods of assessing dimensionality based upon local item dependencies (Tate, 2003).

Evaluating assumptions of local independence. As previously shown, evaluating the number of latent traits needed to obtain local independence provides a powerful tool for assessing test dimensionality. In particular, the traditional IRT conceptualization of dimensionality makes no distinction between major (dominant) and minor dimensions (Nandakumar, 1991). The notion that test performance is governed by a dominant latent trait and several nuisance or nondominant latent traits has necessitated a distinction between *strict* dimensionality and *essential* dimensionality. To illustrate these concepts it is first necessary to consider that IRT models require a dual assumption: local independence and monotonicity. Therefore, dimensionality can be defined as the minimum number of traits necessary to satisfy local independence and monotonicity (Stout, 1990). A *monotonically increasing* function is one that preserves the order with increasing values; if $x_1 \geq x_2$ then $f(x_1) \geq f(x_2)$. Items in an achievement test should have a monotonically increasing item response

functions; the more ability possessed by the examinee, the greater the probability of success (van der Linden & Hambleton, 1997). Using Stout's definition, the difference between strict and essential dimensionality is the difference in conditions of local independence. Strict dimensionality requires the *strong principle of local independence* (SLI).

The strong principle of local independence (SLI), which requires that item responses be statistically independent for fixed values of the traits, is very stringent. It is mathematically defined by McDonald & Mok (1995) as:

$$P\{U = u \mid \theta\} = \prod_{i=1}^{n} P_i\{U_i = u_i \mid \theta\} \tag{2}$$

Notice that the definition of SLI given in Equation 2 is simply the definition of local independence presented in Equation 1. SLI requires that, for fixed values of the traits, not only the covariances be 0, but that all higher-order moments be products of the univariate moments. That is, local independence is a broader assumption than zero correlations; local independence also includes nonlinear and higher- order relationships among the items (Embretson & Reise, 2000). SLI is almost impossible to attain in practice (Stout et al., 1996). Researchers have proposed two weaker forms of SLI: 1) the *weak principle of local independence* (WLI) and 2) an even weaker form called the *principle of essential independence* (EI) (Stout, 1990). Many of the techniques for assessing test dimensionality are grounded in at least one of these three principles of local independence.

The weak principle of local independence (WLI) requires that only the covariances among the items be 0 for fixed values of the traits. McDonald (1981) defined the weak form of local independence by,

$$Cov\{U_j, U_k \mid \theta\} = 0,\ j \neq k$$

When the item responses (conditional on the trait values) have a multivariate normal density, the weak principle implies the stronger principle. McDonald also argued that in practice SLI is rarely violated in cases where WLI holds true.

The principle of essential local item independence (EI) is a weakest form of SLI (Stout, 1990). While WLI requires the covariances to be zero for all θ, EI merely requires that the average of the magnitude (absolute value) of the covariances conditioned on a fixed value of the latent traits converges to zero as the test becomes very long. In other words, essential independence requires that the average value of $\left|Cov\left(U_i, U_j \mid \Theta = \theta\right)\right|$ over all item pairs approach zero for all θ as the test length increases (Nandakumar & Stout, 1993). Assuming essential unidimensionality requires that the items exhibit EI. *Essential unidimensionality* is based on the premise that a dominant trait exists with the possible presence of several minor dimensions. It assumes the dominant dimension is so strong that the trait estimates are not affected by the presence of smaller dimensions (Smith, Jr., 2004).

Dimensionality and Factor Analysis

Another important approach to assessing dimensionality using conditional independence is factor analysis. Factor analysis explores the covariances among items and is an empirical way of studying the construct(s) measured by a test. If the test is designed to measure a certain number of factors (constructs), the items should group themselves according to the factors they were intended to measure. The magnitude of the item loadings across factors may be used to assess the dimensionality of a test. The larger the loading value, the stronger the relationship between the item and the factor. For example, if a set of items is constructed to measure only one construct, then each item should have a large

loading only on that one factor and have weak, almost negligible loadings on any additional factors.

Defining and removing the first factor is roughly equivalent to conditioning on θ. If the residual correlations are all zero, then one factor accounts for the differences in test performance. Or in terms of local independence, there are no local item dependencies since locally dependent items would have nonzero residual correlations after removal of the first factor. However, if a set of items has unusually large residual correlations (indicating local dependences), then two or more factors could also be defined. Factor analytic procedures are available to explore and measure the correlations among factors.

Other Issues Related to Dimensionality

There are three other issues related to dimensionality that will be described in this section. These issues are related to possible correlations among traits or factors. For one, some researchers have suggested that a test may be considered to be unidimensional if the items assess the same combination of skills. This slight deviation of the traditional IRT concept of unidimensionality has been raised by Reckase and Ackerman (1986). A second issue, situated in the factor analytic framework, is test structure. To examine this issue, distinctions among types of factorially simple and types of factorially complex test structures are described. The last issue introduces compensatory and noncompensatory models.

Unidimensionality and Combination of Same Skills

Reckase and Ackerman (1986) have argued that the concept of unidimensionality required by IRT (i.e., all persons with the same estimate of ability have the same probability of a correct response for each item) is not the same as the commonly held conception of unidimensionality. According to Reckase and Ackerman, the IRT definition "does not require that the estimate of ability be a function of a single psychological trait or construct. As an alternative, each item could require the same combination of traits or constructs" (pp. 2-3). Therefore, from an IRT perspective, a test would be unidimensional as long as all of the items required the same combination of skills.

To pursue their alternative conceptualization, Reckase and Ackerman (1986) studied selected mathematics achievement items using a *multidimensional item difficulty* (MID) statistic. This statistic indicates the combination of skills that form a multidimensional space for which the test item provides the best discrimination. It is described by two pieces of information: the direction from the origin of the multidimensional space to the point in the space where the item is most discriminating, and the distance from the origin to that point. Thus, a set of items with the same direction function as if they were unidimensional even though they may require more than one skill to respond correctly. Items that have different directions will form a multidimensional set from the IRT perspective. Once the MID statistics were computed, items were sorted according to direction and then the item sets were analyzed using a specified unidimensional IRT model. The results strongly supported the conception of unidimensionality suggested by a common direction in the multidimensional space for a set of items and the use of MID statistics in forming unidimensional item sets.

The results obtained by Reckase and Ackerman (1986) were supported in further studies of both simulated and real data by Reckase, Ackerman, and Carlson (1988). These authors concluded that "rather than specifying that items need to measure only a single trait, the results presented here show that the unidimensionality assumption implies that items need only measure the same composite of abilities as indicated by multidimensional IRT analysis" (p. 203). However, it seems like a composite is fairly unique to each set of items. Therefore, writing new items, developing test forms and equating procedures that all measure or combine to measure the same, exact initial composite of skills would be extremely difficult.

Factor Analysis and Test Structure

Another topic related to dimensionality is test structure. Recall that in factor analytic terms, test dimensionality is described by factor loadings. A set of items that load on only one factor is called *factorially simple*; it measures one underlying attribute. A conceptual schematic of a factorially simple test structure is shown in Figure 2.1 (a). In Figure 2.1, the items are represented by dots and the distance from the origin to the dot is, in a very simplistic approach, a measure of the magnitude of the item's discrimination. The more highly discriminating items have longer distances. Notice that a unidimensional test is factorially simple; all the items lie along one dimension.

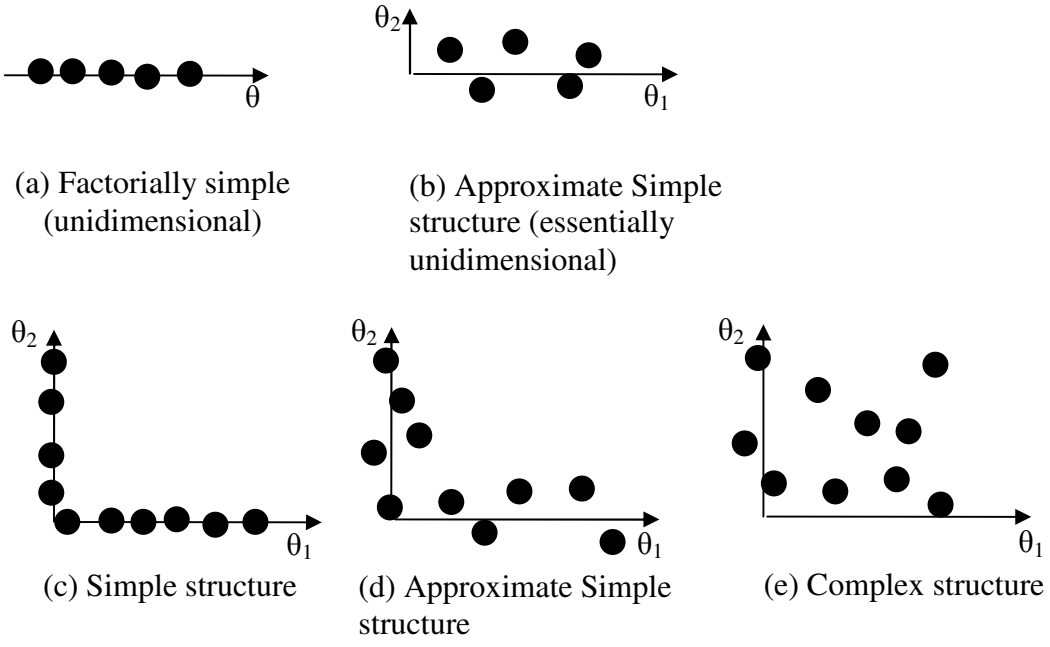


Figure 2.1. Conceptual schematics of test structures.

A concept similar to Reckase and Ackerman's idea of a composite unidimensional trait is a factorially complex structure. If an item or a test measures two or more factors, it is referred to as *factorially complex*. There are three types of factorially complex tests depending on the data: simple structure, approximate simple structure, and complex structure. Simplified schematics for these structures are shown in Figure 2.1 (b) – (e).

If each item on a test measures one, and only one dimension, the test structure is labeled as *exact or simple structure*. Exact or simple structure is defined to exist for a *k*-dimensional test if a *k*-dimensional latent coordination system exists such that all the items lie along the coordinate axes (Stout et al., 1996). A two dimensional (*k*=2) example of simple structure in shown in Figure 2.1 (c). Notice that all the items lie closely along the two dimensional axes. In factor analytic terms, simple structure refers to the situation in which

the factor loadings are either very large, suggesting a clear relationship between the item and factor, or very small suggesting no relationship at all between the item and factor (L. D. McLeod, Swygert, & Thissen, 2001). However, real test data are rarely represented by simple structure.

An alternative to simple structure is *approximate simple structure*, in which an item primarily measures one dimension, with loadings to a lesser extent on the other dimension(s). There are two possible scenarios. One possibility is the presence of a dominant dimension with one or more nuisance dimensions present as in the case of essentially unidimensional shown in Figure 2.1 (b). The items do not lie as closely around θ_1 as they do for the unidimensional case in (a) indicating the correlation of the items to θ_2 as well as θ_1. A second scenario would be the multidimensional equivalent to essential unidimensionality. Two dominant dimensions are shown in Figure 2.1 (d). In this two-dimensional example, the items lie in a close sector around the two-dimensional coordinate axes. In other words, items that display approximate simple structure are mainly sensitive to one trait and only marginally to other traits.

If the items load highly on multiple dimensions, then the structure is referred to as a *complex structure*. Complex structure is typical of most educational testing (Ackerman et al., 2003; Sass & Walker, 2006). A complex structure is displayed when the items lie throughout the two-dimensional coordinate axes (i.e., items measure a range of skills in the $\theta_1\theta_2$ composite) (Ackerman et al., 2003; Gierl, Tan, & Wang, 2005; Stout et al., 1996). This is shown in Figure 2.1 (e). The items along the axes measure one of two dimensions. Also notice the items that appear in-between the two axes. These items measure a composite or combination of both dimensions. For example, a mathematics test could conceivably be

constructed of subsets of items. One subset of items might require algebra skills, θ_1; another subset might involve geometry understanding, θ_2; still another subset might require a combination of both algebra and geometry proficiencies, $\theta_1\theta_2$.

Compensatory and Noncompensatory Models

A third issue to consider when describing multidimensional models is the potential presence of compensation among the abilities required to answer test items correctly. *Compensatory models* assume that high ability on one dimension can compensate for low ability on another dimension in terms of the probability of a correct response. In *noncompensatory models*, sufficient levels of each measured ability are required, and a deficiency in one ability cannot be completely offset through an increase in others. The question of whether the compensatory or the noncompensatory model is more appropriate in applied testing situations is debatable. Either way, some researchers have suggested that there is no way to practically determine whether real data are compensatory or noncompensatory (Way, Ansley, & Forsyth, 1988).

Summary

This portion of the review of literature has summarized the general definitions and topics associated with dimensionality. Dimensionality typically has two connotations: the number of hypothesized, psychological constructs measured by a set of items, and the statistical model needed to describe the interrelationships among item responses. Both perspectives are important to understanding what items/ tests are measuring and evaluating the congruence of the nominal, intended test specifications with the effective test structure.

The foundation of the statistical analyses and test theory models used to describe and thereby assess dimensionality is conditional independence. Forms of conditional independence are the basis for methods such as factor analysis and local independence. The principle of local independence provides a mathematical definition such that once the number of traits is determined and conditioned on, the responses to items are statistically independent. This strong assumption is rarely met in real test data due to the presence of either intentional or unintentional sources of dimensionality. The distinction between intentional (multiple dimensions) or unintentional (perhaps a dominant dimension with nuisance factors) sources is considered in the next section.

Sources of Dimensionality

Achievement test data are often intentionally or unintentionally multidimensional. When an achievement test is purposely designed to measure a constellation of differing knowledge domains, skills, or constructs, multidimensionality is intended, or at least expected. However, in other contexts, multidimensionality may not be intended. Sources of multidimensionality include the many individual differences each examinee brings into a test administration. Multidimensionality can also be the consequence of content complexity. For example, mathematics educators, including the National Council of Teachers of Mathematics (NCTM), have struggled with defining and agreeing on evidence of mathematical proficiency particularly in regards to procedural knowledge and conceptual knowledge. Furthermore, the distinct content areas that comprise grade-level content in mathematics and emphasis on applications and conceptual knowledge create complex mathematics curricula and potentially multidimensional assessments. Other sources of multidimensionality may result from the test development process including test specifications, item difficulty and linking projects such as vertical scaling. The following sections describe in greater detail these potential sources of multidimensionality, with special attention to the presence of multidimensionality in linked assessments.

Differences in Examinees

As previously mentioned, assessment of dimensionality is related to the concept of conditional independence. In IRT, conditional independence is described in terms of local item independence. The assumption of local independence is not satisfied if the inter-item

correlations differ across subpopulations of examinees with equal latent trait values (θ). That is, items should not perform differently for subgroups of examinees with equal θ values. However there are always other cognitive, personality, and test-taking factors that have an impact on test performance and affect an examinee's observed score (Hambleton & Swaminathan, 1985). For example, poor performance on a mathematics test may be attributed to test anxiety and/or math anxiety. For certain students, *stereotype threat* may impact test performance. The research of Steele and his colleagues suggests that societal stereotypes (i.e. certain groups like females or African-Americans do not do well at math), not solely mathematics proficiency, impairs standardized test performance of females and African Americans (Steele, 1997). Another unintended source of multidimensionality is differences in examinee motivation. Motivational effects can have impact on test performance positively as well as negatively. For example, test developers grapple with differential motivation and effort of students, particularly those of high school age, on field test (or non-operational) items. In addition, some students will be more persistent that others during testing.

There are at least two other factors related to the individual examinees that will produce unintentional test multidimensionality: speed of test taking and willingness to guess (Traub, 1983). If the test administration is timed, then speed of work is introduced as a factor. For examinees with the same score on the latent variable in question (e.g., mathematics computation), a timed test administration will separate those who work quickly to answer all the items in allotted time from those who do not. If test items are amenable to guessing, then local independence assumptions would be violated when examinees with equal standing on the latent construct and different propensities to guess are confronted by an

item for which the do not know the correct response. Failure to consider guessing effects could "produce artificial factors and misleading information as to the number of factors needed to account for the data" (Lawrence & Dorans, 1987, p. 2). Furthermore, in a recent study comparing methods for assessing dimensionality and factor structure for binary scored items in which the examinees may be guessing, Stone and Yeh (2006) found that if guessing is relevant to the testing application, then modeling guessing in the analysis of dimensionality may be important. This supported similar results reported previously by Tate (2003).

In conclusion, examinees vary widely in their cognitive processes, personality, psychological attributes, and educational characteristics. Individuals also differ in terms of cultural factors and out-of-school learning experiences that are relevant to in-school learning and classroom experiences. While tests are intended to measure an individual's achievement, care must be taken to ensure that a test is not measuring other unintended individual differences such as anxiety or motivation.

In addition to individual examinee differences, various classroom circumstances can introduce multidimensionality. These can include "the effects of improved teaching methods on more recent samples of students, changes in emphasis and curriculum that took place between pretest and operational administration, and the ability of examinees to recognize and therefore have different motivation on pretest section" (Stocking & Eignor, 1986, p. 21). This can vary widely from student to student, even within the same class (Traub, 1983). Classrooms in turn are shaped by several factors particularly the curriculum and subject matter. Mathematics content presents many challenging topics so possible sources of multidimensionality are introduced due to current theories of mathematical proficiency, methodological approaches to teaching, and NCTM influences in curricula.

Mathematical Proficiency

Previously in this chapter, consideration was given to defining what constitutes the dimensional structure of a test. Consideration is especially warranted when the construct to be measured is mathematical proficiency (also referred to as mathematics achievement). Mathematics proficiency is a complex and multifaceted domain as shown in Figure 2.2. For example, under the umbrella of mathematical proficiency are distinct subdomains or strands such as Geometry and Algebra that comprise related, but different, knowledge and skills. Within each strand, a mathematically proficient student would need to possess the skills to be able to do computations as well as applications or problems situated within a given context. Underlying these skills are procedural knowledge and conceptual knowledge, respectively. Rittle-Johnshon, Siegler and Alibali (2001) offered the following definitions of procedural and conceptual knowledge:

> We define *procedural knowledge* as the ability to execute action sequences to solve problems. This type of knowledge is tied to specific problem types and therefore is not widely generalizable ... such as counting a row of objects or solving standard arithmetic computations... we define *conceptual knowledge* as implicit or explicit understanding of the principles that govern a domain and of the interrelations between units of knowledge in a domain. This knowledge is flexible and not tied to specific problem types and is therefore generalizable. (p . 346-347)

Thus, procedural knowledge refers to computational skills (e.g., adding two digit numbers), knowledge facts (e.g., multiplication tables) or solving one-step equations. Conceptual knowledge typically includes other processes such as reasoning, reading, problem solving, communication and making connections among topics. The NCTM *Standards* describe these as "process standards".

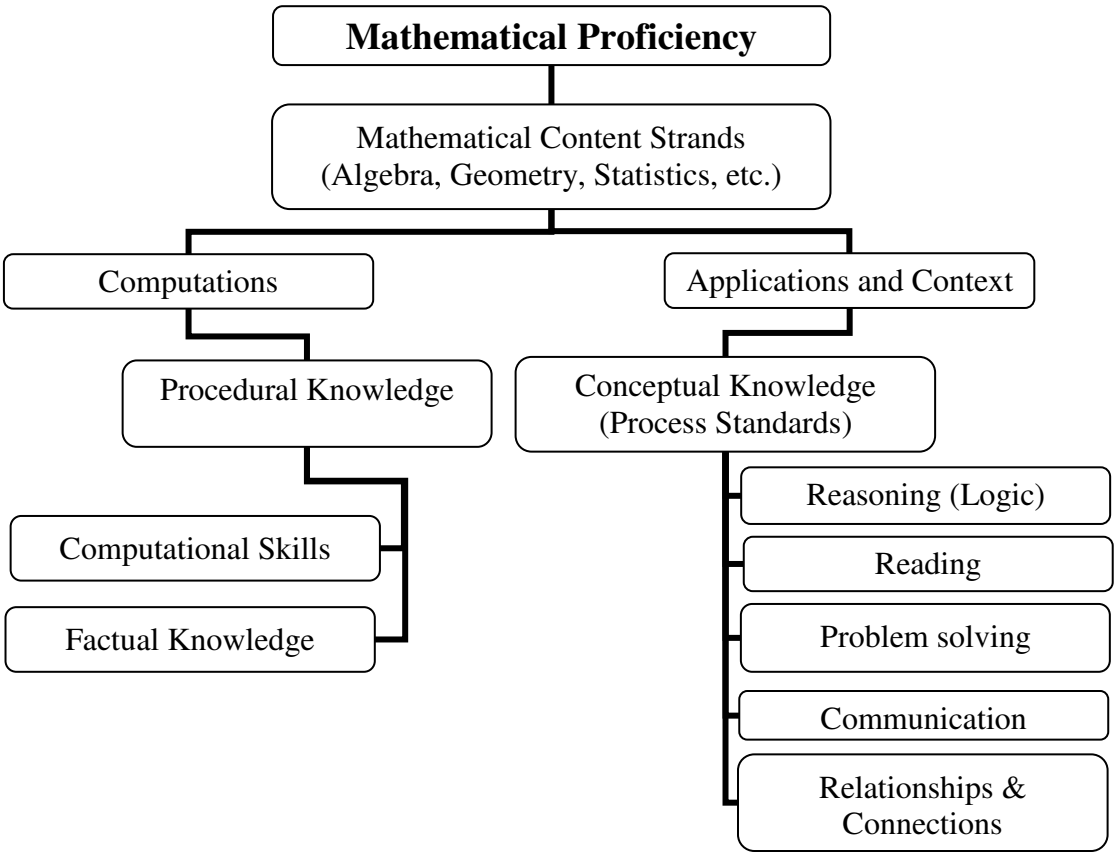


Figure 2.2. Components of Mathematical Proficiency

Educators disagree regarding the introduction and emphasis of procedural knowledge and conceptual knowledge (sometimes referred to as conceptual understanding or learning with understanding). Most theories of the development of procedural knowledge and conceptual knowledge have focused on which type of knowledge first develops in a given domain (e.g. counting, simple arithmetic, adding fractions). Mathematics education researchers debate whether students learn best by memorizing procedural knowledge (such as multiplication facts) or by emphasizing the concepts behind the procedures (i.e., modeling and understanding the reasoning of the multiplication process). According to procedure-first theories, students should acquire knowledge of a procedure first and then later develop the

concepts of why the skills work the way they do. On the other hand, concepts-first theories purport that students initially develop conceptual knowledge and then use this conceptual knowledge to generate procedures. While some researchers (e.g., Carpenter & Lehrer, 1991), reported that there is a mounting body of evidence that supports the importance of learning with understanding from the beginning, other researchers (Rittle-Johnson et al., 2001) hypothesized that student's conceptual and procedural knowledge develop iteratively. That is, an increase in one type of knowledge leads to gain in the other type of knowledge. The reform efforts and standards developed by the NCTM have largely embodied an emphasis on instilling conceptual knowledge before teaching procedure knowledge.

Assessing Mathematical Proficiency

Measuring conceptual knowledge and process skills is challenging. Rittle-Johnson, Siegler, Alibali observed that "to assess conceptual knowledge, researchers often use novel tasks, such as counting in nonstandard way or evaluating unfamiliar procedures" (2001, p. 347). The NCTM *Standards* emphasize the need for students to spend more time on problem solving and reasoning skills particularly with non-routine mathematical items that students would encounter outside the classroom. This shift to a more *authentic* tasks provides a fairly multidimensional vision of what it means to know and understand mathematics (Romberg, 1995). According to Romberg,

> For a task to be considered "authentic," it should not easily fit into neat categories of single content areas and single processes. Solving nonroutine problems usually involves multiple processes and cuts across mathematical domains. Making connections necessarily involves blurring the lines between content and processes. (p. 9)

Nonroutine items that are designed to measure problem solving and reasoning skills are often placed in situational contexts that require students be able to read and comprehend the problems. However, context can add potential cultural loadings to an item or test. Certain words or context may not be understood by all examinees depending on their SES, gender, educational background, personal experiences, region of residency, or other factors. In addition to mathematical proficiency, context inserts another dimension by requiring reading ability. Many mathematics problems require two skills: a verbal skill to determine what is required by the problem and a mathematical skill to solve the problem (Reckase & Ackerman, 1986).

Language ability is, in general, a predictor of math performance. Carpenter, Corbitt, Kepner, Linquist and Reys (1980) reported that 10% to 30% of children perform worse on word problems than on comparable problems presented in numeric format. Many researchers have concluded that the discrepancy between performance on verbal and numeric format problems strongly suggests that factors other than mathematical skill contribute to success in solving word problems and that the interaction between language and mathematics achievement is real (Abedi & Lord, 2001; L. D. McLeod et al., 2001). According to Reckase (1990),

> Numerical computation and verbal reasoning are said to be required to successfully perform on a mathematics story [word] problem. Numerical computation and verbal reasoning are two psychological dimensions that are hypothesized to exist to explain differences in performance on the test item. (p. 4)

To investigate the contribution of these two factors, Abedi and Lord (2001) acquired released items from the National Assessment of Educational Progress (NAEP) and modified the items to reduce their linguistic complexity. These items (both the original and the

modified versions) were then administered to over 1,100 8th grade students. Linguistic modification of test items resulted in significant differences in math performance; scores on the linguistically modified version were slightly higher. The results also showed several other group differences on test performance overall (i.e., both the original and the modified versions of the items). For example, there were differences in math performance with respect to SES but not gender. Students who were English language learners (ELLs) scored lower on the math test than proficient speakers of English. These results were consistent with previous research studies. Reckase, Davey and Ackerman (1989) reported "there is a fairly clear distinction between the arithmetic items that are in story [word] problem format and those items that only require computation or formula manipulation. However, the constructs measured by these two different types of items are highly intercorrelated" (p. 10). A mathematics test that contains some items that are strictly computational and other items that involve verbal material is not likely to be unidimensional (Kolen & Brennan, 2004). This type of mathematics test would, in effect, be an assessment of at least two abilities (i.e., reading and mathematics.

Placing mathematical problems in context is not the only challenge to measuring mathematical knowledge. Guided by the conceptual knowledge framework of the NCTM *Standards*, some large scale testing programs have incorporated polytomous, constructed-response items to measure the communication of mathematical ideas. These items attempt to capture the process of learning in addition to the product (Walker & Beretvas, 2000). In order to do well on items measuring mathematical communication, such items may require that students be able to clearly communicate graphically, numerically, and/or in writing. It

was determined that such items function differently in favor of proficient writers and the data were considered to be multidimensional (Walker & Beretvas, 2000).

Mathematics Standards and Classrooms

Many states, schools and classrooms have modeled their curriculum and methods after the NCTM *Standards.* Other more populous states such as California or Texas use state-developed curricular frameworks that are similar to NCTM *Standards*. The similarity actually stems from the reference of state curricular guidelines during the development of NCTM *Standards*. For example, the California *Framework* (California Department of Education, 1985) was mentioned frequently during the writing of the NCTM standards (D. B. McLeod, Stake, Schappelle, Mellissinos, & Gierl, 1996). The California *Framework* contains five content strands (Number Sense; Algebra and Functions; Measurement and Geometry; Statistics, Data Analysis, and Probability; and Mathematical Reasoning). Notice these strands are very similar to the NCTM *Standards* strands but with a slight variation. The NCTM *Standards* split Measurement and Geometry into two strands and refer to Mathematical Reasoning as a process skill that is integrated throughout the curricula and grades (National Council of Teachers of Mathematics, 2000). Regardless, as indicated previously, many aspects of the instructional environment can introduce multidimensionality. One source is the curriculum itself which delineates the coverage of topics, the depth of those topics and when (grade-wise) topics are introduced. It appears that by the current definition of mathematical understanding (i.e., NCTM) and perhaps by its very nature, mathematics proficiency is multidimensional. This is further corroborated by the different content areas as defined by the NCTM *Standards*.

The NCTM *Standards* are descriptions of what mathematics instruction should enable students to know and to do by specifying the understanding, knowledge and skills that students should acquire from prekindergarten through grade 12. The NCTM Standards are divided between content standards and process standards. The content standards explicitly describe the content students should learn in five strands: (1) Numbers and Operations, (2) Algebra, (3) Geometry, (4) Measurement, and (5) Data Analysis and Probability. The process standards emphasize the need for students to spend more time on mathematical problem solving and reasoning, communicating mathematical ideas, making connections among mathematical topics and exploring relationships among representations of mathematical forms. Both the content standards and the process standards may cause unintended sources of test dimensionality.

Mathematics Curriculum

Mathematics is a multifaceted domain which requires a comprehensive curriculum which in turn introduces other possible sources of multidimensionality. The various mathematical topics are reflected in the content standards developed as part of the NCTM *Standards*. The NCTM Standards view a coherent curriculum as one that effectively organizes and integrates important mathematical ideas so that students can see how the ideas build on or connect with other ideas. In other words, the curriculum is designed to deepen conceptual understanding. A key strategy for accomplishing this goal is to address each of the five strands during a school year. The amount of instructional time spent on each strand varies by grade level. Based upon the *NCTM Principles and Standards (National Council of Teachers of Mathematics, 2000)* which were shaped by the theoretical perspectives,

methodologies, and findings of research (Ferrini-Mundy & Martin, 2003), NCTM developed a figure that demonstrates roughly how content strands might receive different emphases across the grade bands. This graphical representation is shown in Figure 2.3. For example, note the amount of coverage that is allotted to Number Sense and Operations shown in Figure 2.3; this strand receives considerable emphasis in the early elementary years but decreases over time. The opposite is true for the Algebra strand; it receives little emphasis in the early elementary years but emphasis increases across the grades.

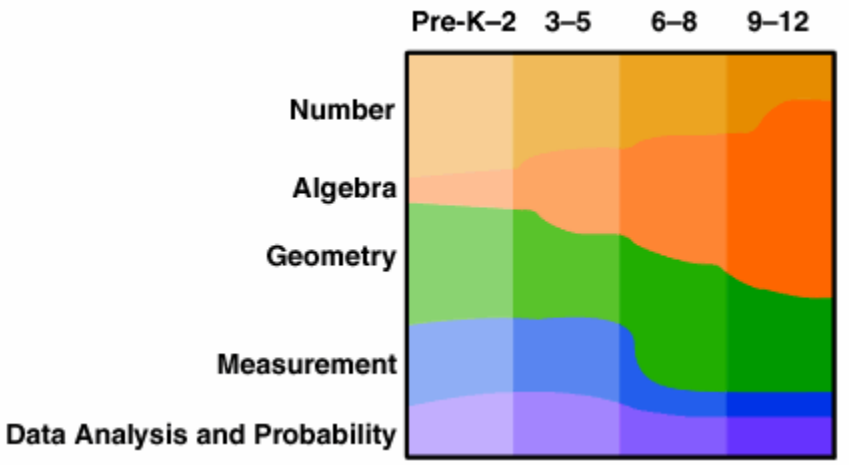


Figure 2.3. NCTM Content Standards Across the Grade Bands (National Council of Teachers of Mathematics, 2000)

Source: (National Council of Teachers of Mathematics, 2000)

Research on curricular differences and dimensionality is rather limited. Phillips and Mehrens (1987) investigated whether linear factor analysis (a commonly used method to investigate dimensionality) was sensitive to measurable curricular differences within a school district. In other words, the authors were interested in whether the curricular differences were great enough to disturb the measurement of the intended ability. The authors

considered curricular differences to be differences in the stress placed on the results of standardized tests by the principals, different textbooks and other curricular materials used, and the differences in amount and focus teachers placed on the test content and results. The analyses used both student test scores and responses of school personnel to rate each school in the district on the match between instruction in the building to the standardized test. Based upon the results, the authors concluded that "curricular heterogeneity appeared *not* to be a potent concern in the possible violation of the unidimensionality assumption of IRT" (p. 14). However, the authors cautioned that several important issues should be considered in interpreting the results. Several concerns stemmed from the use of linear factor analysis and tetrachoric correlations (these will be discussed in the assessment section of this chapter) and methods of quantifying curricular differences. Note that the curricular differences did not take into account the content areas across grades and alignment to standards. This is an unexplored area in educational measurement and will be addressed in the proposed study.

To the extent that test developers—including states that produce standards-referenced tests to comply with No Child Left Behind (NCLB)—adhere to the NCTM *Standards*, their tests will reflect the NCTM content emphases (Feuer, Holland, B.F., Bertenthal, & Hemphill, 1999). As illustrated in Figure 2.4, the content strands present two potential sources of dimensionality: the division of mathematical content into the strands themselves and the grade-varying strand emphasis. The division of mathematical knowledge into the strands themselves potentially creates sources of dimensionality. By establishing the different strands, the NCTM *Standards* are indicating special knowledge and skills that are unique for a specific domain (e.g., Geometry compared to Algebra). Therefore, items written to specific strands could potentially create a dimensional structure reflecting those intended strands.

Consequently, reporting one "math" total score across grades could be problematic. For example, a single mathematics score is reported for a Grade 3 examinee and a single mathematics score is reported for a Grade 8 examinee. However, due to the changing strand emphasis over the curriculum, a Grade 3 form is more heavily weighted on number sense and operations while the Grade 8 form reflects the dominant focus on algebraic reasoning and skills. A single total score usually implies a unitary construct and vice versa (Messick, 1993).

The NCTM *Standards* have been widely adopted by states on a general level. However, at a specific level, adherence to them may vary. Differential probabilities of success could be due to lack of emphasis on or lack of introduction of some aspect of mathematical knowledge (Bogan & Yen, 1983). "Achievement tests are typically designed to measure a complex of skills related to a curriculum area. These tests are inherently multidimensional in what they measure. Yet, a single score is often reported to summarize an examinees performance on such a test" (Reckase et al., 1989, p. 9).

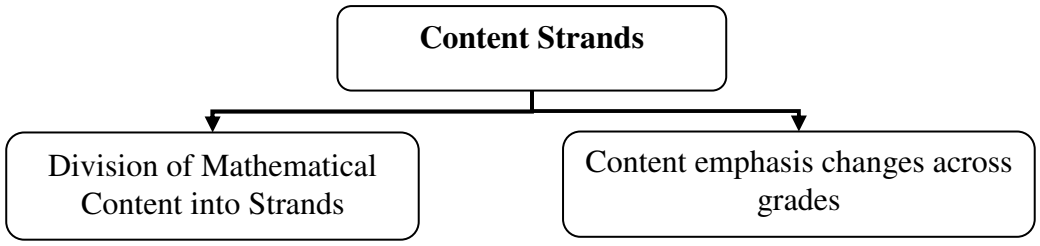


Figure 2.4. Possible Sources of Multidimensionality Related to NCTM Content Strands

Mathematics Assessments and Dimensionality

With the potential for both unintended and intended multidimensionality, it is not surprising that previous research studies have found mathematical assessments to be multidimensional. In a validation study of National Education Longitudinal Study of 1988 (NELS:88), Kupermintz and Snow (1997) demonstrated that achievement on the NELS:88 mathematics test is not adequately represented by a single dimension. They factor analyzed the multiple-choice test in mathematics and the results yielded several interpretable achievement dimensions (two to five dimensions including mathematical reasoning and knowledge). In another recent study, Gierl, Tan and Wang (2005) used several methods to assess the cognitive dimensions that characterize student performance on the SAT. (The specific methods will be described later in this chapter.) The math section of the SAT contains 54 items and covers mathematical concepts in four areas: Number and Operations; Algebra I, II and Functions; Geometry; and Statistics, Probability, and Data Analysis. While both multiple-choice and constructed response item formats were used, both item types were scored dichotomously. Exploratory analyses indicated two dimensions with a moderate correlation among dimensions. The confirmatory analyses also revealed multidimensionality in the SAT data. The authors concluded that there is a "multidimensional basis for test score inferences on the mathematics section of the SAT" (p. 26).

Overall, previous research has shown that mathematics content tends to be multidimensional. The grade-varying strands can cause potential sources of multidimensionality. This multidimensionality must be taken into account for many reasons, including when developing valid scales needed to for growth modeling and value-added

modeling. Possible sources related to the test development process including vertical scaling will be discussed in the next section.

Test Development

Several of the steps in the test development process can introduce sources of multidimensionality. As mentioned previously, some of these sources of multidimensionality can be intentional (i.e., tests developed to report subscores) or more likely, multidimensionality develops unintentionally. Test specifications and linking projects such as vertical scaling can create unplanned sources of multidimensionality.

Test Specifications

After the purpose(s) of the test are clarified, the next step is to specify the attributes of the test which will guide subsequent item development and form assembly. According to Millman and Greene, "the major function of this is step is ... to enhance the ultimate validity of the test-score inferences...foremost among the attributes of a test requiring specification is its content" (1989, p. 338). These specifications describe either the number or the proportion of items from each sub domain that are to be included on the final version of the test. Ironically, this attempt to establish content validity, can also highlight intentional (or unintentional) sources of multidimensionality depending on the test purpose.

Reckase (1979) has observed that "achievement tests are not usually constructed using methodology designed to yield factor pure measures....items are written to match the specifications. The tests produced in this way seldom measure a single trait and often will be

factorially complex" (p. 208). Reckase, Davey and Ackerman (1989) expanded further on this relationship of dimensionality and test specifications by commenting that,

> Achievement tests that are constructed with an emphasis on content specifications are likely not to be unidimensional and it is uncertain whether the current test construction process yields tests that are parallel in a multidimensional sense when that is not specifically stated as a requirement in the test development process. (p. 2)

Linking Methods and Practices

Linking distinct assessments is necessary in many testing programs. In general, linking means putting scores from two or more tests on the same scale. For example, alternate forms of a grade 5 mathematics achievement may be administered but the scores from all versions are reported on the same scale. The process of linking allows test scores obtained on one test to be related or converted to test scores on another test. Linking is commonly compared to the well-known relationship between temperature measured on the Fahrenheit and Celsius scales. The relationship, $F = \frac{9}{5}C + 32$ or equivalently, $C = \frac{5}{9}(F - 32)$ permits a kind of linking between these two temperature scales.

Various techniques are available to link one assessment to another. However, a confusing array of terminology in the literature has been associated with those techniques and the terms are not always used consistently (Kolen & Whitney, 1982). For example, the word *linking* is a generic term that includes a variety of approaches to make results of one assessment comparable to those of another. Efforts have been made to bring coherence to the terminology (Feuer et al., 1999; Kolen & Whitney, 1982; Linn, 1993; Mislevy, 1992). There are four categories or forms of linking which are listed here in terms of the strength of the

resulting linkage: equating, calibration, projection and moderation. *Equating* is the strongest or most demanding form of linking and the one with the most technical support. When a linking relationship is truly "equating", the relationship is invariant across the different populations. That is, the equated scores can be used interchangeably. *Calibration* links test or assessments that are constructed for different purposes and use different content frameworks or test specifications. Many of the statistical methods used in equating can be used in calibration but the resulting relationships are not likely to be invariant across different populations. *Projection* is a unidirectional form of linking that is used to predict or project scores on one test from the scores on another test. There is no expectation or requirement that the two tests are measuring the same construct. *Moderation* is the weakest form of linking and is used when the tests have different blueprints and are given to different, nonequivalent groups of examinees. There are two types of moderation: statistical moderation (often called "distribution matching") and social moderation which involves direct judgments concerning the comparability of performance levels on different assessments.

There also exist subtle differences in the taxonomy of types given above. Consider the second method listed, calibration. There are several connotations of "calibration". The *Standards for Educational and Psychological Testing* refer to this type of linking as "scaling to achieve comparability" (AERA, APA & NCME, 1999, p. 52). Vertical scaling is often referred to as vertical equating but it is typically considered a form of calibration. However some researchers have argued that vertical scaling is more a combination of projection and moderation (Lissitz & Huynh, 2003). Regardless, vertical scaling offers methods for assessing student gains over time and is an important procedure in educational testing.

Vertical Scaling

As a consequence of the NCLB legislation, a pressing issue for many states is demonstrating adequate yearly progress for each student. Adequate yearly progress (AYP) requires a group of students to make substantial academic progress (i.e., growth) every year in every class. Successfully estimating the progress or growth of competence requires modeling the developmental trajectory. Vertical scaling can be used to construct such a developmental scale. The goal of vertical scaling is to place tests that differ in difficulty but are intended to measure similar constructs on the same scale. It implies that the same dimensions are the focus of the teacher's efforts in each grade (Lissitz & Huynh, 2003). However, multiple test levels of mathematics may not be measuring the same construct due to potential sources of multidimensionality stemming from the changing curriculum and content emphasis across grades.

Constructing a vertical scale across grades is very complex due to the difficulty of measuring and modeling student learning. For example, as Kolen and Brennan (2004) observe, "students learn so much during their grade school years, that using a single set of test questions over a wide range of educational levels can be problematic" (p. 372). Modeling the learning process in mathematics is complicated because new topics and skills are being introduced at all grades; it is not learning how to improve one "math" skill but rather expanding and building new knowledge. In a recent study of vertical scaling of science achievement tests Reckase and Martineau (2004), made the following comments:

> Students do not gain knowledge in multiple content areas in a uniform way. Rather, the growth is on different dimensions at different times. The tests also reflect different skills and knowledge at different grade levels. These suggest

> that multidimensional models are needed to reflect the complexities of vertical scaling of science achievement. (p. 18)

Mathematical achievement may not be all that dissimilar from science achievement. Like science, mathematics has multiple content areas (e.g., NCTM strands) and students are learning different skills and knowledge at different grade levels. Many of the procedures for linking score scales assume that the tests are measuring the same construct and that the forms are reasonably parallel in their construction. Neither of these assumptions is met when the tests are designed for different grade levels (Reckase & Martineau, 2004).

Item Difficulty

Obviously, items written for students in upper elementary grades will be more difficult than for students in earlier grades. Item difficulty can introduce another source of multidimensionality, particularly in vertical scaling projects. In IRT, item difficulty is described by a parameter that is sometimes referred to as the *location* parameter. This parameter is symbolized b_i, which represents the difficulty, *b,* for an item, i. In the simplest, one-parameter (i.e., Rasch) IRT model, b_i is the point on the ability scale where the probability of a correct response is 0.5. An item with a higher value of b_i requires a greater ability for an examinee to have a 50% chance of getting the item correct; hence, the harder the item. On one hand, difficult items are useful to distinguish different ability levels. However, the purpose of vertical scaling is to place tests that differ in difficulty but are intended to measure similar constructs on the same scale (Kolen & Brennan, 2004).

Dimensionality can be confounded with item difficulty in several ways. For example, factors might represent items with comparable difficulty levels rather than items that measure

distinct dimensions (Ackerman et al., 2003). Reckase (1990) also observed that when the psychological dimensions are strongly confounded with the difficulty of the test items, the data might appear to be unidimensional. Unidimensionality is assumed because there is little variation in the probability of correct response on the items measuring other dimensions when there is little variation in the probability of a correct response on the items measuring the first dimension.

In another possible scenario, the difficult items in a test may be measuring a different combination of skills than the easy items (Reckase, 2004). This is of particular concern with developing mathematics tests and developing vertical scales. For example, the easier items on a mathematics test assess more arithmetic problem solving or computation skills. The more difficult items tend to be a combination of domains such as solving a coordinate geometry problem using matrix algebra. As a result, differences in scores at the bottom end of the scale are more indicative of differences in computational skills while the differences in the upper portion of the scale reflect differences in skills for manipulating matrix algebra skills (Miller & Hirsh, 1992). The development of vertical scales is also affected by a broader application of this pattern. Grade 3 tests predominantly measure arithmetic skills but by Grade 8 the test emphasis shifts to more problems solving and algebraic skills. Reckase (2004) recently remarked that "it is unlikely that vertically-scaled, grade-level tests have been analyzed to discover the multivariate structure and the relationship of that structure to item difficulty, or that the creation of multiple forms takes these relationships into account" (p.118).

Building a Vertical Scale

As illustrated in Figure 2.5, there are three designs used to collect data for vertical scaling: common item design, equivalent groups design and scaling test design. The *common item design* takes advantage of the overlapping structure of elementary achievement tests. It is comprised of two parts: a block of items that are common between adjacent grades (sometimes above and below grades) and a block of appropriate grade level items. A common item design is shown in Figure 2.6. Item block *b* is the common block of items representing material that overlaps in grade 3 and grade 4 (e.g., adding single digit numbers). Item block *c* links grade 5 to grade 4 and is linked to grade 3 through the grade 4 level using a linking chain. A similar process is used to link the grade 6, 7 and 8 levels to the grade 3 level (which is considered the base level for this example). Note that any grade level can be chosen as the base level and the links would go up/down from the base level. Since the common item design is considered the easiest and most commonly used design and was used to collect the original data that this study is based upon, descriptions of the other data collection designs will be omitted here but can be found in Crocker and Algina (1986) and Kolen and Brennan (2004).

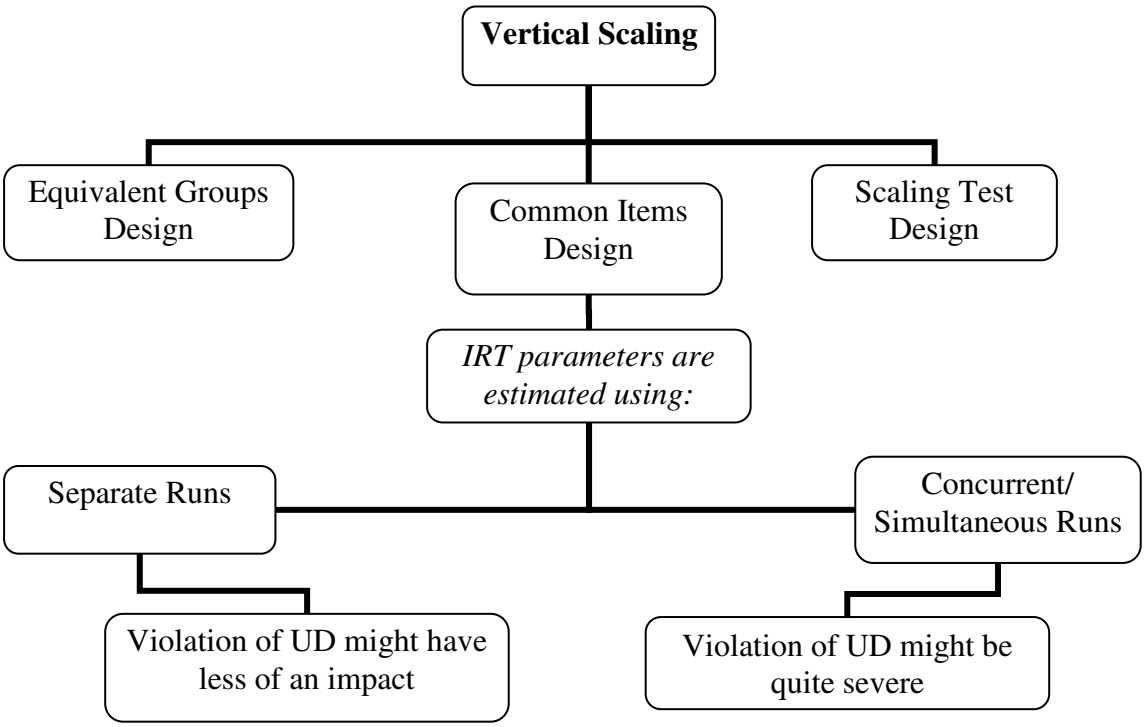


Figure 2.5. Methods and Options for Vertical Scaling

Student Grade Level	Item Block							
		a	*b*	*c*	*d*	*e*	*f*	*g*
	3	√	√					
	4		√	√				
	5			√	√			
	6				√	√		
	7					√	√	
	8						√	√

Figure 2.6. Illustration of a common item design.

The possibility that tests that differ substantially in difficulty might also differ in the traits they measure despite having similar content is an issue that must be considered during in vertical scaling procedures (Crocker & Algina, 1986). These differences in difficulty, content and traits could be possible sources of dimensionality and create unintended multidimensionality. Under the common item design, IRT parameters are estimated either using separate computer runs or concurrent/ simultaneous computer runs (see Figure 2.5). Multidimensionality could have an impact on IRT parameter estimations. According to Kolen and Brennan (2004), minor violations of unidimensionality are possible with separate runs while more severe violations could result during concurrent runs.

Consider first the case of separate runs where IRT parameters are estimated separately at each grade with only a small set of adjacent-grade items included for linking purposes (as in the common items linking design). Violations of unidimensionality tend to be minor and might have less impact since the bulk of the test material is grade appropriate. Nonetheless, above and below grade level items that are present for linking purposes could

create potential sources of multidimensionality. Although it is hoped that students retain content material from the previous year, students perform better on material that was more recently taught so fourth graders responding to third grade items/content (on a grade 4 form) might perform differently than had they responded to the third grade items during the third grade. Above grade items, like fifth grade items on a fourth grade form, present the issue of item difficulty and content coverage. In other words, a fourth grade student might find the fifth grade items more difficult because they have not reached or been exposed to deeper level of fifth grade content. The presence of off grade level items could affect the dimensional structure of the test. One purpose of this study is to investigate the presence of off-items on the dimensionality of a grade level form.

The second method of obtaining estimates of IRT parameters from common item design is concurrent computer runs. Basically, all the data, regardless of grade, are examined simultaneously. The violation of unidimensionality might be quite severe with concurrent estimation. This is of particular concern given a complex content such as mathematics where the curriculum reflects changes in the mathematical strand emphases from grade to grade. "If the curriculum content, and consequently the tests content, change dramatically from grade to grade, a single common dimension is unlikely to be attainable" (Chin et al., 2006, p. 2). However, there has been little research done to investigate potential changes to the dimensional structure of mathematics achievement tests across grades. Given the content, the strands, the changing emphasis from more computational skills to problem solving, this remains an important but unaddressed issue. Another purpose of this study is to explore this potential invariance of the mathematics achievement structure.

Summary

Any factor that influences an examinee's score on a test, other than the intended latent variable threatens the assumption of unidimensionality. Mathematics proficiency stems from a multifaceted content domain. It also presents possible sources of multidimensionality due to the diverse content strands, problem types and formats for assessing conceptual knowledge and the changing curricula emphases over grades. The developmental level at which various cognitive skills are mastered is an important educational issue and necessitates constructing a developmental, vertical scale to place tests that differ in difficulty but measuring similar constructs on the same scale. Appropriate content and item difficulty are significant dilemmas in vertical linking. Tests that measure different dimensions of a content domain must be viewed judiciously in any linkage project. Unidimensionality should never be assumed but should always be verified (Ackerman, 1994). Therefore, procedures for assessing the correct number of interpretable dimensions are a critical element in the test development process.

Procedures for Assessing Dimensionality

Investigating the dimensional structure of a test can be done in many different ways. The goal of this section is to present a brief introduction to some of the more popular procedures that are available for the empirical assessment of test structure. An overview of the procedures is presented in Table 2.1. Note that the table has two categorizations of methods: parametric and nonparametric. Assessment methods can be viewed as members of two different families, one based on parametric models and the other consisting of nonparametric methods. Parametric methods will be discussed first, followed by the nonparametric and then a comparison of the approaches. Note that the many of the procedures (parametric and nonparametric) are based on local item independence as shown in Figure 2.7.

Overview

The difference between parametric and nonparametric methods is the specification of the item response function. In IRT, the probability of success on item *i* is usually presented as $P_i(\theta)$. This function is known as the *item response function* (IRF). This function has also been called the *item characteristic curve* (ICC) and *trace line* (van der Linden & Hambleton, 1997). Parametric methods assume a particular parametric model for the IRF. Nonparametric methods assume only that the IRF is monotonic.

As previously discussed in this chapter, conditional independence in IRT is described by local item independence and the evaluation of the number of latent traits needed to

maintain local independence provides a powerful tool to assess test dimensionality. Therefore, many of the programs used to assess dimensionality are grounded in one of the three forms of local item independence: strict, weak and essential (see Figure 2.1). Recall that the strong local independence (SLI) requires that the items are completely independent once the vector of latent trait(s) is accounted for. The weaker form of local item independence, (WLI), only requires that the covariance between item pairs once the latent trait(s) have been accounted for is zero. Note that procedures using SLI and WLI require IRT model parameters be estimated (i.e., are parametric methods) and both types of procedures will yield goodness-of-fit indices and residual covariances. The weakest form of local item independence is based on the principle of essential item independence (EI). Procedures utilizing EI (i.e., the nonparametric approaches) do not require IRT model parameter estimation and are looking for a dominant factor (in tests designed to be unidimensional) or the possible presence of other dimensions (in tests intended to be multidimensional).

Parametric Methods

The goal of parametric modeling is to provide a parsimonious and quantitative description of data structure. Parametric methods comprise several approaches: classical factor analytic, item factor analytic, IRT, or some combination of the last two. The classical factor analytic approaches and programs refer to the traditional, linear factor analysis of correlation matrices. The item factor analytic perspective is an extension of classical factor analysis. It uses a nonlinear relationship between the probability of a correct examinee response and one or more examinee latent factors or abilities. In this regard, item factor

analysis models are equivalent to MIRT models (McDonald & Mok, 1995). Additional IRT perspectives include analyzing the local item dependencies and using principal components to assess the residuals of data fitted with the Rasch model.

Parametric Methods: Linear Factor Analysis

Historically, test dimensionality has been investigated using linear factor analysis methods. As mentioned earlier, the objective of factor analysis is to uncover the structure that produces the correlations in test data. Factor analysis assumes that correlation among items arises because of their dependency on one or more of the same factors. The influence of the factors on test items is measured by factor loadings. *Factor loadings* are equivalent to regression coefficients, representing the influence of a factor (independent variable) on an item (dependent variable). The numerical value of a factor loading indicates the strength of the influence of the factor on the items (i.e., higher values signify stronger influences, and lower values indicate less influence or no relationship). Typically, statistical applications used for performing factor analyses begin with Pearson product-moment correlations (or covariances) among the variables. The Pearson product-moment correlation coefficient is generally applied in situations where the relationship between two variables is approximately linear and both variables are measured on a continuous scale. Carroll (1945) observed that the Pearson product-moment coefficient tends to decrease as the variability in item difficulty increases. This may produce spurious difficulty factors.

Table 2.1. *Statistical Procedures for Assessing Test Dimensionality with Dichotomous Items*

Approach	Software	Comments
Parametric Methods		
Linear Factor Analysis	Mplus, TESTFACT, SPSS	Analysis of Pearson or tetrachoric correlations
Item Factor Analysis *	NOHARM *	Nonlinear item factor analysis; uses conditional item covariances
Item Factor Analysis	CHIDIM	Based on output from NOHARM
Item Factor Analysis	TESTFACT	Full-information item factor analysis
Local Item Dependencies	IRTNEW	Five different indices of LID are included in the IRTNEW software package.
Principal Components *	WINSTEPS *	The Rasch model is assumed.
Nonparametric Methods		
Assessment of Essential Dimensionality *	DIMTEST *	Tests the null hypothesis that the essential dimensionality of a test is one
Conditional Covariance *	DETECT *	Index of multidimensionality; exploratory only
Hierarchical Cluster Analysis	HCA/CCPPROX	Conducts a cluster analysis of test item proximities to identify dimensionally homogenous clusters of items

* These approaches (and corresponding programs) will be used in this study and therefore more in-depth information about the programs is provided in Appendix A. The rationale for choosing the programs is presented in Chapter 3.

Note the procedures are suitable for a test with a relatively large number of dichotomous items and a relatively large sample size (greater than 1,000).

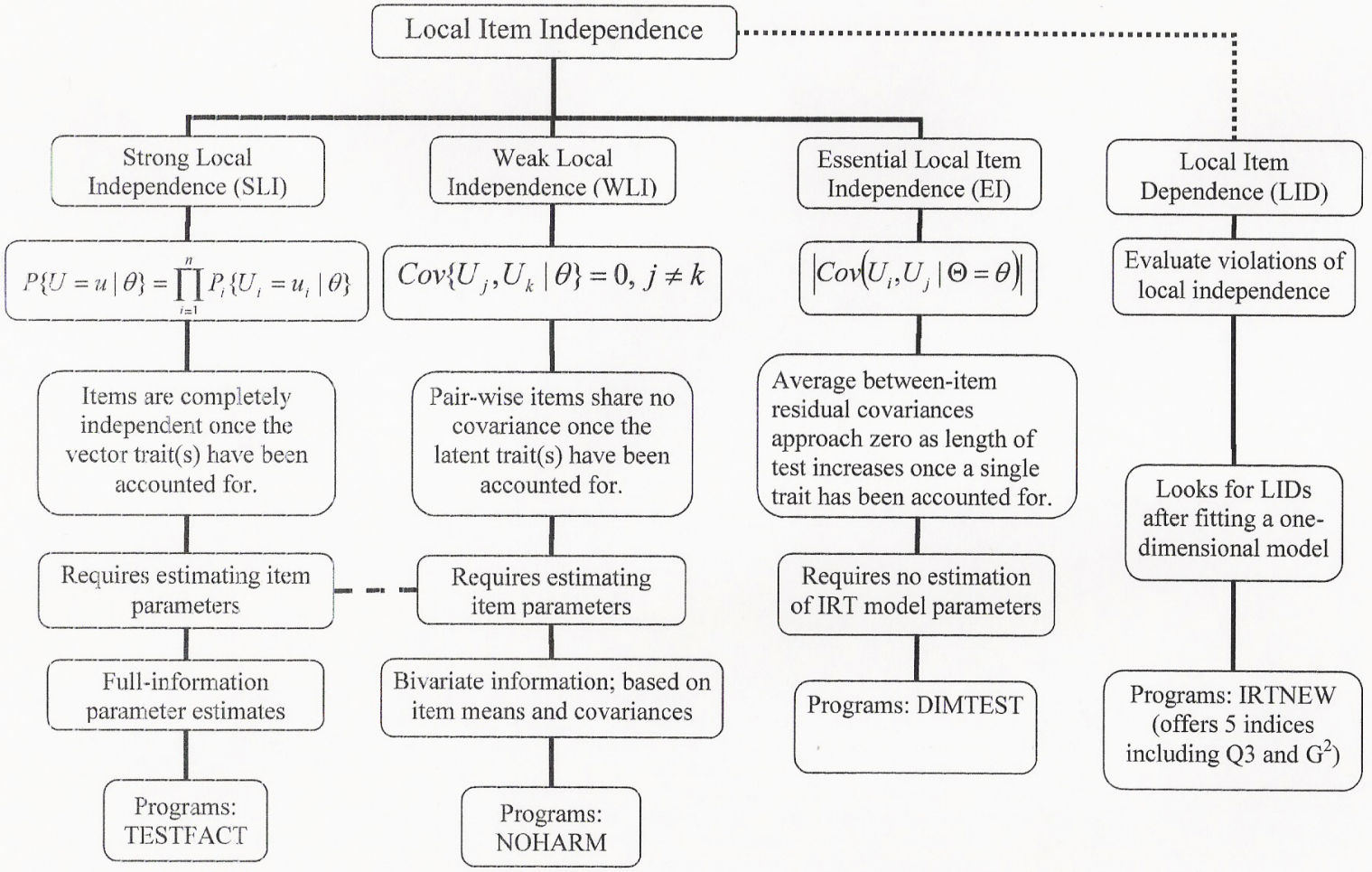


Figure 2.7. Procedures for Assessing Dimensionality using Local Item Independence

To avoid some of the difficulties produced by using Pearson correlations, tetrachoric correlations have been used instead. A *tetrachoric correlation* is another type of correlation coefficient and is applicable when both variables are dichotomies that are assumed to represent underlying bivariate normal distributions, as might be the case when a dichotomous test item is used to measure some dimension of achievement. Several methods of factor analysis based on tetrachoric correlations are available in the Mplus program (L. K. Muthen & Muthen, 1998) and TESTFACT (Bock, Gibbons, Schilling, & Muraki, 1999). One method, the unweighted least squares (ULS) exploratory factor analysis option, has been found to perform well in large-scale applications (Knol & Berger, 1991). Another method, a robust weighted least squares (WLS) procedure, has been proposed as a confirmatory approach for dichotomous variables (B. Muthen, 1993). Confirmatory approaches require hypotheses to guide the model selection and suggest factors or groupings. For example, in a confirmatory factor analysis of a mathematics assessment, content areas (algebra, geometry, etc.), test specifications, skills categories, or item types are obvious factors or groups.

The use of tetrachoric correlations presents several problems. First, tetrachoric matrices for item-level data are often not positive definite (i.e., the matrices are not invertible and therefore are problematic in some factor modeling equations). Second, tetrachoric correlations are difficult to compute when the correlation values approach the extremes (± 1). Third, tetrachoric correlations estimate a correlation based on hypothesized normal variables when, in fact, only binary scores were observed, and normality thus may be an invalid assumption. Fourth, it has been found that linear factor analyses using tetrachoric correlations indicate more factors than are actually present in the data (Hambleton & Rovinelli, 1986; Nandakumar, 1994). Finally, there is no standard approach in factor

analytic theory for determining the number of meaningful factors. This has caused Reckase, Carlson and Ackerman (1985) to conclude that "under fairly common conditions, factor analysis of tetrachoric correlations does not recover the underlying structure of dichotomous data" (p. 1).

As another alternative to factor analysis using tetrachoric correlations, some researchers have turned to item factor analytic, IRT-based and nonparametric methods to assess test structure. The following sections describe these approaches.

Parametric Methods: Item Factor Analysis

Item factor analysis models have been developed as extensions of the classical linear factor analysis. There are two types of item factor analysis that will be discussed in this section: nonlinear and full-information. The nonlinear model is basically a modified linear factor analysis. Both approaches use summary information (i.e., proportions for nonlinear models and correlations in linear models) to model the relationship between the item response and the latent trait(s), however the full-information model uses all the information present in the data to estimate model parameters.

Nonlinear item factor analysis. Several research studies have concluded that the relationship between item performance and underlying latent ability is nonlinear (Ackerman et al., 2003; Hattie, 1984). Nonlinearity can result in a mismatch between model and data. A nonlinear item factor model is a combination of the classical linear factor model with a nonlinear component expressing the probability of a correct answer to an item as a function of an associated latent response variable for the item. In other words, a nonlinear model is

similar in many respects to its linear counterpart with one major exception, the relationship of the observed responses to the underlying trait. As the names imply, the linear factor model assumes this relationship to be linear, and the nonlinear model assumes a nonlinear relationship. It accounts for the nonlinear relationships among the variables by using higher order polynomials in the factor model (e.g., quadratic and cubic terms). As described by Nandakumar, "factor models with linear and quadratic terms were able to fit the data better than models with just linear terms" (1994, p. 32). Nonlinear factor item analysis is similar to linear factor analysis in that it operates on the correlation matrix. Nonlinear factor analysis methods are directly related to MIRT procedures.

Research has been inclusive about the success of nonlinear factor models to accurately determine test structure. Hattie, Krakowski, Rogers, and Swaminathan (1996) found that nonlinear factor models were not as effective in discriminating between unidimensional and multidimensional data sets as their linear counterparts. However Hambleton and Rovinelli (1986), using simulated data, concluded that linear factor analysis overestimated the number of underlying dimensions while nonlinear factor analysis led to correct determination of the item dimensionality. NOHARM and CHIDIM are two programs based upon nonlinear factor analysis.

NOHARM program. A program that utilizes nonlinear factor analysis methods and is commonly employed by the measurement community is NOHARM (*Normal Ogive by Harmonic Analysis Robust Method*). McDonald (1967, 1981) developed the NOHARM methodology, and the NOHARM software program was developed by Fraser and McDonald (1988). Instead of using tetrachoric correlations, NOHARM minimizes the unweighted

least-squares (ULS) difference between observed values (proportions of item pairs that are passed) and expectations (based on a third-degree polynomial function) (Stone & Yeh, 2006). It computes the residual covariances of the items after fitting a model—the user specifies the number of dimensions—and calculates the root mean square of the covariances as an overall measure of misfit of the model to the data. In other words, the residual matrix offers an indication of how well the principle of local independence has been satisfied given the prescribed model. More information about the NOHARM method and program is provided in Appendix A.

Recent research by Tate (2003) found that NOHARM does of good job of identifying the presence of multidimensional data and of recovering the intended factor structure except in the cases where item discrimination parameter values are 1.5 or greater, which are considered to be very large. In a recent study using a Monte Carlo simulation, Finch (2006) compared the factor recovery performance for Varimax and Promax methods of rotation using NOHARM. For his study, Finch used Varimax, a common type of orthogonal rotation, and Promax, a common type of oblique rotation with NOHARM. The results suggested the two approaches were equally able to recover the underlying factor structure, regardless of the factor correlations, although the oblique method was better able to identify the presence of a simple structure.

CHIDIM program. Another nonlinear factor analysis approach to assessing dimensionality is the CHIDIM program (De Champlain & Tang, 1997). It is an extension of the NOHARM method and program; in fact, CHIDIM requires the observed and residual matrices computed in NOHARM as input. Gessaroli and De Champlain (1996) proposed

that an approximate χ^2 statistic based on McDonald's nonlinear factor analytic model could be used to test for the number of dimensions underlying the responses to a set of items. In the context of assessing the results of nonlinear factor analysis, the χ^2 statistic tests the null hypothesis that the off-diagonal elements in a matrix of residual correlations are equal to zero. In other words, CHIDIM is an approximate chi-square test of the fit of an estimated NOHARM model to assess test dimensionality.

Parametric Methods: Full-information Item Factor Analysis

To avoid spurious difficulty factors and other problems associated with factor analysis of correlation coefficients, Bock, Gibbons, and Muraki (1988) proposed another method of nonlinear factor analysis, called *full-information item factor analysis*, to assess test structure,. Full-information factor analysis is a technique based on multidimensional item response theory models. Because it is directly based in item response theory, it uses the data frequencies of all distinct item response vectors and does not require calculation of inter-item correlation coefficients. Stated differently, full-information factor analysis is based upon the concept that all the information available from the entire response matrix is used rather than just the covariance or correlation matrix.

TESTFACT is a full-information-based method that maximizes the likelihood of the factor model parameters give the observed pattern of correct and incorrect item responses (Stone & Yeh, 2006). In other words, TESTFACT utilizes the information from the individual responses directly rather than relying on summary statistics, namely the covariances or correlations, like in NOHARM. The TESTFACT model uses a linear common factor model to relate unobservable response process variables to underlying

factors, and uses a normal ogive IRT model to relate observed item performance to each item's underlying item response process variable (Dorans & Lawrence, 1999). The TESTFACT program (Bock et al., 1999) uses marginal maximum likelihood estimates to provide "full-information" estimates. Tate (2003) found that TESTFACT performed well in recovering data structure in exploratory and in limited bifactor confirmatory approaches. In another study, Schaeffer and Kingston (1988) used TESTFACT to examine the factor structure of the GRE General Test and to appraise the extent to which an analytical factor could be identified that was distinguishable from verbal and quantitative factors. Implications of the results questioned the utility of including analytical reasoning and logical reasoning in the same total GRE score.

Parametric Methods: Local Item Dependencies

In addition to both linear and nonlinear factor analytic approaches to assessing test dimensionality, there are also parametric methods based in IRT. Recall that the dimensional structure of a test can be defined in terms of conditional independence and more specifically in terms of local item independence in IRT. Violations of local independence are also referred to as *local item dependences (LID),* and procedures have been developed to detect these violations.

IRTNEW program. Several of procedures used to detect LID are bundled together in a software package entitled, IRTNEW. IRTNEW (Chen, 1993) provides five different indices of LID and are parametric in the sense that the conditioning is based on the unidimensional IRT model. One of the indices reported in IRTNEW is Yen's (1984) Q_3

index. Q_3 is the correlation over examinees of the residuals for an item pair, where the residual for each item and examinee is the difference between the item response (0 or 1) and the expected probability of correct response to the item for the examinee. The other four indices yielded by IRTNEW were presented and studied by Chen and Thissen (1997). These indices are based on the tables of the observed and expected frequencies of correct and incorrect responses for the item pairs. Two of the unsigned indices are provided by the Pearson's χ^2 statistic and the Likelihood Ratio G^2 statistic; both are distributed approximately as χ^2. The two signed indices, the standardized coefficient difference (or ϕ index) and the standardized log-odds ratio difference (or LOR index) are measures of association between item pairs and expected to be distributed normally with mean 0 and variance of 1.

Chen and Thissen found that the four indices appeared to be sensitive in detecting local dependence (i.e., multidimensionality) among items. However, when compared to the Q_3, the four indices were somewhat less powerful in detecting local dependence caused by the underlying factor structure, but the indices were equally as powerful as the Q_3 in detecting local dependence related to situational behavior, such as students omitting items at the end of a long test due to time constraints. The problem of inflation of family-wise error rate complicates the use of the measures to test for all item pairs (i.e. the omnibus null hypothesis that all conditional associations are zero) and is therefore not practical for assessing an entire test form. However, IRTNEW can be useful in at least two ways: testing a relatively small number of selected item pairs and as an exploratory search for any problematic item pairs by identifying outliers in the distribution of all conditional associations (Tate, 2002).

Parametric Methods: Principal Components Analysis

Another parametric technique for assessing dimensionality is principal components analysis (PCA). The objective of PCA is similar to factor analysis: determine the latent structure underlying a set of variables. However, PCA and factor analysis are not the same or suitable substitutes (MacCallum, 2004). While the differences between and long-standing controversies involving these methods are beyond the scope this study, it suffices to say that PCA analyzes variance and factor analysis analyzes covariance (Tabachnick & Fidell, 2001). To assess the structure of the data, principal component analysis (PCA) applies a model then examines the residuals for any identifiable structure. It extracts the maximum variance from the data set with each component in an attempt to explain the variance. All the variance is distributed to components, including error and unique variance for each observed variable. The first principal component is the linear combination of observed variables that maximally separates examinees by maximizing the variance of their component scores; the first component extracts the most variance. Using the residual correlations, other components are considered. The process is repeated until there is no identifiable pattern remaining in the residuals. Note that for unidimensionality, one component would explain the most variance with little or no distinguished pattern in the residuals (i.e., random noise).

WINSTEPS. According to Smith, Jr., "the use of linear factor analytic models are not appropriate methods for assessing the unidimensionality requirement of Rasch models as these methods assume a normal distribution of the data, whereas Rasch models make no such assumption" (2004, p. 577). Therefore, PCA is widely used as a method for assessing

dimensionality when assuming the Rasch model. This is operationalized in the WINSTEPS program as follows.

WINSTEPS applies a Rasch model and uses PCA to analyze the residual correlation matrix. Note that the Rasch model constructs a one dimensional measurement system regardless of the dimensionality of the data (Linacre, 1998) and, therefore, if the Rasch model fits the data well, then all the information in the data would be explained by the single latent variable and there would be no pattern in the residuals. In other words, the Rasch dimension is hypothesized to be the first dimension and explains most of the variance. Using PCA, WINSTEPS looks for other contrasts that explain remaining variation. Structure in the residuals indicates a possible second dimension. WINSTEPS will reiterate the PCA process again except this second time it is looking for possible patterns in the residuals after the Rasch dimension and the first factor have been applied. This is repeated, if necessary, until no further patterns in the residuals are found. However, preferably one contrast suffices, indicating a good fit of the unidimensional Rasch model.

Nonparametric Methods

Nonparametric approaches to measure test dimensionality were motivated by (1) in some cases, failure of parametric IRT models; and (2) utility of nonparametric methods in situations with small number of items and examinees (Tate, 2003). Nonparametric methods assume only that the item response function (IRF) is monotonic and, therefore, offer the freedom from dependence on highly-prescriptive assumed models by parametric approaches. That is, nonparametric models do not use IRT models and, therefore, do not have to estimate model parameters or be constrained by model specificity. By using a nonparametric method,

one does not confound lack of model fit by a particular unidimensional parametric family of models when working with potentially multidimensional data (Stout, 2002).

The three nonparametric methods described in the following sections are based on local item dependencies (also referred to as conditional item covariances), with the conditioning based on a single test score. These differ from the IRTNEW approach mentioned previously in two ways. First, each is based on a nonparametric computation of conditional item covariances. That is, for each item pair, students are separated into groups with respect to their number correct score on the remaining test items. The covariance of the responses of the two items is computed for each group, and the final conditional item covariance is computed as a weighted or unweighted average of the group values. Second, each of the three nonparametric methods provides a global treatment of all conditional covariances for a given test rather than the simple review of all pair-wise indices employed in IRTNEW. According to Stout et al. (1996), each of the three methods addresses a different aspect of test structure but "together they provide an almost complete summary of the test's dimensional characteristics" (p. 351).

Nonparametric Methods: DIMTEST

DIMTEST (Stout, 1987) is a software program for testing the IRT assumption of local independence for a set of items. Specifically, it is testing the EI form of local independence (see Figure 2.1) that states the average between-item residual covariances after fitting a one-factor model approaches zero as the test length increases. Recall that EI considers the presence of a dominant trait that is so strong that examinee trait levels are unaffected by the presence of smaller, nuisance factors and influences. DIMTEST uses

Stout's *T* statistic for a nonparametric test of unidimensionality. The *T* statistic is used to test the null hypothesis that the essential dimensionality of a set of items is 1. Further description and information about the DIMTEST program, methodology and Stout's *T* statistic are provided in Appendix A.

Overall research studies have reported positively on the ability of DIMTEST to correctly identify unidimensionality/multidimensionality of simulated test data (Tate, 2003). DIMTEST is able to discriminate between unidimensional and two-dimensional tests for a variety of simulated data when the correlation between abilities was as high as .7 (Nandakumar & Stout, 1993; Stout, 1987). Additionally, Nandakumar (1991) found DIMTEST is able to assess essential dimensionality in the possible presence of several minor dimensions; the statistical procedure has good power and functions as a hypothesis test of whether the essential dimensionality is one or exceeds one. However, (Kirisci, Hsu, & Yu, 2001) reported that DIMTEST is sensitive to the methods used to generate multidimensional data and performs poorly with partially compensatory data. A study by Hattie, Krakowski, Rogers and Swaminathan (1996) showed that DIMTEST is not appropriate when the tetrachoric correlation matrix used to identify the AT items is nonpositive definite or when the underlying multidimensional model is not compensatory. The study also reported that the T-statistic was not monotonically related to the underlying dimensionality and, therefore, should not be used as a general index of magnitude of dimensionality (Hattie et al., 1996; Perkhounkova & Dunbar, 1999). It has also been reported that DIMTEST may be more effective when used with larger samples of examinees, has low power for short tests, and is influenced by the number of items in a cluster assessing one trait (van Abswoude, van der Ark, & Sijtsma, 2004). Other concerns have been raised about the performance of

DIMTEST with an increasing mismatch between item difficulties and the ability distribution (Seraphine, 2000).

Nonparametric Methods: DETECT

Zhang (1996) developed a theory of conditional covariances that purported the expected conditional item pair covariances to be highly informative about the dimensional complexity of the latent space required to produce local independence. That is, conditional item covariances provide information about the dimensional complexity of test data. Note that a unidimensional test is the simplest latent structure possible. Therefore, the goal of the Dimensionality Evaluation to Enumerate Contributing Traits (DETECT) index and program (Kim, 1994; Stout et al., 1996; Zhang & Stout, 1999) is to estimate the extent of multidimensionality in the structure underlying test data. It is an exploratory nonparametric dimensionality assessment procedure that searches through various partitions of the test items to maximize the DETECT index. This index is created by computing all item covariances after conditioning on the examinees' scores using the remaining items. For any partition, the DETECT index is defined in terms of the average of all of the signed (+1 if the items are in the same cluster of P, -1 otherwise) conditional item covariances. In other words, each covariance for an item pair in the same partition is multiplied by positive one, whereas each covariance for a pair spanning two different partitions is multiplied by negative one. When the index is maximized, it represents the degree of dimensionality present in a partition. The maximization of the DETECT index also produces item clusters that offer insight into the nature of the test structure. Additional details and explanations of DETECT are presented in Appendix A.

Nonparametric Methods: Hierarchical Cluster Analysis

Cluster analysis, also called data segmentation, is a procedure for grouping or segmenting a collection of objects (e.g., items) into subsets or "clusters", such that those objects within each cluster are more closely related to one another than objects assigned to different clusters. This clustering is based on the notion of degree of similarity (or dissimilarity) between the individual objects being clustered. One major method of clustering is hierarchical clustering. In *hierarchical clustering* a series of partitions takes place, which may run from a single cluster containing all objects to *n* clusters each containing a single object. The former is referred to as *divisive* methods, and the latter procedure is an *agglomerative* method. Agglomerative hierarchical clustering starts with every single object in a single cluster. Then, in each successive iteration, it agglomerates (merges) the closest pair of clusters by satisfying some similarity criteria (in this case the correlation coefficient), until all of the data are in one cluster. In other words, items are clustered into progressively larger groups deemed to be dimensionally homogenous starting with each item constituting its own cluster and concluding with all items in one cluster.

HCA/CCPROX. To assess test dimensionality, two programs have been developed to jointly conduct agglomerative hierarchical cluster analysis, HCA and CCPROX (Roussos, 1992). The core of the approach is a new item-pair conditional covariance-based proximity measure. The resulting cluster analysis progressively forms item clusters based on item/cluster proximity; the clusters of items formed early on in the analysis help to identify those items with the strongest local item dependencies. In simulation studies, when

approximate simple structure holds, the procedure can correctly partition the test into dimensionally homogenous item clusters for very high correlations between latent dimensions (Roussos et al., 1998).

Comparison of Methods

In addition to the previously mentioned studies of individual programs, several studies have been conducted to compare the different approaches to assessing dimensionality of a set of item responses (Hambleton & Rovinelli, 1986; Hattie, 1985; Hattie et al., 1996; Nandakumar, 1994; Tate, 2003). The results have been largely inclusive. Hattie (1985) considered over 80 indices and found that many of the indices based on reliability, component analysis, and linear and nonlinear factor analysis were ineffective in determining the underlying structure of the simulation data particularly when the factors were intercorrelated. However, Hattie concluded that methods based on McDonald's NOHARM could be used as a decision criterion and was a reasonable first step. In contrast, Tate (2003) concluded the following from his recent, exhaustive comparison study:

> For the most part, all methods performed reasonably well over a relatively wide range of conditions. The several exceptions to this outcome occurred when the test data departed significantly from the assumptions or inherent limitations associated with a method, for example, when guessing was present but not allowed in the analysis or when the multidimensional test structure was nonsimple but the goal of the method was to estimate the amount of multidimensional simple structure. (p. 159)

Thus, while it seems that the measurement community cannot agree on either a standard for measuring dimensionality or how the different methods compare, the research does agree on

one thing—correlated factors (i.e., one displaying a non-simple structure) complicate the determination of dimensionality.

Other researchers have used several procedures for assessing dimensionality and compare the results across the selected methods. These procedures can produce two possible results: (1) the methods confirm one another or (2) the methods offer different conclusions which the researcher must resolve. For example, Gierl, Tan, and Want (2005) used both parametric (NOHARM) and nonparametric methods (DIMTEST and DETECT) to identify content and cognitive dimensions on the mathematics and critical reading sections of the SAT. Their methodology extended previous SAT studies which only drew on one procedure (factor analysis) by comparing the results of the different approaches. The comparison yielded similar results that allowed the researchers to conclude that there is a "multidimensional basis for test score inferences on the mathematics and critical reading sections of the SAT" (p. 26). However, in a recent article describing the use of MIRT, Ackerman, Gierl, and Walker (2003) employed three nonparametric methods (DIMTEST, DETECT, and HCA) to illustrate a systematic approach for investigating the dimensionality of test data and found conflicting analyses leaving the researcher to resolve the different results. Additionally, the authors remarked that assessment procedures should be viewed cautiously as "these procedures are only tools" (p.39) that have yielded promising results in simulation studies but produced relatively few published studies using real test data, and should be used in conjunction with substantive judgment (i.e., procedures involving content expert reviewers and/or psychological perspectives).

Summary

An important step in any test development process is to assess the dimensionality of the instrument. Because there is no consensus or standard within the measurement community regarding a procedure for assessing dimensionality, the developer has several tools with which to accomplish this task. One family of methods assumes a parametric approach. Linear and item factor analyses, full-information analysis, and detecting local item dependencies are among the parametric methods. Another family of methods is nonparametric approaches. These methods are based on conditional item covariances (conditioned on a single test score) and do not require a specific model. Regardless of approach, all procedures, parametric or nonparametric, are working with different forms of the same basis of dimensionality--conditional independence of item scores. Interestingly, comparison studies of the different methods have not yielded any one discernable, superior approach; they all seem to work reasonably well. So researchers continue to wrestle with the question of how to best assess dimensionality just as they continue to debate the consequences of dimensionality, which will be considered in the next section.

Consequences of Violations of Dimensionality Assumptions

The previous sections discussed sources of dimensionality and different approaches for assessing the dimensionality of test data. Although there is not one acceptable method for assessing dimensionality, researchers do appear to agree that educational achievement data are often multidimensional. However, many large scale testing programs use unidimensional models which raises the question: what (if any) are the consequences of using multidimensional data with a model that assumes unidimensional data? Much research has

been devoted to investigating the consequences, but the results have been inconsistent, in part due to the disagreement within the measurement community about what constitutes test dimensionality, appropriate evidence for concluding multidimensionality exists, and the seriousness of violating assumptions of unidimensionality. The next section considers the consequences of violations of dimensionality assumptions in three key areas: 1) parameter estimation, 2) vertical scaling, and 3) gathering validity evidence. Because discussion of these three key areas involves both IRT and MIRT, a brief introduction of these models and basic concepts is provided before discussing the literature.

Introduction and Basic Models of IRT and MIRT

The purpose of this section is to present the basic principles of IRT and MIRT models in order to facilitate the review of research on the consequences of violating dimensionality assumptions of the models. It is not meant to capture the history, development and technical depth of IRT and MIRT. There are many book-length treatments, as well as chapters, that discuss the concepts of IRT and MIRT in great detail (Crocker & Algina, 1986; Embretson & Reise, 2000; Hambleton, 1993; Hambleton & Swaminathan, 1985; Hambleton et al., 1991; Smith Jr. & Smith, 2004; Thissen & Wainer, 2001; van der Linden & Hambleton, 1997).

IRT Models

Unidimensional IRT models the relationship between a person's ability and responses to test items. At the center of the theory is a mathematical model of how examinees, at different ability levels for the trait, are expected to respond to a given item. A variety of

mathematical forms have been suggested as models for this relationship, but, by far, the normal and logistic ogive are preferred by theoreticians (Bejar, 1980). IRT rests on two basic postulates: (a) the performance of an examinee on a test can be predicted (or explained) by a set of factors called traits, latent traits, or abilities represented by the symbol theta (θ); and (b) the relationship between examinees' item performance and the set of traits underlying item performance can be described by a monotonically increasing function called an *item characteristic function* or *item characteristic curve* (ICC). In most applications of IRT, the ICC has the shape of an S-shaped curve, with ability level (i.e., θ) plotted on the x axis and the probability of a correct response plotted on the y axis. Examinees with higher values on the trait will have higher probabilities of answering the item correctly than examinees with lower levels of the trait. The height of the curve above any given value of θ represents the proportion of examinees at that ability level who can answer the item correctly. The equation for a normal ogive ICC is typically written as

$$P_i(\theta) = \int_{-\infty}^{w} f(z)\,dz$$

where $P_i(\theta)$ is the proportion of examinees with latent ability θ who answer item i correctly. The expression on the right side is the cumulative normal ogive. It means that the area between $-\infty$ and w under the normal ogive must be calculated. The quantity w is a real number and is determined by the equation

$$w = a(\theta - b)$$

The values of the a and b parameters will vary over items on a test and are denoted as a_i and b_i where the subscript i corresponds to item i on a test. The a_i parameter is called the item discrimination parameter. It is proportional to the slope of the ICC at the point b_i on the

ability scale. As noted earlier, the b_i is referred to as the item difficulty parameter. It is the point on the ability scale where the probability of a correct response is 0.5.

Although early ICC functions utilized the normal ogive, it has been replaced by three logistic models which require simpler computations. The cumulative logistic distribution function has the general form

$$P_i(\theta) = \frac{e^x}{1 + e^x}$$

where e is the base of the natural logarithm and x is a variable that takes on different values depending on the model; the models differ in the number of item parameters used. The *one-parameter logistic model* (1PL) is given by

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

It is one of the most widely used IRT models. An equivalent 1PL model is the Rasch model. From one perspective, Rasch models are special cases of IRT models, although another perspective asserts that Rasch models stem from a distinct paradigm that is model-driven and not data-driven (Andrich, 2004). That is, rather than using the data/items to determine the appropriate model, the Rasch model is assumed and then one finds items that fit the model. An advantage of the 1PL/Rasch model is that the raw score is a sufficient statistic for estimating ability.

Rewriting the equation for the 1PL model and including the a_i and b_i parameters as well as a scaling factor required to approximate the normal ogive yields a logistic function. The *two-parameter logistic model* (2PL) is given by:

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

Similarly the *three-parameter logistic model* (3PL) is given by:

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

Note the addition of additional c_i parameter in the model. This is referred to as the pseudo-chance-level parameter and often inaccurately referred to as the pseudo-guessing parameter.

MIRT Models

Multidimensional item response theory (MIRT) is similar to IRT in that it is also modeling the interaction between a person and a test item. The biggest difference between MIRT and IRT is that rather than using a single ability (θ) to describe a person, MIRT describes the characteristics of the person using a vector of variables representing abilities or hypothetical constructs (Reckase, 1997). In other words, rather than having one distinct θ or ability construct, MIRT models can accommodate any number of constructs or a composite of the constructs. Within a unidimensional framework, an item discriminates, to varying degrees, among all levels of the underlying trait, although there is a range in which the discrimination is optimal. In a multidimensional framework, an item has the capability of distinguishing among levels of many composites, but optimally among levels of just one composite trait (Ackerman, 1996). The goal of dimensionality assessment in MIRT is to identify this composite of abilities or constructs.

There are two types of MIRT models used to describe dichotomously scored item response data—compensatory and noncompensatory. In terms of probability of a correct response, compensatory models allow high ability on one dimension to compensate for low ability on another dimension while noncompensatory models do not permit compensation

among dimensions. Presently, no computer software programs exist to estimate noncompensatory model parameters (Ackerman, 1994; 1996) so only compensatory MIRT models will be discussed. The following is the expression for an *m*-dimensional compensatory multidimensional model:

$$P_{ij}(\theta_{ik}) = c_j + \frac{(1 - c_j)}{1 + e^{-1.7\sum_{k=1}^{m} a_{jk}(\theta_{ik} - b_{jk})}}$$

where θ_{ik} is the ability parameter for person *i* for dimension *k*,

a_{jk} is the discrimination parameter for item *j* for dimension *k*,

b_{jk} is the difficulty parameter for item *j* for dimension *k*, and

c_j is the pseudo chance-level (i.e., "guessing") parameter for item *j*.

Comparing IRT and MIRT

While MIRT addresses the multidimensional nature of many educational data sets, its applicability and usefulness to testing programs is questionable. In a recent presentation, Martineau and his colleagues (Martineau, Subedi, Ward, Li, & Diao, 2006) suggest that while truly unidimensional data are rarely observed in educational achievement tests, MIRT is not a useful choice either: despite its 30+ years of research, MIRT has seen negligible application in educational achievement testing contexts; it is impractical due to its relatively higher cost and availability of software; replication is a problem; and difficulties exist in interpretability of MIRT results. Specifically, the complexity and the uncertainty about the definition of a dimension in MIRT models has caused some researchers to contend that

MIRT cannot be applied in practical testing situations (Kirisci et al., 2001; Luecht & Miller, 1992).

In contrast, when tests measure one latent trait, a single score can be assigned to each examinee, and the interpretation of test performance is unambiguous. According to Reckase and Ackerman (1986), "the more complex the function of the skills required to relate the skills to the total score on the test, the more difficult is the task of interpreting the score" (p. 2). In addition, if a test is truly multidimensional, it becomes impossible to rank order test-takers without implicitly or explicitly weighting the dimensions (Ackerman, 1992).

Although computing power can address some of the complexities of MIRT parameter estimation, the greatest impediments of applying MIRT models are score interpretation (van Abswoude et al., 2004) and difficulty of linking tests that measure composite abilities. When a single score is used to summarize or represent test performance, unidimensional linking procedures are typically used (Hirsh, 1989). If however, multidimensional models are used, then the single score represents a composite of abilities; and thus linking equivalent forms of a test (i.e., equating) or different forms of a test across grades (i.e., vertical scaling) is not feasible. Further, it appears that the development of test items (and test forms) that measure the same composite of abilities is currently an unproven goal.

Given the difficulty of multidimensional score interpretation and multidimensional linking, many researchers have asserted that unidimensional tests may be preferred where possible. Stout (1987) believes that there are at least three reasons why it is essential that a test be unidimensional:

> First, it is often vital that a test that purports to measure the level of a certain ability is in reality not significant [sic] contaminated by varying levels of one or more other abilities displayed by examinees taking the test...second, it is essential that a test designed to be used in the measurement of individual

> differences must in fact measure a unified 'trait'...finally, unidimensionality must be (at least approximately) satisfied if much of the standard response theory methodology is to be trusted as valid. (p. 589-590)

Hattie (1985) also contended that one of the most critical and basic assumptions of measurement theory is that a set of items forming an instrument all measure just one thing in common. This conclusion was based on two assertions. For one, like Stout (1987), Hattie recognized that the unidimensionality assumption provided the basis of most mathematical measurement models. The second argument was more substantive: "To make psychological sense when relating variables, ordering persons on some attribute, forming groups on the basis of some variable, or making comments about individual differences, the variable must be unidimensional" (p. 139). Since most IRT measurement models and score reports assume unidimensionality, researchers have thus focused their attention on investigating the consequences of violating assumptions of unidimensionality, particularly the possible effects on item parameter and ability estimates.

Consequences of Violations on Parameter Estimation

Investigating the dimensional structure of a test should be an important step in any test development process. Violations of dimensionality assumptions can potentially affect model parameter estimates including person parameter estimates used for score reporting. Importantly, the effects of inaccurate item or person parameter estimates are not merely statistical issues, but have substantial practical relevance. For example, many of the state administered end of the year and/or graduation tests are considered to be high-stakes tests, meaning that important consequences are attached to students test performance. Additionally, meeting the requirements of NCLB is also tied to student scores. Therefore,

error in calculating student scores could have tremendous impact on students, teachers, school districts and administrators. However, the research of the effect on model parameter estimates stemming from the use of unidimensional methods with multidimensional data has been largely inconclusive.

When a unidimensional IRT model fits the test data, several desirable features are obtained. In IRT, item and ability parameters are said to be *sample invariant*. This means that examinee ability estimates are not dependent on the sample of test items used to estimate them, and item parameters are not dependent on the ability distribution of the examinees (i.e., item parameter estimates obtained in different groups of examinees will be the same except for measurement error). Another desirable feature of IRT is that estimates of standard errors for individual ability estimates, rather than a single estimate of error for all examinees, can be obtained. However, these properties are only realized when the given data fit the IRT model and assumptions. Misfit of the IRT model through the violation of the assumption of unidimensionality can result in underestimation of the standard errors for the examinee ability parameter estimates (Wainer & Wang, 2001). It can also underestimate the effect size of the difference in means of two grade level tests in a linking project (Reckase & Li, 2006).

Early research on the effects of violating the assumption of unidimensionality on parameter estimation was conducted by Ansley and Forsyth (1985). In a study using data generated to fit a noncompensatory two-dimensional MIRT model, the authors concluded that "violations of the assumption of unidimensionality do have an effect on the parameter estimation for the modified three-parameter logistic model" (p. 47). More specifically, it was found that the $\hat{a}$ values were best considered as averages of the true a_1 and a_2 values where a_1 and a_2 are discrimination parameters for the two dimensions. The $\hat{b}$ values appeared to be

overestimates of the true b_1 values only rather than a combination of b_1 and b_2. (Note that the c parameter was set equal to .2 for each item.) The estimated θ values were most correlated to the averages of the true θ values. These conclusions were similar to the results of a study done by Stocking and Eignor (1986). In that study, simulations using a 3PL model were conducted to study the impact of multidimensionality in the data on preequating. The invariance property of true item parameters suggests that is it possible to equate a test before it is actually administered as long as the true item parameters are known. This procedure is called *preequating* and is used heavily in adaptive testing situations. Multidimensionality was generated in the simulated response data. Responses for some items were generated using a certain true ability and responses to other items were generated using a second true ability. In other words, an examinee would need one ability to respond to some items and another ability to respond to other items on the test. This would result in a multidimensional test. The authors concluded that "the introduction of a particular kind of multidimensionality in the data can have a large impact of estimation precision when the IRT model is unidimensional" (p. 40).

Contrary to these findings, Embretson and Reise (2000) observed the following in a recent review of dimensionality research:

> The effect on parameter estimation of small departures from unidimensionality remains undemonstrated. In fact, some research indicates that IRT model parameter estimation is fairly robust to minor violations of unidimensionality, especially if the latent-trait (factors) are highly correlated or if secondary dimensions are relatively small. (p. 231)

The former observations were primarily based upon the work of Reckase (1979) and Drasgrow and Parsons (1983). Reckase (1979) evaluated the 1PL and 3PL models for use with both real and simulated multivariate data. He concluded the 1PL and 3PL models estimate different abilities when independent factors are present; the 3PL model estimates

one factor while the 1PL model estimates the sum of the factors. Both models estimate the first principal component, when it large relative to the other factors. Although item calibration results will be unstable, Reckase also found accurate ability estimates can be obtained from the models in the presence of a dominant or potent first factor even when the first factor accounts for less than 10% of the test variance (although item calibration results will be unstable). For acceptable calibration, the first factor should account for at least 20% of the test variance. For tests with several equally potent dimensions, the one-parameter ability estimates were best considered as the sum or average of the abilities required for each dimension (Ansley & Forsyth, 1985).

Drasgow and Parsons (1983) used several simulated item pools that ranged from the truly unidimensional to an inconsequential (i.e., very weak) general latent trait. The item pools were used to simulate varying degress of prepotency (i.e. domination) that is required by the software program LOGIST (Wingersky, Barton, & Lord, 1982) in order to recover the general latent trait and not be drawn to a latent trait underlying a cluster of items. Drasgow and Parsons concluded the following:

> If a single dominant latent trait is not sufficiently prepotent [influential], the results presented here clearly show that a unidimensional model is inadequate. However, it is important to note that unidimensional models *do* provide a good description of multidimensional data sets when the dominant latent trait is sufficiently prepotent. (p. 198)

In other words, according to Reckase (1979) and Drasgow and Parsons (1983), the influence of multidimensionality depends on the degree to which there is a dominant latent trait.

Exploring Inconsistent Findings

There are several hypotheses about why the studies of the impact of unidimensionality assumption violations have been inconclusive. One possible reason for discrepancy is differences in the definition of a dimension. More specifically, by one definition, a test is considered to be unidimensional in that it measures only one skill or ability while, from another perspective, a test would be considered to be unidimensional in that the test measures the same composite of several, possibly correlated skills. The impact of multidimensionality appears to be strongly related to the correlations of the dimensions or skills. When the multidimensionality is due almost entirely to the planned test structure and the associated component abilities are at least moderately correlated, typical uses of the test scores may be robust to the violation of the assumption of unidimensionality (Tate, 2002). If correlations vary greatly or are very low ($r \leq .4$), then MIRT should be used. Furthermore "when the component abilities measured by the test are weakly correlated or when there are strong construct-irrelevant factors, the consequences of the violation [of unidimensionality] may be serious for test validity, fairness and score comparability" (Tate, 2002, p. 192).

A second possible reason for inconclusive results of violation studies is the effect of the estimation method used by the estimation program (Kirisci, Hsu, & Yu, 2001). LOGIST and WINSTEPS (Linacre, 2005) use joint maximum likelihood to estimate model parameters. Other common software programs, such as BILOG (Mislevy & Bock, 1984), MULTILOG (Thissen, 1991) and XCALIBRE (Assessment Systems Corporation, 1996), use marginal maximum likelihood procedures. When different estimation programs are used by different studies, inconsistencies in resulting estimated parameters can be expected (Baker, 1987).

Kirisci, Hsu, and Yu (2001) recently investigated the sensitivity of 2PL model parameter estimates derived from BILOG, MULTILOG and XCALIBRE when the unidimensionality assumption was violated and the underlying θ distribution was not multivariate normal. Data with three dimensions were simulated and then six experimental conditions were constructed for each program: two types of dimensionality (one-dimensional, three correlated dimensions) crossed with three θ distributions (normal, skewed, or peaked). They discovered there was an interaction between program and dimensionality indicating that the robustness of the conclusions about the unidimensionality assumption was a function of the estimation program. Although BILOG produced the smallest root mean square error overall, MULTILOG and XCALIBRE showed less variance in model parameter estimation due to the violation of unidimensionality.

A third possible reason for the inconclusive findings about the effects of multidimensionality is the differences in how the multidimensional data were generated for the studies (Kirisci et al., 2001). Multidimensional data are typically generated by one of two methods: a factor-analytic approach or a MIRT model. Studies employing multidimensional data generated by a MIRT model tend to show that violation of the unidimensionality assumption can seriously affect item parameter estimation (Doody, 1985; Kirisci et al., 2001; Reckase, 1987). However, studies using multidimensional data generated by a factor-analytic approach tend to show that a unidimensional IRT model is robust to moderate degrees of multidimensionality. For example, Ansley and Forsyth (1985) criticized both the Reckase (1979) and Drasgow and Parsons (1983) studies for the factor analysis model used to assess dimensionality as well as to generate simulated data. According to Ansley and Forsyth, the relationship between the factor analysis model and the logistic model is not precisely defined,

thus generating data to fit a factor analysis model "might not yield a completely clear picture of the effects of using a unidimensional logistic model with multidimensional data" (p. 39).

Consequences of Violations on Vertical Scaling

In addition to potential influences on model parameters and validity evidence, violations of unidimensionality assumptions can also have an impact on vertical scaling. According to Bogan and Yen (1983), the assumption of unidimensionality frequently does not appear to be met in many testing situations, yet the need for vertical scaling exists. As mentioned previously, vertical scales are used in situations when different examinees are measured with different, purposefully non-equivalent instruments to create vertical or developmental scales for achievement tests (Kolen & Brennan, 2004). Therefore, the issue becomes the robustness of vertical scaling versus the violation of the unidimensionality assumption. This next section describes two groups of research investigating vertical scaling and assumptions of dimensionality. The first group focuses on the 3PL model and the second group utilizes the Rasch model.

Three Parameter Model (3PL) and Vertical Scaling

A research study done by Bogan and Yen (1983) examined how robust three-parameter vertical scaling is to violation of the unidimensionality assumption underlying the linking. Four two-trait data configurations and one unidimensional data configuration were simulated for three differences in mean difficulty between two tests to be vertical scaled. The accuracy of the vertical link was examined by comparing the estimated thetas for all

simulees on the easier test with the estimated mean for all simulees on the harder test. The comparison considered the standardized root mean squared differences (SRMSD), standardized mean differences (SMD), the ratio of the standard deviations, and correlations. The results indicated that vertical linkings for the multidimensional configurations were as good as those for the unidimensional configuration when either the correlations or SRMSD were examined. However, investigation of the SMD, which is an estimate of the overall bias between the estimated thetas for all simulees on the easier test with the estimated thetas for all simulees on the harder test, showed conflicting results. The multidimensional tests usually had less accurate linking (i.e., greater differences in estimated thetas) than the unidimensional tests, particularly when the tests to be linked differed in difficulty.

In another study of the robustness of item and ability parameter estimation to unidimensionality violation using the 3PL, Doody (1985) simulated 10 two-trait and one unidimensional test configurations for a 30-item test and 6,000 simulees. The results of this study indicated that "the poorest item parameter estimates occur for the situation in which one test is unidimensional and one is multidimensional" (p. 64).

In a recent study, Chin, Kim and Nering (2006) examined the following five design and statistical factors on vertical scaling: (1) separation of grade overlap (i.e., ability differences between grades), (2) number of grade levels/forms to be vertically linked, (3) length of the common item block, (4) difficulty range of the common items, and (5) parameter estimation methods. A Monte Carlo simulation was used to study the influence of these five factors on IRT vertical scaling. The test was assumed to have a fixed length of 60 items for each grade level and 10,000 examinees were generated for each grade level group. The five factors were fully crossed, resulting in 108 study conditions. For each condition,

where estimation convergence was obtained, the estimated grade level ability means and within grade level variances were calculated in order to examine whether artificial grade-to-grade growth and/or grade-to-grade variability patterns exist. Root mean square errors (RMSE) and bias were calculated separately for model parameters in order to assess estimation accuracy. The results were inconclusive: the five factor levels interacted with each other and therefore, the authors were unable to make specific conclusions. Concurrent calibration was generally less affected by common block design decisions than separate calibration and it generally "produced satisfying results until the range of the latent trait continuum involved is 'overly' stretched" (p. 15). However, the authors observed a potential limitation of concurrent calibration: concurrent calibration might yield unstable results or possibly not obtain a convergent estimate when the number of forms to be linked is large and the ability/difficulty spectrum extends. In other words, concurrent calibration may be problematic when there are a large number of groups to be vertically scaled and/or when the ability differences among groups are large.

Rasch Model and Vertical Linking

In addition to studies of the 3PL model, the consequences of violating the assumption of unidimensionality on vertical scaling have also been investigated when the Rasch model is employed. Previous research about the application of the Rasch model to vertical scaling has yielded mixed findings. Several studies concluded that the Rasch model does not appear to work well for vertical linking of multiple choice tests (Holmes, 1982; Loyd & Hoover, 1980; Skaggs & Lissitz, 1986; Slinde & Linn, 1978, 1979). Holmes (1982) observed that the unsatisfactory results of vertical linkings with the Rasch model may be due in part to the lack

of fit of items to the model. Holmes also found that the Rasch model does not provide a satisfactory means of vertical equating across the entire ability range. Skaggs and Lissitz (1986) demonstrated that for vertical linking, the Rasch model was less robust to violations of its assumptions than for horizontal equating.

Slinde and Linn (1978) conducted a vertical linking investigation with the Rasch model to link subtests of a 36-item mathematics achievement test. The basis of the subtests was the difficulty level of the items: one subtest contained difficult items and one included easy items. The subtests were administered to a sample of incoming college freshmen that was divided into high, medium and low ability groups based on their performance on the easy subtest. Item difficulty estimates were calculated on the entire test (i.e., combined set of difficult and easy items) for both the high and the low ability groups. The middle group was not used to obtain estimates but instead was used the compare the equivalence of the easy and difficult subtests when the ability estimates were derived from the high and the low groups. Results indicated that an examinee of middle ability would do better to take a more difficult test when the estimates are obtained from the high group, and would do better to take an easier test when the estimates are obtained from a low ability group. More generally, this finding means that the item parameters are not invariant across groups and may depend on the sample used to obtain the estimates. This is not a desirable feature for two tests that are to be vertically linked, and casts doubt on the application of Rasch models for vertical linking. The study was criticized for the division of the sample population into ability groups based on performance on the easy subtest for the same test used for vertical linking (Gustafsson, 1979). However, Slinde and Linn (1979) conducted a reanalysis with another data set and the results generally supported the earlier study.

Loyd and Hoover (1980) also explored the application of the Rasch model to the vertical scaling of levels of the *Iowa Tests of Basic Skills* (ITBS) Math Computations Test. Unlike the Slide and Linn (1978) study mentioned previously, Loyd and Hoover did not form ability groups based on test performance, but used grade groupings that would be more typical to a practical application of vertical scaling. They used three levels of the ITBS and three samples of pupils from the 6^{th} through the 8^{th} grades. Linkings were conducted across adjacent grades (e.g., linking grade 6 and 7 items) and non-adjacent levels (e.g., grade 6 compared to grade 8) using the three samples as separate calibrations groups. If the Rasch model were appropriate for vertically scaling this test, then the calibrations should be consistent in determining ability estimates for the separate ability groups and the vertical scaling of both adjacent and nonadjacent levels should be invariant with respect to groups. Their results supported the Slinde and Linn studies (1978, 1979) in that the linking between any two levels was influenced by the group upon which the linking was based. In other words, the vertical scaling of these levels of the mathematics computation test was not independent of the ability group used in the linking.

Loyd and Hoover (Loyd & Hoover, 1980) observed that a student who takes an easier (lower) level test and has his/her scores linked to a more difficult level will have a higher resultant score than when the linking is based on the higher ability group. Similarly, for students who take a more difficult (higher) level of the test and then have his/her scores linked to an easier (lower) level, the resulting scores will be more favorable (i.e., higher) when the linking is based on the lower ability group. In looking for causes of the inadequate Rasch vertical scaling, Loyd and Hoover considered violations of the underlying assumptions including the assumption of unidimensionality. They then looked at the potential influence

of the mathematical content using a principal axis factor analysis of the total item pool. Specifically, there were concerns that curriculum content across grade levels, particularly in mathematics, might not represent a unidimensional scale. The analysis showed that more than one factor was present in the total item pool. The authors did not explicitly attempt to define these factors, but suggested that mathematics performance may be differentially dependent upon school curriculum. For example, a sixth grade student may have the same probability as an eighth grade student of answering items related to working with whole numbers but the probabilities of answering correctly on a subset of items involving decimals or fractions could differ considerably for the same two students.

Overall, the results of these studies suggest that vertical scaling is most sensible when the instruments to be linked can be viewed as representing a developmental continuum for a subject area and where true scores on the two instruments are functionally related (Harris, 1991). While Chin, Kim and Nering (2006) cautioned that "vertical scaling probably should never be carried out unless there is a satisfying demonstration of unidimensionality across grade levels involved" (p.17), it was also been shown that a dominant factor or highly correlated factors can possibly satisfy the assumption of unidimensionality. The issue of test content across test levels appears to be a critical one and deserves further exploration (Skaggs & Lissitz, 1981).

Consequences of Violations on Validity Evidence

A measurement issue that subsumes considerations of parameter estimation, vertical scaling, and all other dimensionality concerns is that of validity. Test scores are typically used to draw inferences about examinee behavior in situations beyond the testing session

(Crocker & Algina, 1986). Responsible use of test scores requires that the test user be able to justify the inferences (Crocker & Algina, 1986). Such justification requires reliability of the test scores and validity evidence. *Reliability* refers to the consistency of measurements when the testing procedure is repeated on a population of individuals or groups (AERA, APA, NCME, 1999). *Validation* can be viewed as developing a scientifically sound argument to support the intended interpretation of test scores and their relevance to the proposed use (AERA et al., 1999). Validity is the degree to which all the accumulated evidence supports the intended interpretation of the test scores for the proposed purpose. According to the *Standards for Educational and Psychological Testing* (AERA et al., 1999), validity evidence can be based on examination of test content, response processes, internal structure, relations to other variables, and consequences of testing. Validity evidence based on internal structure of a test indicates the degree to which the relationships among test items and test components conform to the construct on which the proposed test score interpretations are based. That is, analyses of internal structure suggest how well the test items and components associate with the construct of interest. Recall that test dimensionality is defined as the minimum number of abilities measured by a set of test items. Therefore, understanding the dimensional structure of a test can provide insight and validity evidence based on the internal structure of a test. Coefficient alpha, factor analysis and other methods are typically included as evidence for the internal structure of an instrument. In addition, use of a theory, such as IRT, that posits unidimensionality can yield evidence of unidimensionality. Negligence in rigorous assessment of test dimensionality may result in construct validity problems; different scores on the test may represent different substantive interpretations in terms of the constructs that underlie them (Jang & Roussos, in press). When unidimensionality is intended but not

realized, the validity of test scores based on IRT models is in question (Messick, 1993) particularly when a single total score is reported. Although MIRT models offer a method to represent multidimensional data, as previously mentioned a significant concern with MIRT models is understanding the meaning of the dimensions which can be a combination of underlying traits thus making score interpretations, linking, and the gathering of validity evidence extremely difficult.

Summary

The consequences of violating the assumption of unidimensionality have important implications on many facets of the test development process including parameter estimation, vertical scaling, and gathering validity evidence. Test items and student performance are analyzed using mathematical models such as IRT or MIRT which assume certain a dimensional structure. Therefore, misdiagnosis or misrepresentation of the dimensional structure can impact model parameter estimates including person ability estimates (i.e., student scores). The dimensional structure of a test is also used to provide one type of validity evidence based upon the internal structure of a test. Because validity refers to the degree to which evidence and theory support the interpretations of test scores, it is a fundamental consideration in test development. Modeling student growth and adequately yearly progress have also become important considerations in a testing program. This has necessitated the use of vertical scales that model the mathematical developmental continuum across grades and content standards. While previous research on the consequences of violating the assumption of unidimensionality has been inconclusive due to differences about

definitions and evidence of dimensionality, it seems that eliminating any error is advantageous with so many high-stakes associated with the test results.

Research Questions

In order to determine item characteristics, student performance levels and link test forms, many of the commonly used item response models assume unidimensionality. However, educational achievement tests often assess more than one skill or ability either intentionally or unintentionally. The nature of mathematical content and understanding introduces several potential, unintentional sources of dimensionality. The issue of dimensionality of mathematical content across grade levels is a critical concern to the validity and development of both end of year achievement tests and monitoring student growth over time. Previous research studies on test dimensionality have been largely based on simulated data sets and, in particular, have been inconclusive about the number of dimension(s) assessed in a typical, statewide-mathematics achievement test. From this review of literature, it appears that it is not accurate to assume that statewide tests measure a well-defined unidimensional construct and more empirical evidence is required to substantiate this claim.

Question 1: To what extent is the dimensional structure of typical statewide mathematics achievement tests aligned to NCTM content strands invariant across grades 3-8?

Question 2: Does the presence of linking items (below and/or above grade level) change the dimensional structure of typical statewide mathematics achievement tests aligned to NCTM content strands?

Question 3: Do different approaches to assessing dimensionality lead to different conclusions about the dimensional structure of typical statewide mathematics achievement tests aligned to NCTM content strands?

CHAPTER 3

METHODOLOGY

Unidimensionality is assumed in many commonly-used IRT models. The presence of unidimensionality can simplify score interpretation, strengthen vertical scaling projects, and provide validity evidence regarding the internal structure of a test that is purported and developed to measure a single construct such as mathematics achievement. However, unintentional sources of multidimensionality may be present particularly in a complex subject like mathematics where mathematics achievement tests typically measure a combination of several subdomains such as algebra and geometry. Therefore, assessment and consideration of the test structure is a critical part of the development and evaluation of large-scale tests. This importance is widely recognized and yet "the question of what to take as evidence of multidimensionality has yet to be answered in a way that is widely accepted by IRT analysts of different theoretical backgrounds" (Traub & Lam, 1985, p. 27).

One purpose of this study was to assess potential changes in the dimensionality and factor structure of mathematics achievement tests aligned to NCTM standards across Grades 3-8. A second purpose was to assess dimensional structure in those tests when out-of-level items are included in the tests for the purpose of establishing a cross-grade (i.e., vertical) scale. Finally, a third purpose was to provide a methodological comparison of methods for assessing the dimensionality and factor structure of these mathematics achievement tests. Given the lack of consensus in the measurement community, four widely applied methods will be used to investigate structure of the data. The four methods will be used to examine the

stability of the test structure across Grades 3-8. The specific research questions to be addressed include:

Question 1: To what extent is the dimensional structure of typical statewide mathematics achievement tests aligned to NCTM content strands invariant across grades 3-8?

Question 2: Does the presence of linking items (below and/or above grade level) change the dimensional structure of typical statewide mathematics achievement tests aligned to NCTM content strands?

Question 3: Do different approaches to assessing dimensionality lead to different conclusions about the dimensional structure of typical statewide mathematics achievement tests aligned to NCTM content strands?

Overview

The proposed study used data collected in February 2004 as part of a large-scale field study. Several school districts across the country agreed to participate in the study which resulted in a large and diverse sample of elementary and middle schools students. Field tests were administered at each grade level. Each field test form consisted of 30 multiple choice items which included a common block of items from out of grade level (below- and above–grade level) for vertical scaling. For example, the grade 4 form included Grade 4 (on-grade) items, and items from Grade 3 and Grade 5 (off-grade) as well. All items in each form were multiple-choice format and dichotomously scored.

Assessment of dimensionality was analyzed using four approaches. The approaches included two parametric approaches (item factor analysis and principal components) and two

nonparametric approaches (assessment of essential dimensionality and conditional covariance). Table 2.1 identifies each approach and the corresponding implementation software. The following section of this chapter describes the participants, instruments, and linking design of the mathematics achievement tests used to collect the data and the specific approaches used to assess the dimensionality of these tests.

Participants

Data were collected on a total of 9,165 students in grades 2 through 9 in 34 schools from 14 districts across six states (California, Indiana, Massachusetts, North Carolina, Utah, and Wisconsin). Table 3.1 shows the field study participation by state, district, school, and number of participating students (*n*). The participants were diverse in their geographical location as well as the size and type of community (e.g., suburban; small town, city or rural communities; and urban). Table 3.2 shows the breakdown of the number and gender of participants by grades.

Measures

Each of the mathematics achievement tests was developed in the same way including attention to content specification, item writing and review, and field testing. The content specifications required that the items be aligned with the five content strands suggested in the National Council of Teachers of Mathematics framework (NCTM, 2000) which are as follows:

1. Numbers and operations

2. Geometry
3. Algebra/Patterns and functions
4. Data analysis and probability
5. Measurement

Table 3.1 *Field study participation by state and school*

State	District	School	*n* (%)
CA	Eureka Union	Ridgeview School	316 (3.2%)
IN	Alexandria Community	Cunningham Elementary	108 (1.1%)
IN	Alexandria Community	Thurston Elementary	309 (3.1%)
IN	Brownsburg Community	Brownsburg	344 (3.5%)
MA	Dennis-Yarmouth Regional	Wixon	125 (1.3%)
NC	Durham Public Schools	Parkwood Elementary	594 (6.0%)
NC	Iredell-Statesville Public Schools	Brawley Middle	238 (2.4%)
NC	Wilkes County Schools	CB Eller	210 (2.1%)
NC	Wilkes County Schools	CC Wright	264 (2.7%)
NC	Wilkes County Schools	East Wilkes Middle	392 (4.0%)
NC	Wilkes County Schools	Fairplains	98 (1.0%)
NC	Wilkes County Schools	Millers Creek	400 (4.1%)
NC	Wilkes County Schools	Moravian Falls	143 (1.5%)
NC	Wilkes County Schools	Mt. Pleasant	144 (1.5%)
NC	Wilkes County Schools	Mtn. View	392 (4.0%)
NC	Wilkes County Schools	Mulberry	208 (2.1%)
NC	Wilkes County Schools	North Wilkes Middle	552 (5.6%)
NC	Wilkes County Schools	Roaring River	144 (1.5%)
NC	Wilkes County Schools	Traphill	85 (0.9%)
NC	Wilkes County Schools	Union	172 (1.7%)

State	District	School	*n* (%)
NC	Wilkes County Schools	Wilkesboro	269 (2.7%)
UT	Cache County Schools	Nibley Elementary	74 (0.8%)
WI	Dalaven-Darien School Dist	Darien Elementary	137 (1.4%)
WI	Dalaven-Darien School Dist	Wileman Elementary	129 (1.3%)
WI	Manitowoc Public School Dist	Jackson Elementary	232 (2.4%)
WI	Manitowoc Public School Dist	Jefferson	346 (3.5%)
WI	Manitowoc Public School Dist	Washington Jr. High School	581 (5.9%)
WI	Manitowoc Public School Dist	Wilson Jr. High School	344 (3.5%)
WI	Sch Dist of South Milwaukee	Blakewood Elementary	253 (2.6%)
WI	Sch Dist of South Milwaukee	Lakeview Elementary	202 (2.1%)
WI	Sch Dist of South Milwaukee	Luther Elementary	128 (1.3%)
WI	Sch Dist of South Milwaukee	Rawson Elementary	277 (2.8%)
WI	Sch Dist of South Milwaukee	South Milwaukee High School	920 (9.3%)
WI	Sch Dist of South Milwaukee	South Milwaukee Middle Sch	717 (7.3%)

Table 3.2 *Field study participation by grade and gender*

Grade Level	*n*	Percent Female (*n*)	Percent Male (n)
2	1,283	48.1 (562)	51.9 (606)
3	1,354	51.9 (667)	48.1 (617)
4	1,454	47.7 (644)	52.3 (705)
5	1,344	48.9 (622)	51.1 (650)
6	976	47.7 (423)	52.3 (463)
7	1,250	49.8 (618)	50.2 (622)
8	1,015	51.9 (518)	48.1 (481)
9	489	52.0 (252)	48.0 (233)

All items were written and reviewed by trained item writers who were experienced mathematics educators and item-development specialists and therefore familiar with mathematical achievement of students at various grade levels. Item writers were also trained in the development of multiple-choice items. Training included materials related to sensitivity issues as represented in the concepts of universal design and fair access (Thompson, Johnstone, & Thurlow, 2002) which emphasize equal treatment of the sexes, fair representation of minority groups, and the fair representation of and access for disabled individuals. Items were then reviewed by content and psychometric experts to ensure quality of the response options and sensitivity issues.

The linking plan called for three forms of 30 items at each grade and employed a common-item design to create the vertical scale. Therefore, some items were administered to the intended grade and were also placed on off-grade forms (above or below one grade).

Items that were placed as below-grade linking items (i.e., Grade 3 items on a Grade 4 form) were specifically chosen to represent fundamental subject matter that an on-grade student would be expected to answer correctly. That is, the below-grade items were items from the previous grade and should reflect material an on-grade student has previously learned. However, the above-grade items were chosen based on strand and could potentially affect the performance of the on-grade students due to anxiety, motivation, lack of exposure to the content, etc. In each form, more of the linking items were below-grade items (typically 2-4 items per form) than above grade items (each form contained two above-grade items). The data from the on-grade items were explored Research Question 1. Data from the both on-grade and off-grade items were utilized in Research Question 2. An illustration of the composition of a Grade 4 form is shown in Figure 3.1.

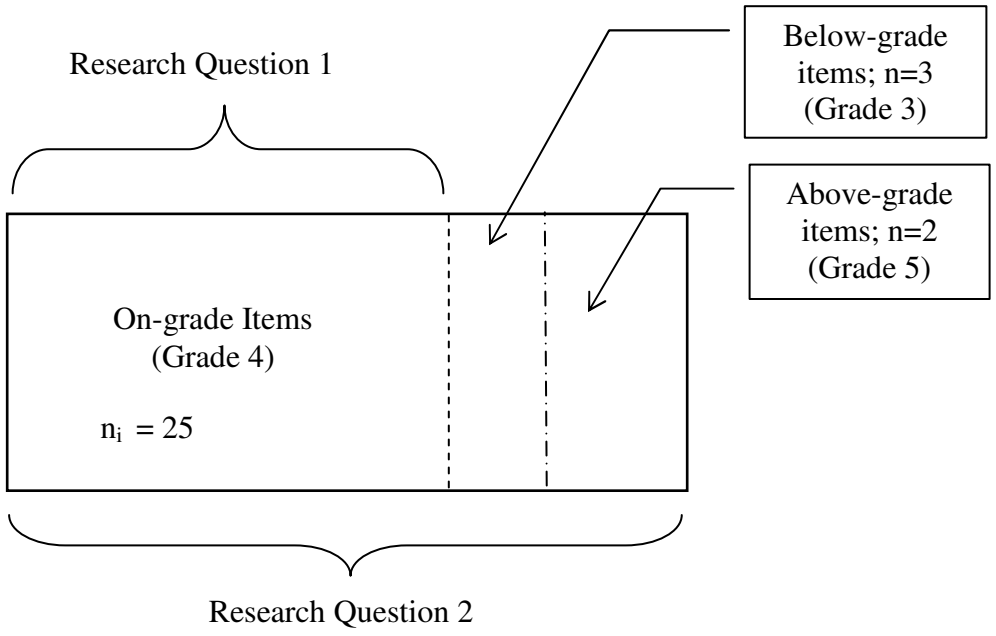


Figure 3.1 Example of a Grade 4 Form

Table 3.3 presents the number of items per strand or a "strand profile" for each grade. Notice that the number of items per strand varies from grade to grade. The Grade 3 form had eight "Numbers and Operations" items, three "Geometry" items, six "Algebra and Patterns" items, three "Data Analysis and Probability" items and six "Measurement" items. The Grade 4 form strand profile was seven, six, two, five, and five items, respectively. Many statewide mathematics achievement tests aligned to NCTM content strands have different test specifications (i.e., different number of items per strand) for each grade to reflect the changing curricular standards across the grades. For example, a state assessment might specify that 35-40% of the total test items should come from Numbers and Operations on a Grade 3 form where only 10-15% of the total test items come from Numbers and Operations on a Grade 8 form.

The last row in Table 3.3 shows the total number of on-grade items that were placed on each form. Notice that the number of on-grade items varied from grade to grade. Consider Grade 3 and Grade 4 again. Twenty-six of the 30 items on the Grade 3 form were Grade 3 items and therefore four items were off-grade level. The Grade 4 form contained 25 Grade 4 items and five items that were from Grades 3 or 5.

Table 3.3 *Number of on-grade items per strand by grade level form*

Strand	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Numbers & Operations	8	7	7	8	6	4
Geometry	3	6	4	4	4	6
Algebra & Patterns	6	2	5	4	5	8
Data Analysis & Probability	3	5	3	4	4	2
Measurement	6	5	6	6	6	4
Total	26	25	25	26	24	24

Selection of Specific Approaches and Software

In the previous chapter, many of the different ways to assess dimensionality were discussed. Researchers do not agree on a single method for assessing test dimensionality. Additionally, Embretson and Reise (2000) in their recent review and critique of dimensionality assessment suggested that:

> "...researchers should now be starting to move away from reporting heuristic indices such as 'variance accounted for by the first factor' or 'ratio of first to second eigenvalue' and start implementing the new procedures that tackle these issues in a much more sophisticated manner. …[W]e recommend more application of Stout's procedure for determining essential dimensionality and/or application of appropriate techniques such are found in TESTFACT, POLYFACT, NOHARM, and LISCOMP." (p. 245)

Using this suggestion and similar research done by Gierl, Tan, and Wang (2005) to identify content and cognitive dimensions on the SAT, this study used the following methods and software packages (see Table 2.1): item factor analysis (NOHARM), principal component analysis (WINSTEPS), assessment of essential dimensionality (DIMTEST), and exploring the conditional covariances (DETECT). All four approaches have been shown to be effective indices of dimensional structure. Recall that DIMTEST and DETECT are both nonparametric procedures. They are popular because they avoid the strong parametric modeling assumption while still adhering to the fundamental principles of item response theory (Roussos et al., 1998). Recall also that NOHARM uses item factor analysis. There are several advantages to a factor analytic approach to multidimensional data structure. One, the multidimensional model allows the correlation between underlying factors to be estimated. Two, the common factor parameterization allows factor analysis interpretative conventions to aid in the interpretation of multidimensional solutions (Gierl et al., 2005).

While the first three programs must be run separately to assess dimensionality and then require another program to estimate item and person parameters, assessing dimensionality in Rasch measurement via WINSTEPS is slightly different. As Smith, Jr., has indicated, "the use of linear factor analytic models are not appropriate methods for assessing the unidimensionality requirement of Rasch models as these methods assume a normal distribution of the data, whereas Rasch models make no such assumption" (Smith Jr., 2004, p. 577). Within the Rasch approach, the fit of the data to the unidimensionality requirement is often addressed internally using, for example, the discrepancy between the observed and the model expected responses. These discrepancies are also referred to as *residuals*. WINSTEPS is a Rasch-based computer program that utilizes principal component analysis (PCA) to assess dimensionality by looking at the residuals. However "criteria have yet to be established for when a deviation becomes a dimension so PCA is indicative, but not definitive, about secondary dimensions" (Linacre, 2005, p. 261). Therefore, several approaches or indicators of dimensionality will be considered. More information about each of the four programs, their assumptions, and the resulting output is presented in Appendix A. Additional information about the decision criteria used in the programs is presented later in this chapter.

Plan of Analysis

For this study, four approaches were used: item factor analysis, principal components, assessment of essential dimensionality and conditional covariances. These approaches offer different lenses from which to view dimensionality and have been shown to be effective indices of dimensionality in previous research (Tate, 2003). The methods are implemented in different software packages. As shown in Table 2.1, item factor analysis is implemented using the NOHARM program, principal component analysis is conducted using WINSTEPS, assessment of essential dimensionality is done using DIMTEST, and exploring the conditional covariances is conducted by DETECT. Table 3.4 summarizes the plan of analysis by research question.

Methods

DETECT can only be run in an exploratory mode and therefore it was used as an initial attempt to identify the dimensional structure of the forms. The results of DETECT also provided clusters that were helpful in understanding how the items work together (or not). Confirmatory analyses using the content strands as the hypothesized structure were then conducted using NOHARM, WINSTEPS and DIMTEST. Only on-grade item data were used for the dimensional analysis of each form (Research Question 1). That is, only the 25 items at the Grade 4 level were used to analyze the dimensional structure of the Grade 4 form. The hypothesis test of essential unidimensionality was assessed. In addition, the content strands served as an organizing principle with which confirmatory analyses were done. Strands represent potential factors or dimensions and thus confirmatory analyses were

conducted looking for five dimensions. Previous research has shown that the strands tend to be highly correlated so the exploratory results using DETECT was also considered to better understand how the items were clustering and the number of possible dimensions to consider in the confirmatory analyses.

Both on-grade and off-grade items were considered for Research Question 2. The analyses were completed using two datasets: (1) the below- and on-grade items and (2) the above-and on-grade items. Referring to the Grade 4 form example again, one set of analyses was conducted using 25 on-grade items (Grade 4) as well as two below-grade items (Grade 3). A second set of analyses were completed using 25 on-grade items (Grade 4) and three items from above-grade (Grade 5). The confirmatory analyses explored whether datasets containing on- and off-grade level items reflect the two different grades represented by the items. For the Grade 4 example, two confirmatory analyses (one for Grade 3 and 4 items, another for Grade 4 and 5 items) tested whether the construct(s) measured differs across the grade levels.

Criteria for Assessing Dimensionality

A summary of the criteria for the different programs and approaches is presented in Table 3.5. In an exploratory analysis, information about the number of dimensions found using DETECT was considered. DETECT output provided several pieces of information. One index (D_{Max}) reflects the degree of multidimensionality and another index (r_{Max}) reports on whether the data are displaying simple or complex structure. Confirmatory analyses were done using DIMTEST, NOHARM and WINSTEPS. DIMTEST calculates a T-statistic and associated *p*-value; the null hypothesis of unidimensionality will be tested at the $\alpha = .05$

level. NOHARM output includes Tanaka's Index (Tanaka, 1993) and the root mean square residual (RMSR). While there are no definitive guidelines to interpret Tanaka's index, a higher value indicates a better fit of a multidimensional model. A RMSR equal to or less than four times the reciprocal of the square root of the sample size, $RMSR \leq 4\left(\frac{1}{\sqrt{n}}\right)$, implies good model fit (Fraser & McDonald, 1988). WINSTEPS divides the variance into explained and unexplained portions. Large values of explained variance compared to smaller amounts of unexplained variance indicate a unidimensional model is fitting the data well. WINSTEPS output also includes eigenvalues.

The programs used different approaches to assess the dimensional structure and therefore different indices and results were reported. The results of each program were compared to summarize the dimensional structure of the mathematics tests. Because this was a multi-method study, "the results from these procedures vary in their ability to discern the signal of the valid skills form construct irrelevant noise leaving the researcher to resolve the different results" (Ackerman et al., 2003, p. 41). In other words, it was expected that the program assessments would most likely be different from one another so consistency in statistical methods as well as substantive judgment guided the conclusions.

Table 3.4 *Outline of Procedures by Research Question*

Research Question	Exploratory Analysis	Confirmatory Analyses
Question 1: Invariance of dimensional structure across grade levels. NOTE: For the analysis, a subset of only on-grade items will be included. Confirmatory analyses will be based upon content strands.	Method/Program • Conditional covariance/DETECT	Method/Program • Assessment of essential dimensionality/ DIMTEST • Item Factor analysis/NOHARM • Principal components/ WINSTEPS
Question 2: Effect on dimensionality of the presence of linking items. NOTE: For this analysis, grade level and linking items are included. The dimensionality of the subtests (above- and on-grade items and below- and on-grade items) will be compared to on-grade items alone.	Method/Program • Conditional covariance/DETECT	Method/Program • Assessment of essential dimensionality/ DIMTEST • Item Factor analysis/NOHARM • Principal components/ WINSTEPS

Research Question	Exploratory Analysis	Confirmatory Analyses
Question 3: Comparison of the different approaches to assessing dimensionality	The number of dimensions suggested and other information generated by the results of four approaches and programs will be compared.	

Table 3.5 *Output Results and Comments by Program*

Program	Output, Reported Indices, or Summary Statistic	Comments/Notes
DIMTEST	• T (DIMTEST test statistic) • *p-value*	• Small *p*-values indicate rejecting the null hypothesis of unidimensionality. • $\alpha = .05$
NOHARM	• Tanaka's Index (1993) • RMSR*	There are no interpretative guidelines for Tanaka's index, other than a higher value implies a better fit.
WINSTEPS	• Residual analysis • Variance explained/ unexplained • Eigenvalues	• Large values of variance explained indicates unidimensionality. • Eigenvalues < 1.4 indicate "noise" in the residuals and all factor(s) are accounted for by the model.

*Root mean square residual (RMSR)

Program	Output, Reported Indices, or Summary Statistic	Comments/Notes
DETECT	• $D(P^*)=D_{Max}$ index • r_{Max} index	Classifications suggested by (Kim, 1994): • $D_{Max} < .10$ indicates that the data can be considered unidimensional • $.10 < D_{Max} < .50$ suggests a weak amount of multidimensionality. • $.51 < D_{Max} < 1.0$ suggests a moderate amount of multidimensionality. • $D_{Max} > 1.00$ indicates a strong amount of multidimensionality • r_{Max} is an index of simple structure; $r_{Max} > .80$ suggest data display ~ simple structure while $r_{Max} < .80$ implies a complex structure. • Cluster examination

*Root mean square residual (RMSR)

Summary and Limitations

The psychometric models used in the context of many mathematics achievement tests assume a unidimensional construct is being measured. However, mathematics achievement tests reflect a complex subject that spans five content strands and skills that build from grade to grade. In order to better understand the measurement of mathematics achievement and possible sources of unintended multidimensionality, this study addressed several key questions. The first question considered the dimensional structure of mathematics achievement tests across grades. That is, does the within-grade dimensional structure of a mathematics achievement test change from a Grade 3 to Grade 8? The second question addressed potential effects of the presence of off-grade items that were included in on-grade forms for the purpose of creating a cross-grade (i.e., vertical scale). In particular, this question addressed whether out-of-level (i.e., above- and below-grade) items affect the dimensional structure of an on-grade form. The last research question was directly tied into the methods of assessing dimensionality and explored whether the different approaches to assessing dimensionality led to different conclusions regarding the dimensional structure. There are several approaches to assessing dimensionality and this study utilized four of the most commonly used methods: item factor analysis (NOHARM), principal component analysis (WINSTEPS), assessment of essential dimensionality (DIMTEST), and exploring the conditional covariances (DETECT).

There are several limitations of this study. First, this study was based upon analyses of real test data (as opposed to simulated data) and therefore the "true" underlying factor structure was unknown. Second, it is possible that the instructional and curricular emphases would result in weak factor changes across grades. Third, the mathematics content strands

might be so highly correlated that the dimensional structure could be considered essentially unidimensional even in the presence of confirmed multidimensionality. There are also several limitations associated with the field test design. For example, the data offer a limited number of available items, particularly the off-grade items on each form. The off-grade items could also affect the performance of on-grade students. While the off-grade items were specifically chosen to represent approachable material for an on-grade student, the presence of off-grade items may affect student performance due to anxiety, lack of motivation, or curricular differences.

CHAPTER 4

RESULTS

Using real test data and applying a variety of popular dimensionality assessment methods, the test structures of mathematical achievement tests were examined across Grades 3-8. Both exploratory, confirmatory or a combination of both approaches were used when appropriate. The first research question required analyses using on-grade items only. Therefore, only Grade 3 items were considered for the assessment of the Grade 3 test structure, only Grade 4 items for the Grade 4 test, etc. The second research question included off-grade level items which is typical of a vertical scale linking design. The results related to the first research question (on-grade items) are presented first, followed by those for research question two (off-grade items). The final section in this chapter offers a comparison of the different solutions and approaches as stated in research question three.

Results for Dimensional Structure across Grades

The following section presents the results of Research Question 1: the dimensional structure across mathematical achievement tests Grades 3 through 8. Each set of on-grade items were analyzed for possible sources of dimensionality related to five mathematical content strands. The analyses were also used to compare test structures across grades. The original expectation was the tests would be essentially unidimensional or would exhibit only modest amounts of multidimensionality due to the different strands.

Conditional Item Covariance and On-Grade Items

The first method used to assess potential changes in dimensional structure across the grade levels studied was an analysis of conditional item covariances. Conditional item covariances provide information about the dimensional complexity and structure of test data. The DETECT program was used to investigate conditional item covariances. The program provides three pieces of summary information that bear on test structure: (1) the DETECT Index (D_{max}) which indicates the amount of multidimensional simple structure; (2) the r_{max} index which indicates whether the data are displaying simple or complex structure; and (3) the number of clusters needed to maximize D_{max} where the number of clusters is theoretically equal to the number of dominant abilities or dimensions of the test. However, one condition must be noted about the relationship between dimensions and clusters: the number of dominant abilities measured by the test is indicated by the number of clusters *only* in the optimal partition of items for a test that is essentially multidimensional and exhibits simple structure.

The results from DETECT for on-grade items are shown in Table 4.1. The second column presents the D_{max} index, which ranged from 0.4148 to 0.6536. D_{max} values greater than 0.10 but less than 0 .50 suggest a weak amount of multidimensionality and D_{max} values greater than 0.51 and less than 1.0 suggest a moderate amount of multidimensionality (Kim, 1994). Therefore, the values obtained in these analyses indicate a weak amount of multidimensionality in Grades 3-6 and a moderate amount of multidimensionality in Grades 7 and 8.

The r_{max} index, shown in third column of Table 4.1, ranged from 0.4998 to 0.6197. An r_{max} value greater than 0.80 suggest the data display approximate simple structure while an r_{Max} value less than 0.80 implies a complex structure (Kim, 1994). The magnitude of the values of r_{max} shown in Table 4.1 generally indicates that the test forms analyzed exhibited complex structure. The last column in Table 4.1 presents the number of clusters DETECT used to calculate the D_{max} and r_{max} indices for each test. Four of the six forms exhibited five clusters (Grade 3, 5, 6, and 8) while items from Grades 4 and 7 were partitioned into four clusters.

Table 4.1 *Results of Conditional Covariance Analysis (DETECT) of On-Grade Items*

Grade	D_{max}	r_{max}	No. of Clusters
Grade 3	0.4558	0.5534	5
Grade 4	0.4905	0.6032	4
Grade 5	0.4148	0.4998	5
Grade 6	0.4550	0.5204	5
Grade 7	0.6536	0.6119	4
Grade 8	0.5631	0.6197	5

Zhang and Stout (1999) found that while the clusters partitioned by DETECT are more accurate when r_{max} is greater than 0.80 (i.e., approximate simple structure), DETECT is still very informative when approximate simple structure fails to hold. Therefore, the clusters were examined further but caution should be exercised when interpreting the cluster results. The clusters for the Grade 3 on-grade items are shown in Table 4.2. The first row displays the total number of items per cluster. The subsequent rows show the number of

items per strand in each cluster. For example, Cluster 1 consisted of 12 items (out of 26 items on the form). Four of those items were from the Numbers and Operations strand, one item from the Geometry strand, five items from the Algebra and Pattern Recognition strand, and one item each from the Data Analysis and Probability strand and the Measurement strand. Recall that each item was written to a specific content strand and the test specifications required items from all five strands. The clusters however do not match item designated strands indicating that the item clusters do not appear to be based on the content strands. For example, as can be seen in the table, the eight items that were designated as being in the Numbers and Operations content strand were identified by DETECT as failing to cluster together as intended, but were distributed across three clusters: Cluster 1, Cluster 2, and Cluster 3. The clustering of items for the other grades were similar to the clusters for Grade 3 in that item clusters did not appear to be strand-based. The results for the other grades are presented in Appendix B.

Table 4.2 *Distribution of Strand-Designated Items by Cluster and Content Strand for Grade 3*

	Distribution of Strand-Designated Items by Cluster					
Content Strand	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Total
Numbers & Operations	4	3	1	0	0	8
Geometry	1	0	0	1	1	3
Algebra & Patterns	5	0	1	0	0	6
Data Analysis & Probability	1	2	0	0	0	3
Measurement	1	2	0	2	1	6
Total Number of Items in Cluster	12	7	2	3	2	26

Assessment of Essential Dimensionality of On-Grade Items

The second method used to assess potential changes in dimensional structure across the grade levels studied was an assessment of essential dimensionality. DIMTEST uses Stout's *T* statistic for a nonparametric test of unidimensionality. The *T* statistic is used to test the null hypothesis that a set of items is essentially unidimensional. The p-values from applying confirmatory DIMTEST (based on strands) are presented in Table 4.3. To control for family-wise error rate when testing five comparisons (i.e., five content areas), the False Detection Rate (FDR) was utilized. According to Benjamini and Hochberg (1995), FDR has

higher power than the Bonferroni method, and it controls for Type I errors better than testing without adjustment than Bonferroni and other post-hoc comparison techniques. To implement the FDR method, the p-values associated with the content areas are ordered from smallest to largest values within each grade level. The smallest p-value is compared to a critical value of 1*.05/5, or .01. If the smallest p-value is larger than the critical p-value, no further comparisons are necessary and the null hypothesis of interest is retained. As can be seen in Table 4.3, the smallest p-values for Grades 3, 4, 6, 7 and 8 exceed the critical value of .01 and are therefore not significant at the .01 level. Thus, the null hypothesis of essential unidimensionality cannot be rejected for these comparisons. However, since the smallest p-value for Grade 5 is significant, then the second smallest p-value is considered. According to the FDR technique, the second smallest p-value is tested against 2*.05/5, or .02. This is not significant so no further tests are warranted. The DIMTEST results for Grade 5 test suggest that the Numbers and Operations items are dimensionally different from the other items.

In summary, when a set of items based on content strand was compared to the items on the rest of the test, the null hypothesis of essential unidimensionality could not be rejected for all strands in Grades 3, 4, 6, 7 and 8. In other words, subsets of items based on content were not dimensionally different from the remaining items suggesting that the data are essentially unidimensional. However, Grade 5 results displayed a slightly different story. The items designated as Numbers and Operations for Grade 5 suggest a potentially different dimension than the remaining Grade 5 items from the other four strands.

Table 4.3 *P-values from DIMTEST Using On-grade Items*

Content Strand	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Numbers & Operations	0.8497	0.1290	0.0660	0.1961	0.2605	0.1218
Geometry	0.2742	0.3558	0.0154	0.3299	0.6492	0.2822
Algebra and Patterns	0.1133	0.1122	0.3674	0.4419	0.0354	0.6955
Data Analysis & Probability	0.9863	0.4373	0.1655	0.8989	0.6453	0.1827
Measurement	0.1038	0.6310	0.4281	0.4253	0.0243	0.9407

Nonlinear Item Factor Analysis of On-Grade Items

The third method used to answer Research Question 1 (that is, potential changes in dimensional structure across grade levels) was an item factor analysis approach. This approach was performed using the software program NOHARM which is based on a nonlinear factor analytic approach. NOHARM computes the residual covariances of the items after fitting a model (the user specifies the number of dimensions) and calculates the root mean square of the residual covariances as an overall measure of misfit of the model to the data. In other words, the residual matrix offers an indication of how well the principle of local independence has been satisfied given the prescribed model.

Initially, a confirmatory analysis was conducted in NOHARM. The hypothesis of five dimensions (based on content strands) was tested. The results for each grade are shown in Table 4.4. The root mean square residual (RMSR) is an indicator of model fit; RMSR=0 indicates a perfect model fit and increasingly higher values indicate worse fit (Kline, 2005). The RMSR values were relatively small across the grades, ranging from 0.0089 to 0.0174,

signifying very little misfit of the data to a five-dimensional model. Tanaka's index is another fit index and it ranges from 0 to 1; while there are no specific interpretive guidelines, better fit is indicated by values closer to 1 (Tanaka, 1993). Tanaka's index was higher in Grades 3 and 4 than Grade 5-8 indicating a better fit for a 5-dimensional model in the lower grades than the higher grades.

Table 4.4. *Confirmatory Nonlinear Item Factor Analysis Results (NOHARM) for On-Grade Items (Five-Dimensions)*

Grade	RMSR	Tanaka's Index
Grade 3	0.0101	0.9568
Grade 4	0.0089	0.9556
Grade 5	0.0174	0.8950
Grade 6	0.0151	0.9098
Grade 7	0.0174	0.8931
Grade 8	0.0142	0.9159

Note. The five dimensions were based on the five mathematical content areas.

To further investigate the structure of the tests, exploratory analyses were then conducted with NOHARM to allow the number of dimensions to vary. That is, each form was analyzed in NOHARM using one, two, three, four and five dimensions in turn. The results for each grade are displayed in Table 4.5 and Table 4.6. Overall, each condition at each grade results in a small RMSR and a high Tanaka's Index, indicating a good model fit. To determine the estimated number of dimensions using the exploratory NOHARM, the percent decrease in RMSR was calculated. These results are shown in Table 4.6. For this study, the assessment of test dimensionality in an exploratory analysis was based on

consideration of the degree of improvement of model fit with increasing dimensionality of the model. Following the previous research of Tate (2003), test dimensionality was defined as the highest dimensional model that still produced an approximately 10% or greater decrease in the RMSR over the preceding model.

Interpretation of the results shown in Table 4.6 can be illustrated by considering the first row which shows the results for Grade 3. The RMSR associated with the unidimensional solution is 0.0105; when a second factor is added to the model, the RMSR is 0.0087, a difference of 0.0018, which represents a 17.1% decrease. Continuing an examination of the Grade 3 results, a third dimension is associated with a 0.0077 which results in a decrease of 11%. The addition of a fourth factor, the RMSR is 0.0070 which only decreases the RMSR by 9% so a four-factor solution is not considered. Using this 10% decrease in RMSR criteria, the estimated number of dimensions for each grade level is given in the last column. Based on interpreting the amount of decrease in the RMSR, at least one dimension is possible for the Grade 5 form, Grades 3, 4 and 6 forms could have two or three dimensions, Grade 7 items could have as many as four dimensions and the Grade 8 items exhibited at least 5 dimensions. It is important to note that many of the comparisons hovered around the 10% decrease in RMSR used as a criterion. Use of this criterion in an equivalent sample of data would likely produce different results and therefore the interpretations from these results should be considered with caution.

Table 4.5 *Exploratory Nonlinear Factor Analysis (NOHARM) Tanaka's Index for On-grade Items*

Grade	One Dimension	Two Dimensions	Three Dimensions	Four Dimensions	Five Dimensions	Change in Tanaka's Index (5D-1D)
Grade 3	0.9536	0.9681	0.9748	0.9796	0.9830	0.0294
Grade 4	0.9547	0.9661	0.9724	0.9764	0.9804	0.0257
Grade 5	0.9646	0.9710	0.9770	0.9812	0.9853	0.0207
Grade 6	0.9511	0.9622	0.9691	0.9751	0.9808	0.0297
Grade 7	0.9522	0.9644	0.9739	0.9783	0.9828	0.0306
Grade 8	0.9501	0.9662	0.9758	0.9829	0.9876	0.0375

Table 4.6 *Comparison of Exploratory Nonlinear Factor Analysis Results (NOHARM) for On-Grade Items*

	One Dimension	Two Dimensions		Three Dimensions		Four Dimensions		Five Dimensions		Change in RMSR (1D-5D)	Estimated # of Dimensions
Grade	RMSR	RMSR	% RMSR Decrease	RMSR	% RMSR Decrease	RMSR	% RMSR Decrease	RMSR	% RMSR Decrease		
Grade 3	0.0105	0.0087	17%	0.0077	11%	0.0070	9%	0.0063	9%	0.0041	2 or 3
Grade 4	0.0090	0.0078	13%	0.0071	10%	0.0065	8%	0.0059	9%	0.0031	2 or 3
Grade 5	0.0103	0.0093	10%	0.0083	11%	0.0075	10%	0.0066	11%	0.0036	1
Grade 6	0.0112	0.0098	12%	0.0089	10%	0.0080	10%	0.0070	12%	0.0042	2 or 3
Grade 7	0.0117	0.0101	14%	0.0086	14%	0.0079	9%	0.0070	11%	0.0047	4
Grade 8	0.0110	0.0090	18%	0.0076	15%	0.0064	16%	0.0055	15%	0.0055	5+

To further investigate these potential multidimensional findings, the factor loadings produced by NOHARM in an exploratory five-dimensional case were examined for patterns among the factor loadings and content strands. In addition, to the initial factor loadings NOHARM also produces rotated sets of factor loadings. Rotation is ordinarily used after factor extraction to maximize high correlations and minimize low ones. A most commonly used orthogonal rotation is varimax. The goal of varimax is to simplify factors by maximizing the variance of the loadings within factors—loadings that are high after extraction become higher after rotation and loadings that are low become lower thereby making factor interpretation easier. An orthogonal rotation was selected to explore distinct, uncorrelated dimensions that would be expected if the content strands represented different constructs or abilities. Correlated factors make interpretation of the factor loadings difficult. Furthermore, in a recent study using a Monte Carlo simulation, Finch (2006) compared the factor recovery performance for Varimax and Promax methods of rotation using NOHARM. His results suggested the two approaches were equally able to recover the underlying factor structure, regardless of the factor correlations.

The varimax rotated factor loadings for an exploratory five dimensional structure in the Grade 3 items are shown in Table 4.7. The content strand for each set of items is given in the first column. The highest factor loading for each item is in bold. For example, Item 1 loads highest on Factor 1. Items 1 through 9 are specified as items assessing Numbers and Operations. However, notice the highest factor loadings for Items 1, 7 and 8 are on Factor 1; Items 2, 3 and 9 load on Factor 2; and, Items 4 and 6 load on Factor 4. Interestingly, all items for Algebra and Patterns load on the second factor in addition to several items from the other strands. There are several items (Items 6 and 21) whose highest loadings are negative

indicating an inverse relationship to the factor. The items do not load according to the content strand as expected if multidimensionality was due to differences in skills or content specific to that strand. Grades 4- 8 NOHARM factor loadings for the 5-dimensional model are presented in Appendix C. The loadings were similar to those presented for Grade 3 in that the items do not load according to the content strands.

To simplify the factor loadings table and look for potential patterns, a summary of the number of items by content strand and factor is shown in Table 4.8. The last row of each grade level table displays the number of total items that load on each factor obtained in these analyses. The other rows in each table show the number of items that load on each factor by the intended content strand. For example, consider the portion of Table 4.8 that shows results for Grade 3. As can be seen in the last column of that table, 8 of the 26 items on the test were intended to measure the Number and Operations strand. However, as can be seen in the first row of the table, three of those items loaded on Factor 1, three items loaded on Factor 2, and two items loaded on Factor 3. Overall, the tables illustrate that, across Grades 3 through 8, the items do not tend to load according to the content strands as expected if a potential source of multidimensionality was due to differences in skills or content specific. Looking at the Algebra and Patterns strand across grades shows that while the items on Grades 3-6 tend to load on the same factor, the item loadings spread across all factors at Grades 7 and 8. Therefore the results from the nonlinear item factor analysis indicate content strands do not appear to be potential sources of multidimensionality in the test structure of mathematics achievement tests in Grades 3–8 (Research Question 1).

Table 4.7. *Nonlinear Item Factor Analysis (NOHARM) Factor Loadings for Grade 3 (i=26)*

		Varimax Rotated Factor Loadings				
Strand	Item #	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
Numbers and Operations	1	**0.415**	0.377	0.069	0.085	0.114
	2	-0.014	**0.473**	0.154	0.208	-0.087
	3	0.193	**0.495**	0.029	0.104	0.134
	4	0.234	0.278	0.012	**0.645**	-0.238
	6	0.187	0.127	-0.094	**-0.215**	0.082
	7	**0.379**	0.305	0.271	0.125	-0.115
	8	**0.324**	-0.068	-0.094	0.020	-0.041
	9	0.197	**0.434**	0.086	0.078	0.218
Geometry	10	-0.004	**0.534**	-0.133	0.042	-0.030
	11	0.175	0.309	-0.282	**0.377**	0.141
	12	0.093	0.078	0.011	0.118	**0.239**
Algebra & Patterns	14	-0.028	**0.413**	0.086	0.132	-0.017
	15	-0.013	**0.563**	-0.006	0.186	-0.098
	16	0.080	**0.173**	-0.017	0.120	0.051
	17	-0.042	**0.677**	0.090	0.122	0.294
	18	0.088	**0.663**	0.067	0.030	-0.209
	19	0.125	**0.667**	0.055	-0.173	0.108
Data Analysis & Probability	20	0.076	**0.734**	-0.151	0.120	0.230
	21	0.085	0.025	-0.001	0.037	**-0.558**
	22	**0.535**	0.106	0.052	0.297	0.176
Measurement	23	0.111	0.055	0.044	**0.315**	0.073
	24	**0.261**	-0.057	0.167	0.167	0.023
	25	**0.688**	0.207	0.394	0.096	-0.101
	26	-0.066	0.319	**0.714**	0.082	0.093
	28	0.094	-0.061	**0.302**	0.006	-0.001
	30	0.071	0.253	0.042	**0.431**	0.150

Table 4.8. *Summary of NOHARM Factor Loadings by Content Strand*

Grade 3	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Total
Numbers & Operations	3	3	0	2	0	8
Geometry	0	1	0	1	1	3
Algebra & Patterns	0	6	0	0	0	6
Data Analysis & Probability	1	1	0	0	1	3
Measurement	2	0	2	2	0	6
Total	6	11	2	5	2	26

Grade 4	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Total
Numbers & Operations	4	0	3	0	0	7
Geometry	2	1	2	1	0	6
Algebra & Patterns	0	1	0	0	1	2
Data Analysis & Probability	2	0	1	0	2	5
Measurement	3	0	0	0	2	5
Total	11	2	6	1	5	25

Grade 5	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Total
Numbers & Operations	3	0	0	3	1	7
Geometry	0	2	1	1	0	4
Algebra & Patterns	0	4	0	1	0	5
Data Analysis & Probability	0	2	1	0	0	3
Measurement	0	1	1	1	2	5
Total	3	9	3	6	3	24

Grade 6	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Total
Numbers & Operations	3	3	0	0	2	8
Geometry	2	0	0	0	2	4
Algebra & Patterns	0	1	0	0	2	3
Data Analysis & Probability	2	0	1	0	1	4
Measurement	1	0	0	2	2	5
Total	8	4	1	2	9	24

Grade 7	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Total
Numbers & Operations	3	2	1	0	0	6
Geometry	1	0	1	2	0	4
Algebra & Patterns	2	2	0	0	1	5
Data Analysis & Probability	1	1	2	0	0	4
Measurement	1	1	2	1	0	5
Total	8	6	6	3	1	24

Grade 8	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Total
Numbers & Operations	1	2	0	1	0	4
Geometry	3	2	0	1	0	6
Algebra & Patterns	3	0	3	2	0	8
Data Analysis & Probability	0	0	0	0	2	2
Measurement	1	2	0	0	1	4
Total	8	6	3	4	3	24

Principal Components Analysis of On-Grade Items

The final method used to explore potential changes in dimensional structure across grades (that is, Research Question One) was a principal components analysis of residuals. Using principal components analysis (PCA), the software program WINSTEPS identifies secondary dimensions in the data by the decomposition of the observed residuals. Residuals are deviations in obtained data from what is predicted based on application of a statistical model--in this case, Rasch model. Note that WINSTEPS applies a Rasch analysis (i.e., a one-dimensional measurement system) regardless of the dimensionality of the data. High correlation of residuals for two items indicates that they may not be locally independent. That is, both items may be measuring some other shared dimension.

The results of the PCA analyses of residuals are presented in Table 4.9. The table is divided into sections by grades. The initial row in each section presents the on-grade items (the other rows pertain to the off-grade linking items and will be discussed in the next section). Eigenvalues are the variances of the principal components. Because the principal components analysis was conducted on the correlation matrix, the variables are standardized, which means that the each item has a variance of 1, and the total variance is equal to the number of items used in the test, in this case (for Grade 3), 26. In other words, there is one unit of information per item so the eigenvalues sum to the number of items. The first component will always account for the most variance (and hence have the highest eigenvalue), and the next component will account for as much of the left over variance as it can, and so on. Hence, each successive component will account for less and less variance. Each residual factor is then measured by the strength of the residual dimension in eigenvalue units; the more eigenvalue units, the stronger the residual dimension. Previous simulation studies have shown that random data (i.e., noise) can have eigenvalues of size 1.4 therefore WINSTEPS and PCA analysis use 1.4 as a cutoff value (Linacre, 2005). That is, a residual factor with an eigenvalue greater than 1.4 could potentially be a valid factor (i.e., enduring or repeatable structure) but if its eigenvalue is less than 1.4 then it most likely noise, random error, etc.

Columns three through six in Table 4.9 show the eigenvalue units for each residual factor or dimension. The first dimension identified in the WINSTEPS analysis is the primary (i.e., intended unidimensional) structure in the data as posited by the Rasch model (Linacre, 2005). PCA is used to analyze the residuals to determine any possible secondary dimensions that could explain residual variation beyond that which is accounted for by the model. For

example, in the Grade 3 form, the first factor after fitting a Rasch model has the strength of two items. The PCA analysis is applied again and the results are presented in second residual factor column of Table 4.9. Therefore, continuing with the Grade 3 example, the findings suggest the second residual factor after removing the primary dimension and the first residual factor has the strength of 1.5 items. The next column denotes that a possible third residual factor after fitting a Rasch model and accounting for the two previous residual factors. This third residual factor has the strength of 1.4 items. Since random data (i.e., noise) can have eigenvalues of size 1.4, there is little evidence of an enduring structure (Linacre, 2005) and therefore WINSTEPS ends the analysis.

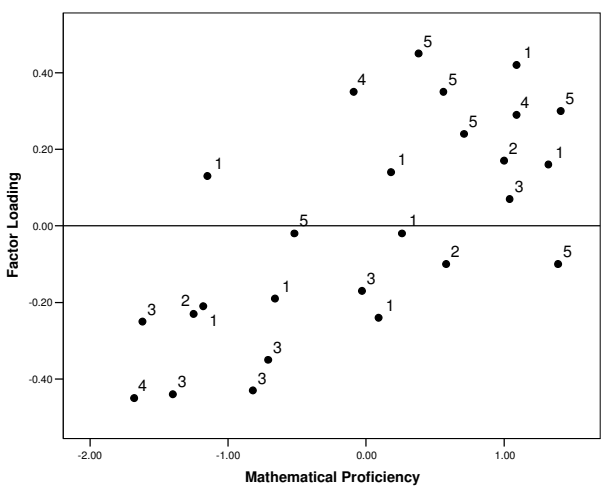
Overall, as shown in the third column of Table 4.9, the first residual factors do not show much strength; the subsequent factors show even less strength. The first residual factor of Grades 4 and 7 accounted for the most unexplained variance (2.2 eigenvalue units), followed by Grades 3 and 8 (2 eigenvalue units) and Grades 5 and 6 (1.6 eigenvalue units). This indicates that after the unidimensional model has been applied to the data, there is little evidence of structure--that is, additional dimensions--in the residuals.

WINSTEPS output also includes principal components factor plots of the standardized residuals. Figure 4.1 a-c shows the first, second and third residual factor plots respectively for Grade 3 on-grade items. The X-axis is the measurement axis (i.e., the posited single dimension). This dimension has been extracted from the data prior to the analysis of the residuals. The items are labeled with their content strand designation: (1) numbers and operations, (2) geometry, (3) algebra and patterns, (4) data analysis and probability and (5) measurement. The trend in Figure 4.1 (a) shows a positive correlation

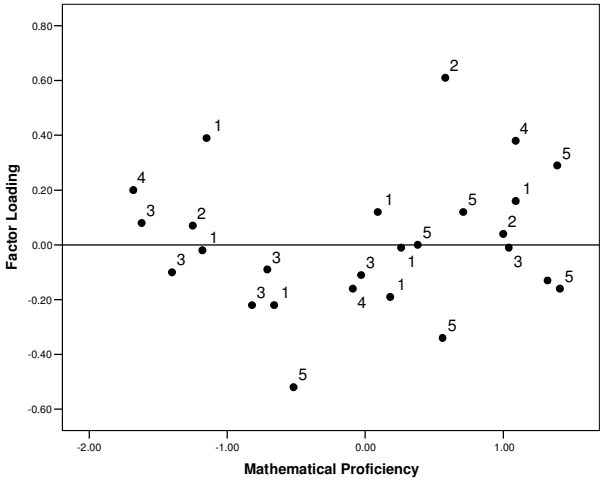
between Rasch item measures and factor loadings. However, notice that this trend disappears as the second and third factors are analyzed (Figure 4.1 b and c).

Table 4.9 *Results of PCA Analysis of Residuals*

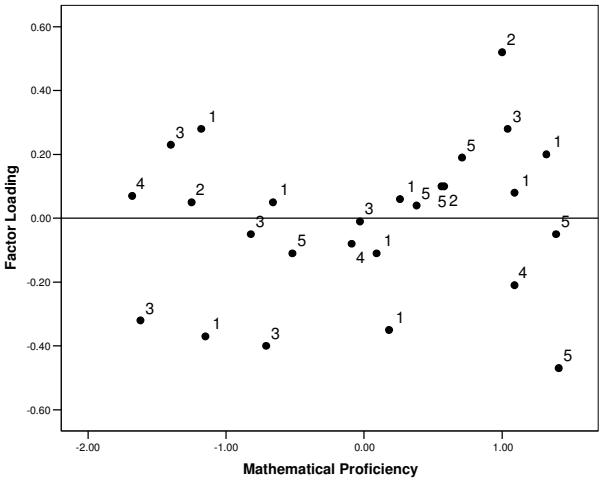
Grade	Total Unexplained Variance (Eigenvalue units)	1st Residual Factor (Eigenvalue units)	2nd Residual Factor (Eigenvalue units)	3rd Residual Factor (Eigenvalue units)	4th Residual Factor (Eigenvalue units)
Grade 3 Items Only	26	2	1.5	1.4	na
Grade 3: G2&3 Items	28	2.1	1.4	1.5	1.4
Grade 3: G3&4 Items	28	2	1.5	1.5	na
Grade 4 Items Only	25	2.2	1.6	1.5	na
Grade 4: G3&4 Items	28	2.3	1.7	1.5	1.3
Grade 4: G4&5 Items	27	2.2	1.6	1.5	na
Grade 5 Items Only	24	1.6	1.5	1.4	na
Grade 5: G4&5 Items	28	1.6	1.5	1.5	na
Grade 5: G5&6 Items	26	1.6	1.6	1.4	na
Grade 6 Items Only	24	1.7	1.5	1.4	na
Grade 6: G5&6 Items	28	1.8	1.6	1.5	1.5
Grade 6: G6&7 Items	26	1.7	1.6	1.5	na
Grade 7 Items Only	24	2.2	1.5	1.4	na
Grade 7: G6&7 Items	28	2.4	1.5	1.5	1.4
Grade 7: G7&8 Items	26	2.2	1.6	1.5	na
Grade 8 Items Only	24	1.9	1.6	1.4	na
Grade 8: G7&8 Items	28	2.1	1.7	1.5	na
Grade 8: G8&9 Items	26	2.1	1.7	1.5	na



(a) First Factor



(b) Second Factor

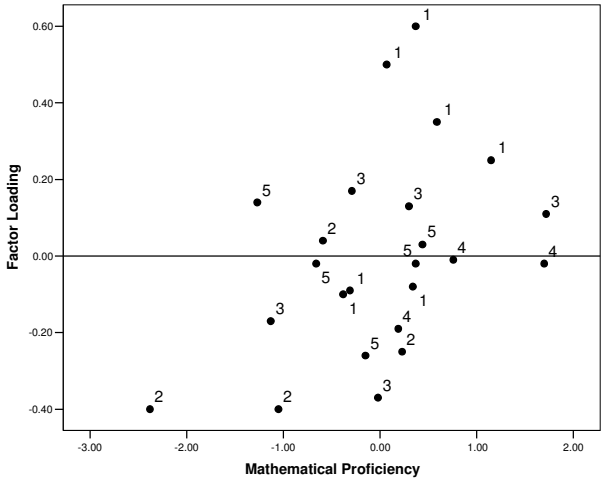


(c) Third Factor

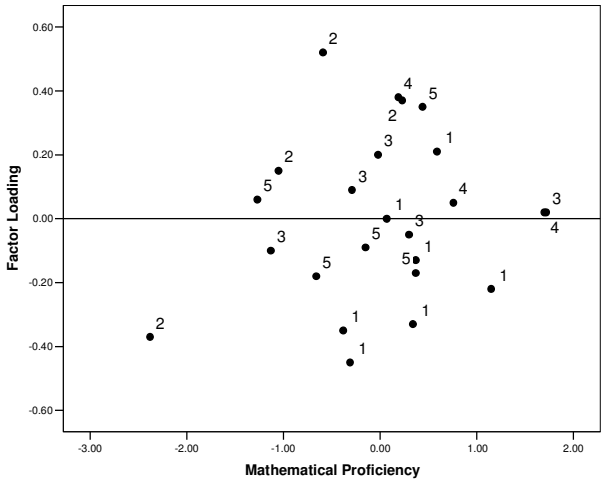
Figure 4.1. Principal Components (Standardized Residual) Factor Plots of Grade 3 On-Grade Items

The WINSTEPS results for Grade 4 and Grade 7 were similar to Grade 3 results shown in Figure 4.1 and are presented in Appendix D. The residual factor plots for Grades 5, 6 and 8 were similar to each other but slightly different from the Grade 3 results. Figure 4.2 shows the results for Grade 5. Notice that the data points of the first residual factor plot

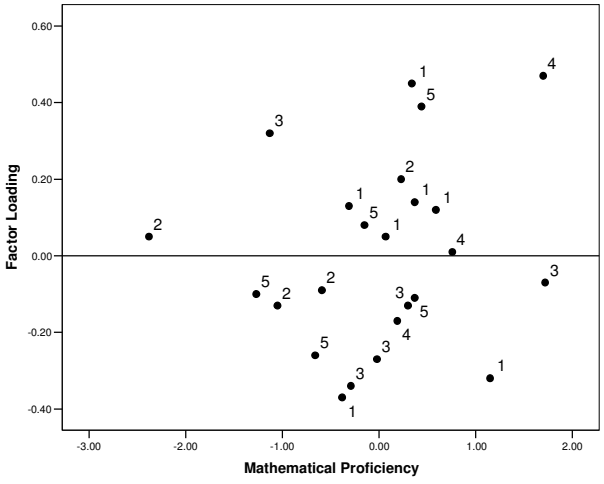
(Figure 4.2 a) appear to be more random and do not show the positive correlation between mathematical proficiency and factor loading as did the first factor in Grade 3 (Figure 4.1a). The second and third residual factor plots also display random placement. The residual factor plots for Grades 6 and 8 are shown in Appendix D.



(a) First Factor



(b) Second Factor



(c) Third Factor

Figure 4.2. Principal Components (Standardized Residual) Factor Plots of Grade 5 On-Grade Items

Summary of Dimensionality of On-grade Items across Grades 3-8

Research Question 1 focused on potential dimensional changes across mathematics achievement tests in Grades 3-8. In the preceding sections, conditional covariance, assessment of essential dimensionality, nonlinear item factor analysis and principal component factor analyses were performed to evaluate whether the test structure. Overall, results applying a conditional covariance analysis approach using the software program DETECT indicated that the on-grade items exhibit weak to moderate amounts of multidimensionality and a complex structure. Recall that when a test exhibits complex structure, some item responses are effectively determined by more than one ability. If each item on a test measures one, and only one dimension, the test structure is labeled as exact or simple structure. If the items load highly on multiple dimensions, then the structure is referred to as a complex structure. Item factor analysis using NOHARM and principal component analyses using WINSTEPS show some evidence of multidimensionality but the results from the assessment of dimensionality employed in DIMTEST purport that the multidimensionality does not appear to be related to the five mathematical content strands. The number of potential dimensions seems to vary slightly and randomly across Grades 3- 8. That is, there does not seem to be relationship among the number of potential dimensions and grade level. However, the results suggest that overall the five content strands are not possible sources of dimensionality of mathematics achievement tests for Grades 3-8.

Results for Inclusion of Linking Items

The second research question considered the possible change in dimensional structure within a grade level test due to the inclusion of off-grade level linking items. There are two types of off-grade items: items from a grade below and items from a grade above the level of a form. The inclusion of off-grade items is a widely used method for developing a vertical scale to span two or more grades. The following sections present the results of analyses where off-grade items (i.e., items written to other grade levels) are included on a grade level form. The number of off-grade items included in the grade level forms examined in this study was very small (typically two to four items), although this, too, is typical of vertical scaling designs in K-12 educational achievement testing.

Conditional Item Covariance and Inclusion of Linking Items

The first method used to assess potential changes in dimensional structure due to the inclusion of linking items was an analysis of conditional item covariances. Exploratory DETECT was applied to off-grade item data using two different runs. First, on- and below-grade items were explored and then data for on- and above-grade items were examined. The results are presented in Table 4.10. When below-grade items were included in the DETECT analyses, D_{max} ranged from 0.3794-0.6595 indicating weak to moderate multidimensionality. The r_{max} index ranged from 0.4843-0.6074 signifying complex structure. The number of clusters ranged from 4-6. As shown in Table 4.10, these results were similar to the findings for the on-grade items alone (shown in the first column). Including above-grade items showed similar results to the on-grade items alone as well as the inclusion of below-grade

items: D_{max} ranged from 0.4199-0.5724, r_{max} ranged from 0.4760-0.5976, and the number of clusters ranged from four to five. Again, applying the guidelines mentioned previously regarding the magnitudes of the Dmax and rmax indices, these analyses reveal that the inclusion of off-grade items results in data that display weak to moderate multidimensionality and complex structure.

The number of clusters and the make-up of the clusters differed depending on which items were included. Further exploration of these clusters from Grade 3 is shown in Table 4.11. The first three rows display the item numbers per cluster based on the grade level of the items: the first row contains Grade 3 items only; the second row displays Grades 2 and 3 items and the third row pertains to Grades 3 and 4 items. The off-grade item numbers are bolded and underlined. These items tended to be dispersed throughout the clusters. The fourth row shows the common items across clusters. For example, Items 2, 3, 9, 10, 15, 18, 19, and 20 were placed in the same cluster across all three conditions (i.e., on-grade, above-grade, and below-grade). The last row in Table 4.11 displays the test specifications of items and content strand. That is, items 1-9 were intended to measure the Numbers and Operations strand. It is interesting to note, however, that the clusters do not generally follow the intended content strands. If the clusters were indeed representing different dimensions based on content then Cluster 1 should contain only items 1-9. The cluster analyses presented here for Grade 3 were typical for the DETECT clusters in the other grades. Results for the other grades are provided in Appendix E.

Table 4.10 *Comparison of Test Structure for Including On-Grade and Off-Grade Items Using Conditional Item Covariances (DETECT)*

	On-Grade Items Only			Below- and On-Grade Items			Above- and On-Grade Items		
Grade	D_{max}	r_{max}	Number of Clusters	D_{max}	r_{max}	Number of Clusters	D_{max}	r_{max}	Number of Clusters
Grade 3	0.4558	0.5534	5	0.4381	0.5569	5	0.4238	0.5188	5
Grade 4	0.4905	0.6032	4	0.4599	0.5524	5	0.4582	0.5638	5
Grade 5	0.4148	0.4998	5	0.3794	0.4843	4	0.4222	0.4965	4
Grade 6	0.4550	0.5204	5	0.4432	0.5003	4	0.4199	0.4760	5
Grade 7	0.6536	0.6119	4	0.6595	0.6074	4	0.5683	0.5572	4
Grade 8	0.5631	0.6197	5	0.4770	0.5424	6	0.5724	0.5976	4

Table 4.11 *Cluster Results for Including On-Grade and Off-Grade Items Using Conditional Item Covariances (DETECT)*

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Grade 3 items only	1, 2, 3, 9, 10, 14, 15, 17, 18, 19, 20, 26	4, 7, 8, 21, 22, 24, 25	6, 16	11, 23, 30	12, 28
Grade 3: G2 & 3 Items	1, 2, 3, 5, 9, 10, 15, 18, 19, 20, 27	4, 11, 12, 22, 23, 30	6, 14, 16	7, 8, 21, 24, 25	26, 28
Grade 3: G3 & 4 Items	2, 3, 9, 10, 11, 14, 15, 17, 18, 19, 20, 29, 30	1, 4, 7, 8, 21, 22, 24, 25	6, 12, 13, 16	23	26, 28
Grade 3 Common Items (across clusters)	2, 3, 9,10,15, 18, 19, 20	4, 7, 22	6,16	na	28
Theoretical Clusters by Content Strands	Items 1-9: Numbers & Operations	Items 10-13: Geometry	Items 14-19: Algebra & Patterns	Items 20-22: Data Analysis & Probability	Items 23-30: Measurement

Note: Off-grade item numbers are underlined.

Assessment of Essential Dimensionality When Off-Grade Items Are Included

A second approach to answering Research Question 2 regarding the potential affects on dimensionality of including linking items involved assessing the essential dimensionality of the data via the computer program DIMTEST. The results of applying DIMTEST when off-grade items are included are shown in Table 4.12. The results for below- and on-grade items are shown in the shaded rows; results for the above- and on-grade items are presented in the unshaded rows. The first column in the table provides the grade and item combinations and the second column specifies the number of off-grade items included in each grade level form. The last column provides the p-values associated with the T statistics that DIMTEST calculates. As seen in the last column, the p-values generated by DIMTEST do not permit the null hypotheses of unidimensionality to be rejected. That is, for none of the grade levels does the inclusion of off-grade items result in a test that is dimensionally distinct from one that is constructed of on-grade items only.

Table 4.12. *Assessment of Essential Unidimensionality (DIMTEST) Including Off-Grade Items*

Item Levels	No. of Off-Grade Items	p-value
Grade 3: G2&3 Items	2	0.3278
Grade 3: G3&4 Items	2	0.5584
Grade 4: G3&4 Items	3	0.212
Grade 4: G4&5 Items	2	0.1075
Grade 5: G4&5 Items	4	0.5300
Grade 5: G5&6 Items	2	0.6125
Grade 6: G5&6 Items	4	0.9924
Grade 6: G6&7 Items	2	0.4672
Grade 7: G6&7 Items	4	0.4349
Grade 7: G7&8 Items	2	0.5921
Grade 8: G7&8 Items	4	0.3675
Grade 8: G8&9 Items	2	0.8157

Nonlinear Item Factor Analysis When Linking Items Are Included

Another approach to examining the presence of linking items on dimensional structure (i.e., Research Question 2) is nonlinear item factor analysis. It was hypothesized that there would be two dimensions related to the grade level: one dimension representing on on-grade items and a second dimension resulting from the off-grade level items. Therefore, confirmatory factor analyses using the software program NOHARM and a priori specification of two dimensions was applied to the datasets containing on- and off-grade

items. The results for the two-dimensional analyses are presented in Table 4.13. The results for below- and on-grade items are shown in the shaded rows and the above- and on-grade items are presented in the unshaded rows. The RMSR were small, ranging from 0.0092 to 0.0122 for below- and on-grade items and from 0.0093 to 0.0117 for the above-and on-grade items. Tanaka's Index ranged from 0.9475 to 0.9598 and 0.9414 to 0.9609, respectively. Recall that interpretation is rather limited because currently there are no specific guidelines for RMSR or Tanaka's Index. In general, a good model fit is indicated by a small RMSR (i.e., close to zero) and a high Tanaka's index (closer to 1).

Table 4.13. *Confirmatory Nonlinear Item Factor Analysis (NOHARM) for Off-Grade Items (Two Dimensions)*

Item Levels	RMSR	Tanaka's Index
Grade 3: G2&3 Items	0.0103	0.9541
Grade 3: G3&4 Items	0.0106	0.9503
Grade 4: G3&4 Items	0.0092	0.9511
Grade 4: G4&5 Items	0.0093	0.9513
Grade 5: G4&5 Items	0.0101	0.9598
Grade 5: G5&6 Items	0.0104	0.9609
Grade 6: G5&6 Items	0.0112	0.9455
Grade 6: G6&7 Items	0.0114	0.9462
Grade 7: G6&7 Items	0.0122	0.9414
Grade 7: G7&8 Items	0.0118	0.9476
Grade 8: G7&8 Items	0.0110	0.9475
Grade 8: G8&9 Items	0.0117	0.9408

Exploratory analyses were also conducted to determine where the off-grade items would load on a two-factor solution if NOHARM selected the factor loadings. The results for the analyses of Grade 3 data are shown in Table 4.14. The underlined values emphasize the off-grade items. In left panel of Table 4.14, items 5 and 26 are the below-grade (Grade 2) items administered with the Grade 3 form. In right panel of Table 4.14, items 13 and 29 are above-grade items (Grade 4) included on the Grade 3 form. The bolding denotes the largest factor loading for each item. The off-grade items do not appear to form a separate factor in either the below- or above-grade items and even appear to load on different factors. The clusterings appeared to be random and no observable pattern in the item types was

distinguished. The results for Grades 4-8 were similar to Grade 3 and are presented in Appendix F. These results indicate that the presence of a small number of linking items do not appear to change the dimensional structure of the test forms.

Table 4.14. *NOHARM Factor Loadings for Grade 3: On-and Off-Grade Items*

Grades 2 and 3			Grades 3 and 4		
Item #	Factor 1	Factor 2	Item #	Factor 1	Factor 2
1	**0.431**	0.344	1	**0.417**	0.358
2	**0.517**	0.083	2	**0.480**	0.093
3	**0.523**	0.137	3	**0.523**	0.155
4	0.337	**0.399**	4	0.329	**0.399**
5	***0.643***	*0.121*	6	**0.114**	-0.001
6	**0.082**	0.002	7	0.321	**0.443**
7	0.329	**0.438**	8	-0.061	**0.228**
8	-0.06	**0.251**	9	**0.472**	0.211
9	**0.512**	0.163	10	**0.543**	-0.064
10	**0.515**	-0.101	11	**0.364**	0.191
11	**0.321**	0.136	12	**0.149**	0.097
12	**0.151**	0.09	*13*	***0.474***	*0.089*
14	**0.417**	0.02	14	**0.427**	0.016
15	**0.564**	0.017	15	**0.556**	0.019
16	**0.185**	0.091	16	**0.207**	0.080
17	**0.726**	-0.053	17	**0.723**	-0.038
18	**0.577**	0.054	18	**0.604**	0.074
19	**0.638**	-0.028	19	**0.625**	-0.015
20	**0.776**	-0.056	20	**0.778**	-0.024
21	-0.022	**0.097**	21	-0.035	**0.066**
22	0.21	**0.531**	22	0.190	**0.569**
23	0.105	**0.223**	23	0.113	**0.242**
24	0.017	**0.355**	24	-0.007	**0.350**
25	0.259	**0.715**	25	0.231	**0.698**
26	***0.353***	*0.18*	26	**0.324**	0.152
27	**0.622**	0.241	28	-0.026	**0.174**
28	-0.027	**0.196**	*29*	*0.291*	***0.385***
30	**0.318**	0.207	30	**0.321**	0.248

Principal Components Analysis for Inclusion of Off-Grade Items

The final method used to assess Research Question 2 (potential influence of off-grade level items on the dimensional structure) was a principal components analysis. The principal components analyses for the off-grade items using WINSTEPS with Grade 3 items is shown in Table 4.15. For comparison purposes, the first row contains the results from on-grade items only. The next two rows show the eigenvalue units for off-grade items. Note that the amounts of unexplained variance explained by additional factors are similar to the corresponding results for the on-grade items. The residuals from the Grade 2 and 3 items displayed a fourth factor but it the eigenvalue is very small.

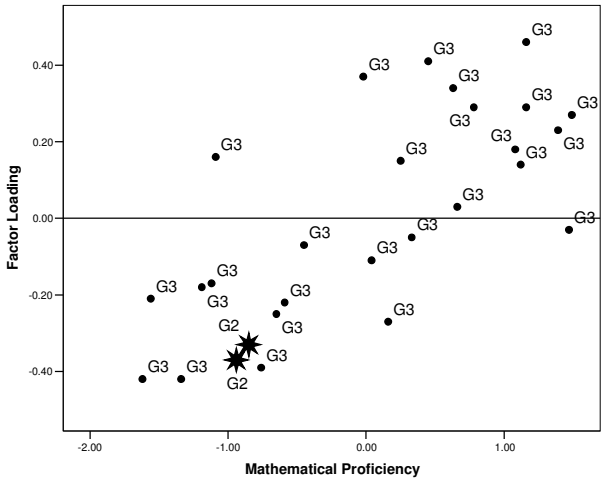
Table 4.15. *Principal Components Analyses Results for Grade 3 On- and Off-Grade Items*

Grade	Total Unexplained Variance (Eigenvalue units)	1st Residual Factor (Eigenvalue units)	2nd Residual Factor (Eigenvalue units)	3rd Residual Factor (Eigenvalue units)	4th Residual Factor (Eigenvalue units)
Grade 3: G3 Items Only	26	2	1.5	1.4	na
Grade 3: G2 & 3 Items	28	2.1	1.4	1.5	1.4
Grade 3: G3 & 4 Items	28	2	1.5	1.5	na

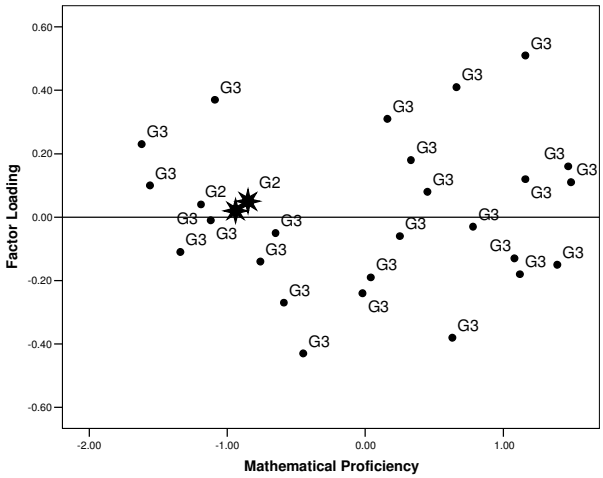
The factor plots of the residuals based on the inclusion of Grade 2 items on the Grade 3 form are shown in Figure 4.3 a-d. The item labels show the grade level of the item (G2 or G3). The Grade 2 items are also marked with an asterisk (") in the figures. These plots were basically identical to the plots for the on-grade items only presented previously in Figure 4.1 a-c. The first factor (after extracting the primary dimension) plot shows a positive correlation between the mathematical proficiency and the factor loading (Figure 4.3 a).

The other plots of the residuals in Figure 4.3 (b-d) display residuals that are more random and do not appear to follow a trend which suggests that there is no further important or enduring structure in the data. That is, a unidimensional model appears to fit the data well.

Analyzing Grades 3 and 4 items on the Grade 3 form using WINSTEPS produced the factor residual plots shown in Figure 4.4 a-c. Notice that the positive trend seen in the first factor of both on-grade and below/on grade items does not appear when items from Grades 3 and 4 are used (Figure 4.4 a) and the residuals are more dispersed. This random pattern is also seen in the second and third factors. The residual factor plots from the other forms using the respective off-grade items are shown in Appendix G.

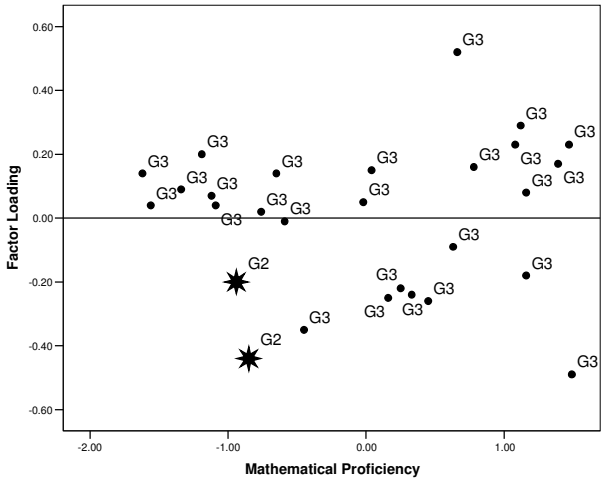


(a) First Factor

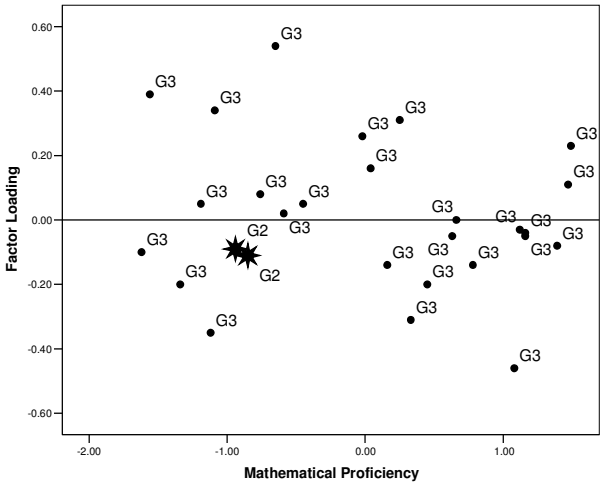


(b) Second Factor

Note: Off-grade items are designated with a ” symbol

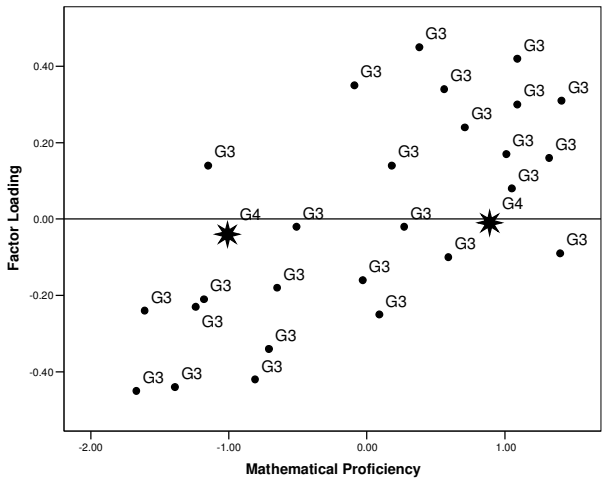


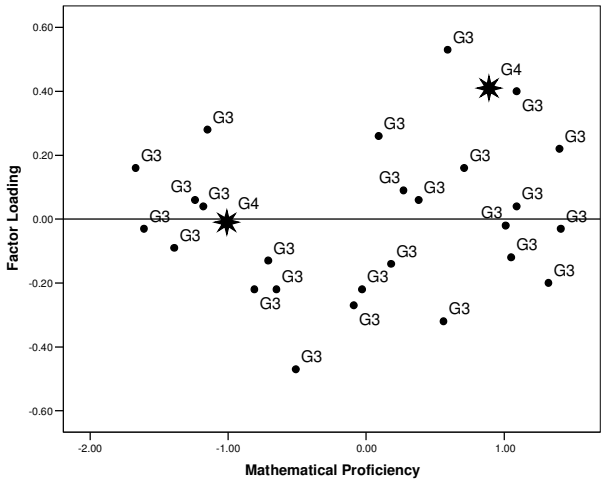
(c) Third Factor



(d) Fourth Factor

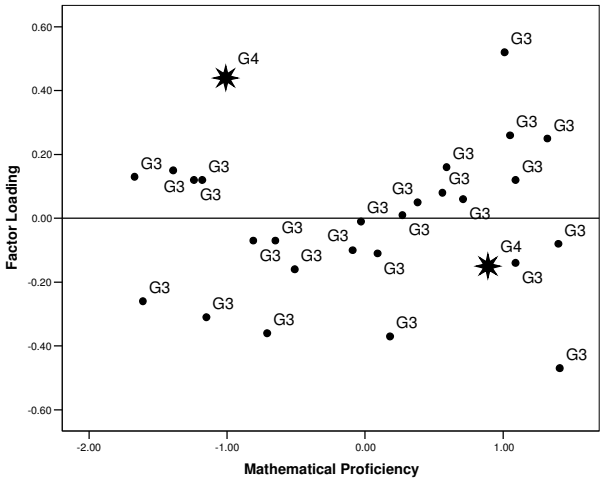
Figure 4.3. Principal Components (Standardized Residual) Factor Plots of Grade 3: Grade 2 and 3 Items





Note: Off-grade items are designated with a ” symbol.

(a) First Factor *(b) Second Factor*



(c) Third Factor

Figure 4.4. Principal Components (Standardized Residual) Factor Plots of Grade 3: Grade 3 and 4 Items

Summary of Investigation of Including Off-Grade Items on Test Dimensionality

Research question 2 considered possible changes in the dimensional structure of on-grade level form when off-grade items are included on the test form. Overall, the inclusion of off-grade items in the test structure analyses did not appear to change the dimensionality results. As in the analysis previously reported regarding the dimensional structure of on-grade items (i.e., research question 1), the software used to gauge dimensionality (DETECT) again identified weak to moderate multidimensionality and complex structure. The inclusion of off-grade items tended to change the clustering of items compared to the clustering that was obtained from analysis of on-grade items alone. According to the results produced by the software program designed to assess essential unidimensionality (i.e., DIMTEST), off-grade items were not dimensionally different from on-grade items which was evidenced by the factor loadings obtained by the nonlinear item factor analysis approach using NOHARM. The principal components analysis of residuals found little structure in the residuals that would suggest the presence of multidimensionality when off-grade items are included in a grade level form.

Comparison of Methods

Research question 3 concerned possible differences in the results of dimensionality analyses yielded by the various approaches and software programs. This last section describes comparisons of those different results. As expected, the different methods and programs lead to different conclusions about the test structure not only regarding the number of dimensions but also regarding the items that comprise those dimensions. In addition, the unique pieces of information offered by each program can be combined together to better understand the data structure.

Previously the results for exploratory approaches to the investigation of test structure for on-grade items using DETECT and NOHARM were presented. DIMTEST is also capable of doing an exploratory analysis as well. When DIMTEST is used in this way, rather than the researcher specifying the initial subtest, the program "ATFIND" is used to determine the most homogenous subtest from all the items on the form. Table 4.16 presents the results of exploratory DIMTEST using the on-grade items. Recall that DIMTEST tests the hypothesis of essential unidimensionality. According to the exploratory DIMTEST, essential unidimensionality holds for Grade 4 and Grade 6 forms. However, while DIMTEST does not find evidence that Grades 3, 5, 7 and 8 display essential unidimensionality, it cannot determine how many more dimensions are present.

Table 4.16. *Results of Exploratory Assessment of Essential Unidimensionality (DIMTEST) Using On-Grade Items*

Grade	p-value	Result
Grade 3	0.0222	Reject Ho
Grade 4	0.0456	Retain Ho
Grade 5	0.0067	Reject Ho
Grade 6	0.0955	Retain Ho
Grade 7	0.0081	Reject Ho
Grade 8	0.0011	Reject Ho

A comparison of the number of resulting dimensions from DETECT, DIMTEST, NOHARM and WINSTEPS for on-grade items are shown in Table 4.17. This table was created by combining the results of the exploratory analyses presented previously in Tables 4.1, 4.3, 4.6, and 4.9. Across the grade levels studied, the clustering of items produced by DETECT suggests that the number of dimensions ranges from four to five. The hypothesis tests in DIMTEST reject essential unidimensionality in four of the six grades. The number of dimensions estimated using NOHARM ranges from one to over five. The principal components analysis (PCA) of residuals using WINSTEPS did not identify any pattern in the residuals once the unidimensional (Rasch) model had been fitted to the data.

Table 4.17. *Summary of Overall Exploratory Analyses Using On-Grade Items*

Grade	Conditional Item Covariance (DETECT)	Assessment of Essential Unidimensionality (DIMTEST)	Nonlinear Item Factor Analysis (NOHARM)	PCA Analysis of Residuals (WINSTEPS)
Grade 3	5	>1	2 or 3	1
Grade 4	4	~1	2 or 3	1
Grade 5	5	>1	1	1
Grade 6	5	~1	2 or 3	1
Grade 7	4	>1	4	1
Grade 8	5	>1	5+	1

Additional output produced by the software programs DETECT, DIMTEST and NOHARM was considered more in depth. For example, both DETECT and NOHARM indicate that there are at least five dimensions in the Grade 8 data. (Note: while Grade 3 examples were presented throughout this chapter, Grade 8 was chosen for this particular example because both DETECT and NOHARM suggest that there are five dimensions and therefore the five dimensions suggested by DETECT could be compared with the five dimensions suggested by NOHARM for Grade 8. Grade 3 does not allow that five-five comparison. In Grade 3, DETECT suggests there are five possible clusters but NOHARM results suggests two or three.) However, the clusters and factors are not comprised of the same items. Consider the comparison of the Grade 8 results shown in Table 4.18. The items that make up the clusters and the items with the largest factor loadings are not the same items. Cluster 1 and Factor 1 both contain eight items but the items are not the same eight items. There are only three items (Items 1, 10 and 22) that are common in both Cluster 1 and

Factor 1. Comparisons for the other grades are found in Appendix H. These comparisons also showed a disparity between the DETECT item clusters and the NOHARM factor loadings.

Another comparison can also be made using the ATFIND items from DIMTEST. Recall that the ATFIND procedure in DIMTEST finds the most homogenous subset of items from the entire form. The items found by ATFIND for the Grade 8 form are listed in the third row of Table 4.18. Note that the "most homogenous items" found by ATFIND/DIMTEST are not the same set of items clustered by DETECT or loaded on factors by NOHARM.

Table 4.18. *Comparison of Exploratory Results from Grade 8 On-Grade Items by Software Program*

Software Program	Results				
DETECT	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
	1, 2, 3, 8, 10, 13, 15, 22, 27	4, 7, 12, 16, 18, 29	11, 21, 25, 28	14, 17, 20, 30	20
NOHARM	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
	1, 10, 11, 12, 20, 21, 22, 30	2, 3, 8, 13, 27, 29	14, 15, 16	4, 7, 17, 18	24, 25, 28
DIMTEST	AT Subtest				
	4, 5, 13, 15, 17, 20, 22, 23				

Summary

Three research questions were explored using data from typical mathematics achievement tests for Grades 3-8. The exploration was conducted using four different approaches: conditional item covariances, assessment of essential unidimensionality, nonlinear factor analysis, and principal components analysis. Research question 1 considered possible influence of five mathematical content areas on the dimensional structure. While the data did display small to moderate amounts of multidimensionality and was complex in nature, this did not appear to be generated by the five content areas. Research question 2 explored the use of off-grade items in a linking project. The scope was rather limited with so few off-grade items but the available data did not appear to be influenced by the inclusion of off-grade items. In regards to Research Question 3, each of the software programs designed to provide information relevant to assessment of test structure appears to offer a unique piece of information to the bigger picture of dimensionality. For example, DETECT estimates the amount of multidimensionality and complexity of the data structure and this information is helpful in interpreting the NOHARM factor loadings where each item loads on each factor (implying a complex structure).

Overall, determining test structure is a complicated endeavor particularly when the data display complex structure as is typical of educational data. Therefore, the question of dimensionality is not appropriately viewed as a "yes/no" question, but as a question of "how much?". How much multidimensionality can be present before parameter estimates become affected? How much multidimensionality can be permitted before validity evidence is threatened? How much correlation is needed between factors before they constitute a single dimension? The next chapter expands on these future research questions in light of the

findings of this study to move to a better understanding of dimensionality of educational test data.

CHAPTER 5

CONCLUSIONS AND DISCUSSION

Assessing the dimensionality of test data is an important yet difficult task, particularly when working with real test data where the true underlying factor structure is unknown. As Ackerman, Gierl, and Walker have observed, "working with real test data is never easy and rarely are the interpretations straightforward" (2003, p. 38). The psychometric community must attend to this caution and carefully evaluate the results of dimensionality assessment with substantive interpretation. For example, the results of this study suggest that the test structure for the Grades 3 – 8 mathematics achievement tests are complex and display weak to moderate amounts of multidimensionality. However, that primary finding is only part of the story. The rest of the story unfolds when other factors, related to the process of learning, factors affecting tests in general and mathematical tests in particular, are considered before making confident claims about test structure.

The following sections of this chapter will first summarize some of the key findings of this study and interpretations of those findings, following a brief review of the study's limitations. Next, the concept of dimensionality itself will be examined, and a refined process for examining dimensionality will be proposed. Finally, the chapter will conclude with suggestions for future research in this area.

Research Summary and Interpretations

Before beginning a summary of the key findings of this research, it is important to review some limitations of the study sample, design, and analysis. One limitation of this study was the length of each test (24 -28 items). This limitation is particularly important in regards to Research Question 2 (i.e., the inclusion of off-grade level items on the dimensional structure). Due to the linking study design, each on-grade form contained only a few off-grade items (2-4 items). This linking design was a limitation because more off-grade items could potentially exhibit dimensionality due to content exposure, curricular and/or difficulty factors. In addition, the item format used for all of the mathematics items studies was limited to four-option multiple-choice items; therefore, the results can not be extended automatically to different item formats. In this study, four methods were used for investigating dimensional structure. Each of the four dimensionality assessment methods and programs introduces its own set of limitations as well. For example, two of the approaches (conditional item covariance and assessment of essential unidimensionality) are nonparametric approaches and two methods are parametric (nonlinear factor analysis and principal components analysis). Parametric methods assume a particular parametric model for the IRF while the nonparametric methods assume only that the IRF is monotonic. Therefore, assuming a particular parametric model might or might not fit the data well. One parametric model in particular, the Rasch model (1-PL), has additional limitations. Other IRT models include parameters for differences in item discrimination (2-PL) and guessing (3-PL) but WINSTEPS only employs the Rasch model. It is a possibility that some findings in the study would have differed or other interpretations been plausible had additional

parameters been included in item calibrations (e.g., guessing, discrimination). Appendix A contains more information about each program used in this study.

These limitations notwithstanding, this study yielded insights into what is known about the dimensionality of mathematics achievement tests, how that dimensionality is affected when out-of-level linking items are embedded in mathematics achievement tests for the purpose of creating vertical (i.e., across-grade) scales, and how various procedures for assessing dimensionality perform in these contexts. These findings correspond to the three main research questions addressed in this study and the following summary of findings is organized according to those research questions.

The first research question explored the dimensional structure across mathematical achievement tests for Grades 3 through 8, in which only on-grade items were considered. Overall, the results suggested each grade level test form displayed a complex structure with weak to moderate degrees of multidimensionality. While the number of potential dimensions seems to vary slightly and randomly across Grades 3- 8, the results suggest that overall the five content strands are not possible sources of dimensionality of mathematics achievement tests for Grades 3-8. Potential sources of multidimensionality could be related to item difficulty as well as differing item demands such as reading loads introduced by highly contextualized problem situations, interpreting graphs or figures, inclusion of math vocabulary and/or symbols, whether a tool such as a protractor or ruler is needed or the number system involved in the item (whole numbers, decimals, fractions, positive/negative integers, etc.).

The second research question considered the possible change in dimensional structure within a grade level test due to the inclusion of off-grade level linking items (i.e., items from

a grade below and items from a grade above the level of a form). The analyses of this study did not indicate that the inclusion of off-grade items resulted in a test that is dimensionality distinct from one that is constructed of on-grade items only. This conclusion was consistent across Grades 3 – 8.

Research question 3 explored possible differences in the results of dimensionality analyses yielded by the various approaches and software programs. The different methods and programs lead to different conclusions about the test structure not only regarding the number of dimensions but also regarding the items that comprise those dimensions. Although different results were produced, it was also learned that the specific pieces of information offered by each program could be integrated together to better understand the data structure. For example, DETECT clusters could be compared to the factor loadings determined by NOHARM to determine which items seem to be working together. Overall however, as expected, the results produced by the various approaches suggested the mathematics tests analyzed in this study displayed complex structures with weak to moderate amounts of multidimensionality. The extent and implications of this multidimensionality are interpreted in the following sections.

Complex Structure

The results of the conditional item covariance and DETECT's r_{max} index and the factor loadings yielded by the nonlinear item factor analysis operationalized by NOHARM suggested a complex test structure in the mathematics achievement tests across grades 3-8 (see Tables 4.1 and 4.7 respectively). Recall from Chapter 2 that if each item on a test measures one, and only one dimension, the test structure is labeled as exact or simple

structure (see Figure 2.1). If the items load highly on multiple dimensions, then the structure is referred to as a complex structure. When a test exhibits complex structure, some item responses are effectively determined by more than one ability or construct. When complex structure is observed, the type of test, the overall content, and the substantive and cognitive aspects of mathematics curriculum, instruction, language, and other assessment issues must be considered.

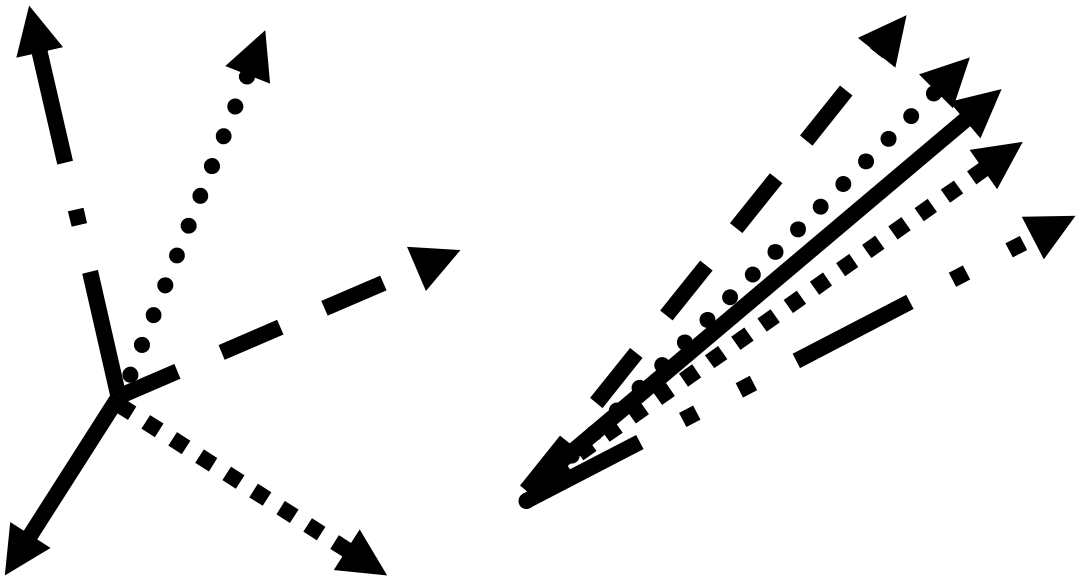
Many mathematical skills span content strands and are used in conjunction with other skills and/or in subsequent skills. Mathematics is often conceptualized as being made up of separate strands (as shown in Figure 2.3 of the NCTM content standards across grade bands) but this tends to be more an organizing principle for curriculums and textbooks rather than an indication of the structure of multidimensionality in the mathematical achievement construct. The results of this study did not show a relationship between dimensionality and the content strands. Additionally, these findings support the NCTM Connections Standard which proposed that all students (prekindergarten through Grade 12) should be able to make and use connections among mathematical ideas and see how the mathematical ideas interconnect. According to NCTM, "mathematics is not a collection of separate strands or standards, even though it is often partioned and presented in this manner" (National Council of Teachers of Mathematics, 2000, p. 64).

There is, however, a great amount of overlap and correlation in mathematical topics, skills and strands. For example, consider basic addition of whole numbers which is classified as a skill in the Numbers and Operations strand. Knowing addition facts leads to other skills such as (1) subtraction facts (also in the Numbers and Operations strand), (2) finding the mean of a set of data (Data Analysis and Probability strand) and (3) determining whether

angles in a figure are complimentary or supplementary (Geometry strand). The last illustration (3) is particularly interesting. There tends to be more distinction or difference between algebra and geometry particularly when geometry involves learning basic shapes, properties of figures or spatial reasoning. However, at some point the content strands intertwine again, as geometry problems require students to use the four basic operations (addition, subtraction, multiplication and division) to find perimeters, areas and volumes or basic algebra skills and algebraic thinking to solve for a missing angle or side length. Thus, given the complex nature of mathematical skills and their correlations, the complex nature of the test structure is not surprising; indeed, it should be expected. The study results reflect the interconnectivity of the strands.

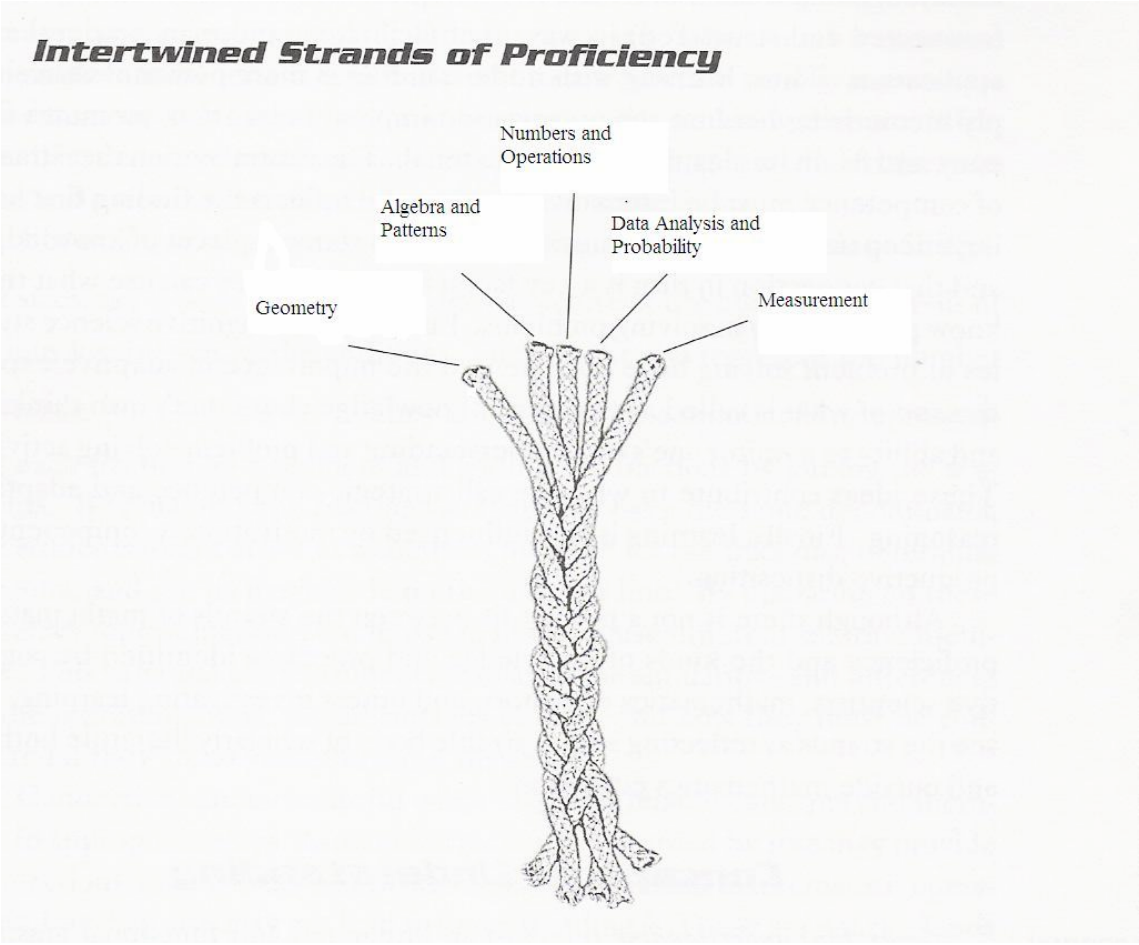
While the determination of complex structure in the data does not indicate the number of dimensions, it does suggest something about interaction of the dimensions. Figure 5.1 illustrates two possible relationships of factors of a complex structure. Figure 5.1(a) illustrates less correlation among five factors while Figure 5.1(b) displays five factors that are more correlated. Regarding the highly correlated factors observed in the mathematics achievement test data analyzed in this study, a relevant analogy, or image is that of a rope. A rope is made up of different fibers or strands that can be distinguished but are wound together to produce one rope as illustrated in Figure 5.2. If the constructs of a test are represented by fibers of the rope, this analogy shows how several dimensions might seem distinct and yet are woven together so tightly (i.e., correlated) that the minor dimensions blend into a single more prominent cable. Therefore, the complexity of the data structure along with the known overlap of mathematics skills perhaps suggest that mathematics achievement tests could represent a fundamentally unidimensional construct. Importantly, it should be noted here

that the phrase, “essential unidimensionality”, is being avoided as it denotes a specific statistical model developed by Stout and Nandakumar (Nandakumar, 1991, 1993; Stout, 1987, 1990; Stout et al., 1996).



(a) Distinct Dimensions (b) Highly Correlated Dimensions

Figure 5.1. Graphic Representations of Complex Structure and Multidimensionality



Adapted from Kilpatrick, Swafford and Findell (2001).

Figure 5.2. Relationships among Mathematical Strands

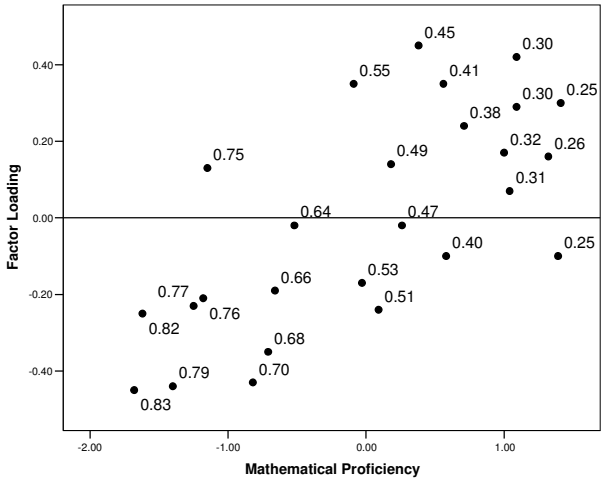
Interpretation of Multidimensionality

Although the complex nature of both the mathematical content and mathematical achievement test structure must be acknowledged, it is also important to evaluate the evidence of weak to moderate amounts of multidimensionality in the test data. The response to an item is often dependent upon several secondary dimensions in addition to the hypothesized primary dimension or proficiency (Traub, 1983). Dimensionality is a property of both the test and the examinee population taking the test (Hattie, 1985; Nandakumar &

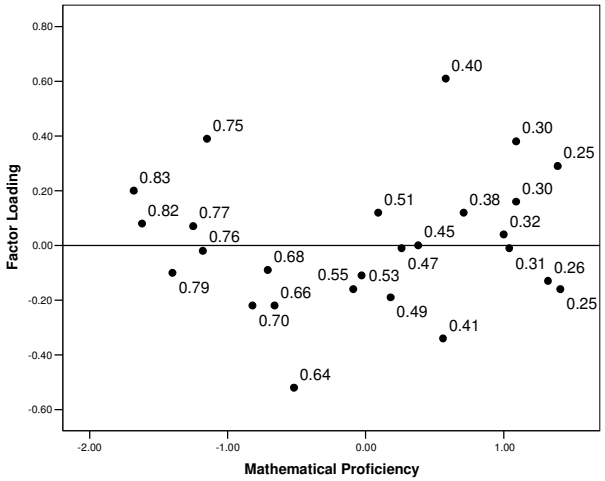
Stout, 1993; Reckase, 1990; Tate, 2002). There are several important features that are examinee-by-instrument interaction that can possibly confound dimensionality: namely, item difficulty and reading demand of mathematical items.

Item Difficulty and Dimensionality

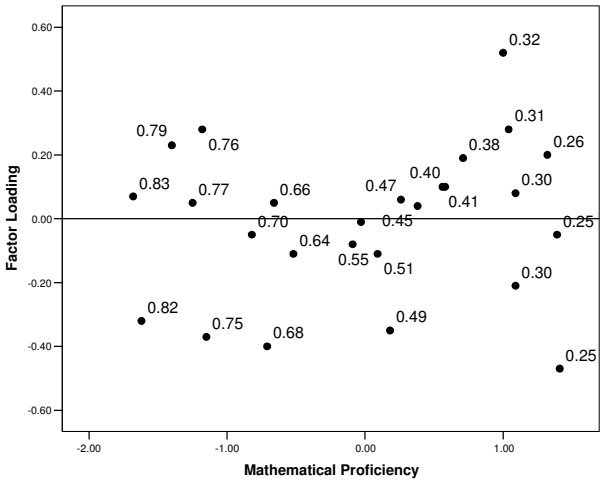
Dimensionality can be confounded with item difficulty if the factors represent items with comparable difficulty levels as opposed to items that measure distinct dimensions (Ackerman et al., 2003). In order to examine possible effects of item difficulty, p-values were used as measures for item difficulty. The WINSTEPS standardized residual plots were modified so that the item labels displayed the item p-values. Figure 5.3 (a-c) shows the modified WINSTEPS residual plots for Grade 3 items that were originally presented in Figure 4.1(a-c). The first residual plot (after removing the first, predominant factor) presented in (a) shows a positive correlation between the harder items (those items with lower p-values indicating fewer students answered the items correctly) and the easier items (higher p-values). However, a similar relationship (i.e., a positive correlation) is not readily apparent in the second and third residual plots (Figure 5.3 b and c). This would indicate that item difficulty explains much of the variance in the first residual plot and once it is removed or accounted for, then there is little, if any, remaining structure in the residuals. The other grades (4-8) showed similar results and the plots for these grades are presented in Appendix I.



(a) First Factor



(b) Second Factor



(c) Third Factor

Figure 5.3. Principal Components (Standardized Residual) Factor Plots of Grade 3

The clusters of items identified by the conditional item covariance analysis conducted using DETECT showed evidence of a difficulty factor as well. Consider the Grade 3 results shown in Table 5.1. The majority of the easiest (highest p-values) items are clustered in Cluster 1. The mean p-value of the items in Cluster 1 is 0.68 whereas the mean p-values of items in Clusters 2, 3, 4 and 5 are 0.44, 0.29, 0.34 and 0.37 respectively. The results for Grades 4-8 showed similar patterns and are presented in Appendix J.

Table 5.1 *DETECT Cluster Results with Item P-Values for Grade 3*

Item #	P-Value	Item #	P-Value	Item #	P-Value
Cluster 1		*Cluster 2*		*Cluster 3*	
20	0.83	4	0.75	16	0.31
15	0.82	21	0.55	6	0.26
17	0.79	7	0.49	Mean	**0.29**
10	0.77	24	0.45		
3	0.76	22	0.30	*Cluster 4*	
19	0.70	8	0.30	11	0.40
18	0.68	25	0.25	23	0.38
2	0.66	Mean	**0.44**	30	0.25
26	0.64			Mean	**0.34**
14	0.53				
9	0.51			*Cluster 5*	
1	0.47			28	0.41
Mean	**0.68**			12	0.32
				Mean	**0.37**

Similar patterns, suggesting a difficulty factor, were evident in the exploratory factor analysis loadings produced by NOHARM. The p-values and factor loadings are shown for a two-factor solution for the Grade 3 form in Table 5.2 and a three-factor solution in Table 5.3. It appears that in each solution there is a factor where most items with high p-values load. For example, in the two-factor solution most of the items with the highest p-values load on

the first factor. The mean of the p-values of the items in the first factor is 0.57 compared to the mean of the second factor which is 0.43. The differences in the means are even greater when considering a three-factor solution as shown in Table 5.3. The second factor contains items with a mean p-value of 0.66 compared to first and third factors where the mean p-values are 0.38 and 0.47 respectively. Similar results were found for the other grades and those results are given in Appendix K.

Table 5.2 *NOHARM Factor Loadings for Grade 3 Two-Factor Solution*

First Factor			Second Factor		
Item	Factor Loading	P-value	Item	Factor Loading	P-value
20	0.763	0.83	4	0.400	0.75
15	0.564	0.82	21	0.102	0.55
17	0.717	0.79	7	0.433	0.49
10	0.519	0.77	24	0.358	0.45
3	0.537	0.76	28	0.197	0.41
19	0.633	0.70	23	0.216	0.38
18	0.615	0.68	22	0.527	0.30
2	0.488	0.66	8	0.245	0.30
26	0.332	0.64	25	0.718	0.25
14	0.427	0.53		*Mean p-value:*	*0.43*
9	0.482	0.51			
1	0.428	0.47			
11	0.362	0.40			
12	0.136	0.32			
16	0.201	0.31			
6	0.104	0.26			
30	0.343	0.25			
	Mean p-value:	*0.57*			

Table 5.3 *NOHARM Factor Loadings for Grade 3 Three-Factor Solution*

First Factor			Second Factor			Third Factor		
Item	Factor Loading	P-value	Item	Factor Loading	P-value	Item	Factor Loading	P-value
4	0.485	0.75	20	0.736	0.83	7	0.410	0.49
24	0.254	0.45	15	0.550	0.82	21	0.143	0.55
11	0.496	0.40	17	0.710	0.79	25	0.662	0.25
23	0.281	0.38	10	0.512	0.77	26	0.432	0.64
12	0.179	0.32	3	0.501	0.76	28	0.287	0.41
16	0.155	0.31	4	0.267	0.75	*Mean p-value:*		*0.47*
8	0.262	0.30	19	0.646	0.70			
22	0.603	0.30	18	0.619	0.68			
30	0.325	0.25	2	0.486	0.66			
Mean p-value:		*0.38*	14	0.426	0.53			
			9	0.446	0.51			
			1	0.366	0.47			
			6	0.096	0.26			
			Mean p-value:		*0.66*			

Reading Demand and Dimensionality

Currently, many mathematics achievement tests consist of both decontextualized computation and moderately to highly contextualized problem-solving items. The problem solving items contain more verbiage that could require an additional ability (i.e., reading) not essential for the solution of the more decontextualized mathematical computation items. Consider the contrast of items shown in Figure 5.4. The mathematics item shown in the first panel of Figure 5.4 requires more reading and understanding of the context than does the item shown in the second panel. In a recent simulation study, Beretvas and Williams (2004) found that a hierarchical generalized linear models (HGLM) showed promise as a method for detecting this type of differential item functioning (i.e., strong readers' mathematical

performance is different than the performance of students with lower reading ability). This suggests directions for future research that are discussed in the next section.

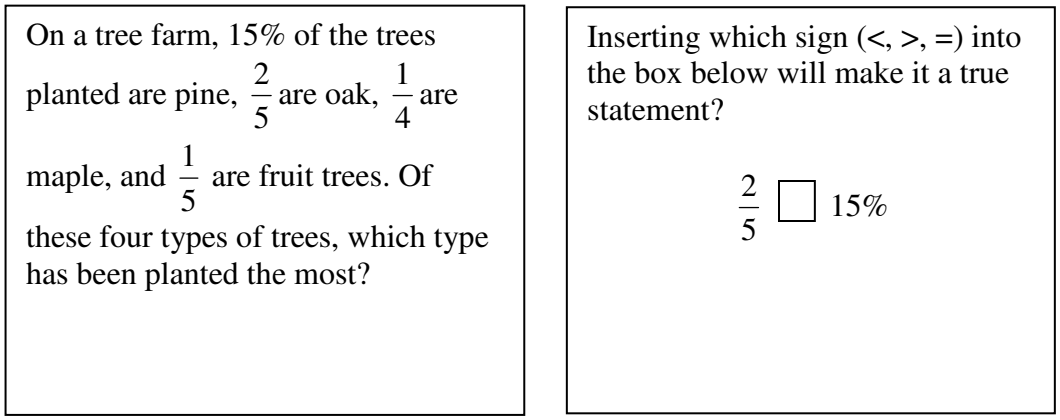


Figure 5.4. Sample Mathematics Items

Multidimensionality introduced by reading and language issues may have particular impact on English language learners. According to Hofstetter (2003), numerous factors account for the differential performance between English learners and non-English learners. For example, given equal knowledge of mathematics content and procedures, students with less proficiency in English are more likely to be assigned to a lower level mathematics class than their English peers which could limit their exposure to the mathematics content typically found on standardized tests. Depending on their level of fluency, English learners take longer to complete tests as they engage in decoding and encoding strategies between their native language and English. In general, "assessments administered in English tend to measure English learners' language proficiency rather than content knowledge" (p. 162). This likely (unintended) presence of multidimensionality attributable to language proficiency would clearly threaten the validity of the inferences based on such tests.

Refining the Definition of Dimensionality

Over 20 years ago, Drasgow and Parsons (1983) recommended viewing dimensionality as a continuum. This call was reiterated more recently by Smith, Jr. (2004) in the following statement:

> Therefore unidimensionality should not be viewed as a dichotomous yes or no decision, but rather as a continuum. A relevant research question then becomes, 'At what point on the continuum does multidimensionality threaten the interpretation of the item and person estimates?' (p. 576)

With ongoing efforts to refine the assessment methods of detecting statistical dimensionality and with advances in cognitive modeling procedures, it seems that now is the time to move away from a dichotomous view of dimensionality and move toward *detectable dimensionality* and the integration of several current research areas.

The proposed term, *detectable dimensionality*, refers to the number of dimensions such that items work together cohesively and research is moved towards constructing a theory about the learning process. Unlike Stout's definition of essential dimensionality which relies solely on a statistical model of dimensionality, the definition of detectable dimensionality requires the user to refer to the data analysis to inform the measurement process thereby moving towards a theory of content learning (Burdick, Stenner, & Kyngdon, 2007). The basis of detectable dimensionality is the observed patterns and integration of both statistical and psychological dimensionality frameworks. According to Briggs and Wilson (2004), "the art of assessing dimensionality is to find the smallest number of latent ability domains such that they are both statistically well-defined and substantively meaningful" (p. 323).

Recall from Chapter 2, a distinction is often made to the meaning of the term dimensionality. One common application of the term dimensionality refers to the number of

hypothesized psychological constructs believed to account for performance on a test (psychological dimensionality); it emphasizes the actual test content and cognitive processes required by examinees to respond to items on the test and could also be considered as the substantive hypotheses and interpretations. Another use refers to the minimum number of variables that are needed to summarize a matrix of item response data (statistical dimensionality) (Reckase, 1990). Statistical dimensionality uses quantitative analytic methods to assess the interrelationships of the item responses. However, these two definitions often result in identifying different numbers of significant dimensions. Most researchers would agree with Reckase that "psychological processes have consistently been found to be more complex than they first appear" (1997, p. 25). However, it also seems like there is a fine line between assuming too few and too many dimensions. According to Stone and Yeh (2006), "if dimensionality is overestimated, more parameters are estimated, which in turn increases estimation error" (p. 194). Similarly Tate (2003) notes, "when analyzing real test data, analysts have long recognized that the number of factors of some practical importance may often be smaller than the number that can be supported statistically" (p. 197).

Refining the definition of dimensionality to encompass both statistical and psychological substantiation has its foundation in the previous work of Camilli, Wang and Fesq (1995) who argued that statistical procedures alone (such as factor analysis) provide an incomplete conceptualization of dimensionality because dimensionality is dependent not only on the set of items and a particular set of examinees, but also on test use. Tate (2002) similarly recommended that determination of dimensionality should be guided by substantive considerations based on the content and purpose of the test. Therefore, the final assessment

of dimensionality should incorporate both judgments about test content and psychological processes, as well as, statistical evidence.

Under the definition of detectable dimensionality, the tests used in this study could be categorized as "detectably unidimensional." That is, there is evidence of a complex structure with multidimensionality but substantively there are no theoretical, a priori explanations for dimensions beyond the first. The dimensions are not consistently reproducible especially when item difficulty and reading ability are taken into account. Referring to the rope image in Figure 5.2, the rope could be seen as "detectably unidimensional" and weak to moderate amounts of multidimensionality could be described as detecting an area where the strands of the rope become more pronounced-as in a figure/ground context where the one or more strands become more salient or perceptible, yet are still a part of the whole. In other words, a distinct dimension may be a perspective issue, or may depend on how much one part of the rope is examined.

Detectable dimensionality can also be thought of in terms often used in applied statistics. Traditional investigations of dimensionality look for statistical significance (of factors, etc.) while detectable dimensionality is analogous to considering the practical significance. It integrates the statistical significance with content areas and cognitive theory. Many of the results from both the statistical models and the current cognitive diagnostic models are exploring responses at an atomic level. This is helpful and informative information and yet it still does not address whether there are truly distinct or correlated skills being measured. Detectable dimensionality allows both types of information to be placed within a content area and used to inform the measurement process.

Implications for Practice

The results of this study have several implications for test development and reporting. First, the results of this study support the use and development of vertical scaling. Inclusion of off-grade items used in the common item design does not appear to be potential sources of multidimensionality. Specifically, the results of this study showed that the inclusion of up to four common items, administered above or below one grade, does not substantially alter the dimensional structure of a test. In addition, dimensionality does not appear to be related to content strands for Grades 3-8. Thus, modest changes in the curriculum across grades, in test specifications for contiguous grade levels, or in content standards purposefully developed with the aim of vertical articulation (such as these characteristics were represented in the test development procedures for the tests studied here) should not present a major impediment to the ability to implement a vertical scale.

Second, the results of this study demonstrated a lack of relationship between dimensionality and the intended mathematical content strands. In terms of score reporting, this finding suggests that the common practice of reporting separate strand-based scores (i.e., a score for Numbers and Operations, another score for Measurement, etc.) does not have strong psychometric support. Alternatively, some researchers have recently suggested that accumulating information from items outside of those within an intended content strand shows promise as a means of enhancing the validity and utility of strand-based scores (Edwards & Vevea, 2006). Regardless of the eventual contribution of augmentation approaches, it is clear that content strands are useful for organizing curriculums and test specifications and therefore have utility independent of their dimensional structure.

The lack of relationship between dimensionality and the intended mathematical content strands suggest that the NCTM *Connections* standard may be functioning as intended. That is, the items developed for the mathematics tests used in this study appear to require students to make connections across the five different content strands. These results should encourage teachers, schools, and curriculum materials to continue to emphasize and build upon these connections to deepen students' mathematical reasoning skills and conceptual understanding. Rather than teach a skill one time and typically out of context, it should be reviewed when it comes up again and particularly when it is used in a context. For example, students learn how to add, subtract, multiple and divide integer numbers (numbers and operations strand) and are typically taught these as stand alone skills. However, working with integers becomes critical when learning to solve one- and two-step algebraic equations and integers are important when finding distances in the coordinate plane during a geometry lesson. It is important that the curriculum and textbooks work with teachers to build these connections for the students. It is also important teachers have a chance to explore these connections either with other mathematics teachers in group or lesson discussions or during professional development workshops which focus on the developmental, essentially unidimensional nature of mathematics.

The results of this study also emphasize the connectedness of mathematical topics such that knowing how mathematical skills build and relate to one another could be useful in other ways. Diagnostic information and determination of a potential need for early intervention strategies would be greatly aided by knowing how to approach mathematical skills and topics by bringing in related skills that a student better understands or feels more

confident. It is vital to prevent students from falling behind in their mathematical proficiency, becoming frustrated or math anxious or a combination thereof.

Suggestions for Future Research

The present study provides some initial answers to the questions about the relationship of five mathematical content strands and dimensionality for on-grade items and about the inclusion of off-grade level items. However, many questions remain unanswered. Therefore, additional studies should be undertaken to evaluate dimensionality of other mathematics achievement tests with and without off-grade level items. The outcomes from these additional studies will provide researchers and practioners with a better understanding of the dimensional structure of mathematics achievement tests and the relationships among mathematical skills. In addition, future studies could yield better guidelines for whether an IRT or a MIRT model is appropriate.

Reading and Mathematics

Further exploration of the reading and mathematics connection is warranted as this appears to be a multi-faceted relationship. For example, there are literature books which incorporate mathematics as part of the story and there are story or word problems in mathematics. But there are also more overlapping areas and skills such as vocabulary. While vocabulary is part of the reading and comprehension process, how does mathematical vocabulary relate to mathematical learning? Figure 5.5 illustrates different ways in which mathematical vocabulary can be embedded in a question. All three questions are asking the

student to determine the "*perimeter*" so a student has to recognize what the word perimeter means in (a) and (b) as how one calculates the perimeter. The question in (c) requires students to recognize the context in which perimeter is applicable. Item (a) however not only uses the word "rectangle" but illustrates it as well. Items (b) and (c) require that a student know the word rectangle and what it describes. In all three items, students would need to know basic properties about rectangles (i.e., four sides, pairs of congruent sides, etc.) in order to answer the questions correctly. Understanding how to *read* mathematics is an important concept in the development of mathematical learning. More dialogue and research within the content areas can help inform the "reading" necessary in mathematics.

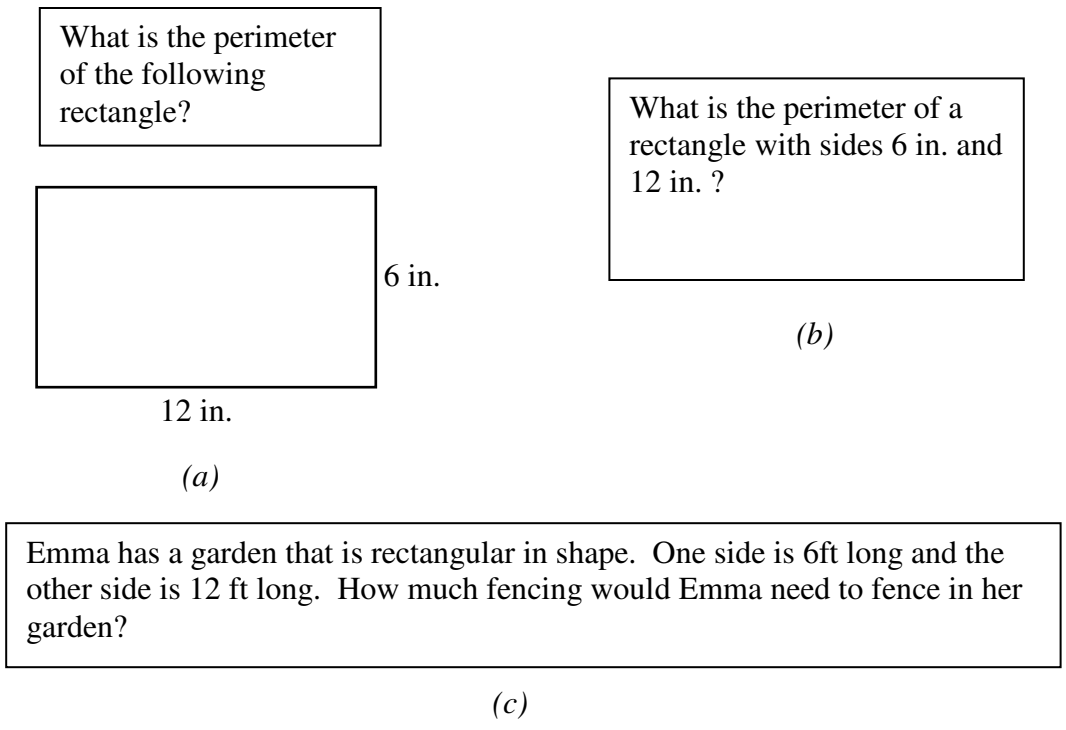


Figure 5.5. Mathematical Vocabulary Examples

Distinguishing the reading demand of mathematics items would aid in the exploration of the similarities and differences of reading and mathematical learning. Future research studies are necessary to look at the possible relationship between reading demands and dimensional structure. That is, are highly contextualized problems causing unintentional sources of dimensionality on mathematics achievement tests? Items could be rated for the amount of reading, the type of reading required (i.e., word problems, graphs, charts, etc.) and the mathematical vocabulary necessary to successfully solve a question. The data are then explored using the assessment methods of this study to determine whether dimensional differences result from items with higher reading loads. In addition, it is also necessary to investigate the potential impact of including differences in item discrimination and guessing.

Beyond Grades 3-8 Mathematics

The results of this study regarding the inclusion of off-grade level items should be extended to consider when more linking items are used. That is, what dimensional changes (if any) are introduced when the linking design requires more off-grade items to be included? It would also be interesting to consider the difficulty of the off-grade items and compare parameter estimates—is there evidence of differential item functioning (DIF)? This would be particularly relevant for the above-grade items. Is this material that an on-grade student has the mathematical background for or is it due to differences in instructional and curricular emphases?

Future studies should also extend dimensionality studies to high school and college level mathematics. There is very little research exploring the dimensional structure of upper level mathematics. For example, research is needed to assess whether an end of course type

of test administered for a Geometry course is unidimensional. Geometry courses in the past focused on formal proofs and reasoning skills. However, these skills are de-emphasized in recent books and curricula and more attention is given to problem solving. Consider the item presented in Figure 5.6. The "geometry" needed to *set up* a solution for this item is recognizing a triangle and that the sum of the interior angles of a triangle is 180°. The actual solution additionally involves algebraic skills- that is, solving the equation, $110 + 40 + x = 180$.

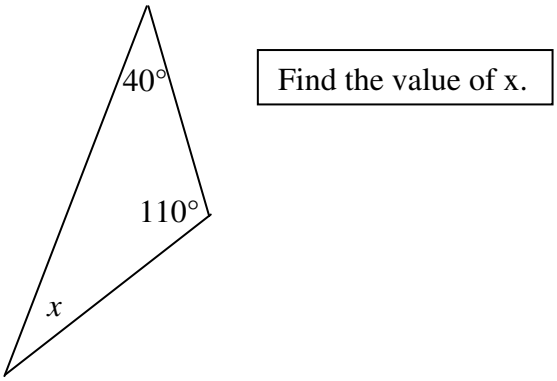


Figure 5.6. Geometry Item Example

Item format is another possible source of intentional or unintentional dimensionality. This research study was limited to a four-choice format. But as states try to create more authentic tasks for mathematical assessments, more types of formats such as gridded, open-ended responses are being used to capture process skills such as problem solving, critical reasoning and communication. Perhounkova and Dunbar (1999) used real test data with DIMTEST and Poly-DIMTEST to explore the potential influence of item format on dimensionality of tests. They found that "combining items of different formats may introduce additional complexity into the dimensionality structure of the composite test" (p.

29). In order to test students' mathematical communication skills, the state of Washington has included items on the Washington Assessment of Students' Learning (WASL) which require students to write responses to mathematical problems. Walker and Beretvas (2000) have found that these types of tests are multidimensional: "one representing an examinee's ability to communicate about mathematics and another representing an examinee's ability to solve mathematical problems" (p.7). Additionally, van der Linden and Hambleton (1997) commented that the inclusion of polytomous response data are not the only feature introduced by a new format. They suggest that "many of these new formats often require examinees to use more than one skill such as problem solving, critical thinking, reasoning, organization and writing" (p. 221). There has also been an ongoing line of research indicates that a change in test format actually changes the measured construct (see Perhounkova and Dunbar, 1999). These are important questions and issues that impact test development, particularly what is taken as validity evidence. Much research has been done inquiring about the teaching and learning of mathematics. It is important that assessment methods integrate these ideas and concepts in item development.

Better understanding of mathematical thinking and types of assessment can also extend beyond mathematics into other content areas such as science. Science is not only similar to mathematics by the use of specific content vocabulary (i.e., density, velocity, etc.) and the use of contextualized/ decontextualized items, it also incorporates many mathematical and reading skills. Future studies are needed to better understanding the dimensional structure of science test data especially how it relates to the mathematical and reading demands of the items.

Modeling and Assessing Dimensionality

Continued work on methods for determining and assessing statistical dimensionality is needed. There is evidence that unidimensional IRT models are robust to some departure from unidimensionality but further research is needed to determine how much departure is unacceptable. Ongoing research is also needed to further explore the application and use of MIRT models. In a recent presentation, Martineau and his colleagues (Martineau et al., 2006) suggested that while truly unidimensional data are rarely observed in educational achievement tests, MIRT is not a useful choice either: despite its 30+ years of research, MIRT has seen negligible application in educational achievement testing contexts; it is often considered to be impractical due to its relatively higher cost and availability of software; replication is a problem; and difficulties exist in interpretability of MIRT results. Specifically, the complexity and the uncertainty about the definition of a dimension in MIRT models has caused some researchers to contend that MIRT cannot be applied in practical testing situations (Kirisci et al., 2001; Luecht & Miller, 1992).

Perhaps the greatest impediments of applying MIRT models are score interpretation (van Abswoude et al., 2004) and difficulty of linking tests that measure composite abilities. If multidimensional models are used, then the single score represents a composite of abilities; and thus linking equivalent forms of a test (i.e., equating) or different forms of a test across grades (i.e., vertical scaling) is not feasible. Further, it appears that the development of test items (and test forms) that measure the same composite of abilities is currently an unproven goal.

Finally, ongoing research should continue to suggest new and better ways to measure dimensionality. For example, Bejar (1983) expanded the definition of unidimensionality to

include items functioning in unison. Burdick, Stenner and Kyngdon (Burdick et al., 2007) have recently begun work on a similar type model where dimensionality is defined by items that rank order students in the same way. More research is also needed to better understand the correlation of dimensions and its affect on measuring and detecting one or more dimensions. In particular, future research studies should address the correlation requirements needed for detectable dimensionality.

Conclusions

This research study, like other studies involving educational data, shows how important the assessment of dimensionality is to a testing program and yet how intricate and complex the task is. It does not however preclude a testing program from periodically assessing "whether the test assembly process is producing tests that are in accord with the test construction blueprint" (Dorans & Lawrence, 1999, p.5) or from conducting periodic checks of the stability of a common scale over time as proposed in Standard 4.17 of the *Standards for Educational and Psychological Testing* (American Educational Research Association et al., 1999). Detectable dimensionality integrates two important characterizations of dimensionality: psychological meaning and statistical fit. It is only when these two components support one another that the true test structure can be assessed and interpreted and perhaps more importantly that the implications for the educational process be clarified.

APPENDIX A: PROGRAMS FOR ASSESSING TEST DIMENSIONALITY

This appendix provides greater detail about the technical information of approaches and software programs used to assess test dimensionality described in Chapters 2 and 3: item factor analysis (NOHARM), principal component analysis (WINSTEPS), assessment of essential dimensionality (DIMTEST), and exploring the conditional covariances (DETECT). All four approaches have been shown to be effective indices of dimensional structure. NOHARM and WINSTEPS are parametric methods and DIMTEST and DETECT are nonparametric methods. The difference between parametric and nonparametric is the specification of the item response function. In IRT, the probability of success on item *i* is usually presented as $P_i(\theta)$. This function is known as the *item response function* (IRF). Parametric methods assume a particular parametric model for the IRF. Nonparametric methods assume only that the IRF is monotonic.

NOHARM

One of the most widely used nonlinear factor-analytic approaches is the Normal-Ogive Harmonic Analysis Robust Method (NOHARM). NOHARM refers both to a model that was developed by McDonald (McDonald, 1967) and to a program written by Fraser and McDonald (Fraser & McDonald, 1988) which uses the NOHARM model. The model can be presented as either the latent trait or the common factor parameterization. The program was written to fit unidimensional and multidimensional normal ogive models of latent trait theory to dichotomous data, as presented by McDonald (McDonald, 1967). NOHARM can be run either in exploratory or confirmatory mode and provides Varimax and Promax rotated factor solutions. It does not used tetrachoric correlations but instead minimizes the unweighted

least squares (ULS) difference among observed proportions that pairs of items are passed and expectations based on a third-degree polynomial function implied by the factor model. NOHARM outputs a residual matrix of differences among observed and expected proportions, as well as the root mean square residual(RMSR) as an overall index of model fit (Stone & Yeh, 2006).

NOHARM uses a *k*-dimensional normal ogive model and is given by the following equation:

$$P = N\{\beta_{i0} + \beta_{i1}\theta_1 + \beta_{i2}\theta_2 + \cdots + \beta_{ik}\theta_k\}$$

where

N is the normal ogive function,

θ is the latent ability the vector,

$\beta_{i0} = \frac{t_i}{\sqrt{\psi_i}}$ and

for the *k*th dimension $\beta_i = \frac{\lambda_i}{\sqrt{\psi_i}}$.

Note that in the equations above, *t* is the threshold value such that if an examinee's proficiency is beyond *t* then they will get the item correct; if not, the item will be incorrect. λ is the common factor loading and ψ is the explained item variance.

The NOHARM output file includes several sections. In the first section NOHARM summarizes the input data such as the title and the number of items, dimensions, and subjects. It also includes the sample correlation matrix, the fixed guessing parameters, pattern matrices and initial value matrices. The second section contains results for both the latent trait and common factor parameterization. The results of the latent trait parameterization include the following:

- final item parameter estimates (the item location, β_{i0} and item discriminations β_i),
- correlations among factors,
- the residual matrix,
- two summaries of the matrix (sum of squares of the residuals and the RMSR), and
- Tanaka's unweighted least squares goodness of fit index (Tanaka, 1993).

The common factor parameterization for the factor-analytic model is defined as:

$$y_i = \lambda_{i1}\theta_1 + \ldots \lambda_{ik}\theta_k + \delta_i$$

where

y_i is conceptualized as a continuous latent response pr opensity (for each item score there exists an underlying item-specific threshold that corresponds to the difficulty level of the item where the examinee must exceed this threshold to get the item correct),

$\lambda_i = [\lambda_1, \lambda_2, \ldots, \lambda_k]$ is the factor-loading vector,

$\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_k)$ is the examine trait vector having mean 0 and covariance Φ and,

δ_i is a residual term distributed (0, ψ_i).

Referring to the common factor model, NOHARM also includes the following values for the common factor model reparameterization:

- threshold values,
- unique variances,
- factor loadings and
- the Varimax (orthogonal) and Promax (oblique) factor loadings and factor correlations (exploratory mode only).

One of the advantages of the nonlinear factor analytic approach is the interpretative assistance. For example, the λs estimated by NOHARM can be interpreted as factor loadings which can be used to identify those items that appear to cluster together. This in turn can be helpful in identifying the nature of the underlying latent trait being measured by the items. The proportion of *y*s (i.e., common factor model) not accounted for by the dimensions is represented by the ψs, a measure of the item uniqueness. The *t*s can aid in the understanding of item difficulty. The inverse normal transformation of the item difficulty level is represented by $t=N^{-1}(p_i)$ meaning that positive *t*s indicate easier items and negative *t*s represent harder items.

The following methods are available for assessing model-data fit (Ackerman et al., 2003, p. 44):

1) Difference between observed and reproduced correlation matrix is small producing small sum of squares of residuals and root mean square residual (RMSR).

2) Compare unidimensional model to multidimensional model by comparing the residuals of each model
3) Chi-square fit statistic (also based on residual matrix)

The chi-square fit statistic is based on testing the null hypothesis that the off-diagonal elements in the residual correlation matrix produced by the factor analysis are equal to zero (Gessaroli & De Champlain, 1996). If the null is not rejected, the fitted model adequately approximates the observed correlations among the items. Therefore if the fitted model was unidimensional, then the null hypothesis of unidimensionality would not be rejected.

WINSTEPS

WINSTEPS is based upon a Rasch model and uses joint maximum likelihood estimation (JMLE) procedures to estimate item and person parameters. JMLE is more flexible for missing data than is conditional maximum likelihood estimate (CMLE) or modified maximum likelihood estimation (MMLE) (Linacre, 2005); it does not assume a person distribution. WINSTEPS begins with a central estimate for each person measure and item calibration unless predetermined values are provided by the analyst. An iterative version of the PROX (normal approximation) algorithm is used to reach a rough convergence to the observed data pattern. The JMLE method is then implemented to refine the estimates using proportional curve fitting.

The Rasch model constructs a one dimensional measurement system regardless of the dimensionality of the data (Linacre, 1998). Ideally if the unidimensional Rasch model fits well, then all the information in the data would be explained by the single latent variable. The residuals (the differences between what a model predicts and what is observed) could

then be considered noise, would be independent of each other and when standardized would follow a normal distribution. Therefore all elements of an inter-item residual correlation matrix would be zero if the data fit the model. WINSTEPS asserts that all the residual variance is due to common factors, and places 1's in the diagonal of the inter-items residual correlation matrix, and the empirical correlations among the standardized residuals in the off-diagonal elements (Linacre, 2005).

In order to check that all items share the same dimension, WINSTEPS identifies substructures in the data by performing a principal-components/contrast decomposition of the observation residuals (Linacre, 2005). *Principal components analysis* (PCA) is a technique for simplifying a dataset. It is a linear transformation that transforms the data to a new coordinate system such that the greatest variance by any projection of the data aligns on the first coordinate (called the first principal component), the second greatest variance on the second coordinate, and so on. The manual for WINSTEPS cautions users that Rasch-residual-based PCA is not to be interpreted as a usual factor analysis. "The [PCA] components show contrasts between opposing factors, not loadings on one factor" like factor analysis (p. 261). In typical factor analysis, the researcher is looking for shared factors and to assign the items to the factors in a way that is as meaningful as possible; it is aimed at explaining common variance. The purpose of PCA is not to construct variables but to explain total variance. The Rasch dimension is hypothesized to be the first dimension. The residuals are then analyzed; the researcher is looking for the contrast in the residuals that explains the most variance. If the contrast is very weak (i.e., noise), then there is no second dimension. If there is structure to the residuals, then the contrast is considered the second dimension in the data and similar procedures are followed for exploring for a third

dimension, etc. In Rasch analysis, it is hoped that contrasts are not found and if there are, then the fewest number of contrasts are desired.

In PCA, components are assigned eigenvalues. Basically, the eigenvalues of an inter-item correlation matrix is often used as an indication fo the number of factors underlying the item responses. More specifically, "an eigenvalue is equal to the sum of the squared loadings of the indicators on the component or the factor with which the eigenvalue is associated" (Pedhazur & Schmelkin, 1991). The variance that the solution accounts for is associated with the eigenvalue. Simulation studies indicate the eigenvalues less that 1.4 are at the random level (i.e. noise) (Smith & Miao, 1994) or can sometimes be as high as 2.0 (Raiche, 2005). In addition, "Ben Wright recommends that the analyst split the test into two halves, assigning the items, top vs. bottom of the first component in the residuals...cross-plot the person measures. If the plot would lead you to different conclusions about the persons depending on the test half, then there is a multidimensionality" (Linacre, 2005, p. 266).

DIMTEST

DIMTEST is an asymptotically justified non-parametric procedure that provides a test of hypothesis of unidimensionality of a test data set. The program was developed and written by Nandakumar and Stout (1993) and is based upon Stout's concept of essential dimensionality (Stipek, 1987; Stout, 1990) which emerges from the theory of essential local independence (Nandakumar, 1991). The use of DIMTEST however does not require acceptance of Stout's concept of essential dimensionality –it can be viewed as a technique to detect sizable lack of fit of a locally independent unidimensional latent trait model (Nandakumar & Stout, 1993). "A test is considered essentially unidimensional when the

average between-item residual covariances after fitting a one-factor model approaches zero as the length of the test increases" (Embretson & Reise, 2000, p. 230). Since its initial development, DIMTEST has undergone two major revisions: the first by (Nandakumar & Stout, 1993) and the second by Froelich and Habing (2001). This most recent version will be used in the analysis.

Essential dimensionality is based upon the concept of essential independence. An item pool *U* is said to be *essentially independent* (EI) with respect to the latent variable θ, if *U* satisfies

$$D_N(\theta) = \frac{\sum_{1 \le i \le j \le N} \left| Cov\left(U_i, U_j \middle| \Theta = \theta\right) \right|}{\binom{N}{2}} \to 0 \; as \; N \to \infty$$

The essential dimensionality (d_E) of an item pool U is then defined as the minimal dimensionality necessary to satisfy the assumption of essential independence. When d_E=1, essential unidimensionality is said to hold. Essential unidimensionality holds when only one dominant dimension influences the examinees performance on a set of test items.

DIMTEST assesses the relationship among subsets of items based on conditional item covariances. A small subtest of items is referred to as the assessment subtest (AT) because its responses will be used to assess the test's dimensionality. The larger set of remaining items is used to partition the examinees into groups for a stratified analysis and is referred to as the partitioning subtest (PT). If a test is unidimensional then the conditional covariance between any two items on as the AT is zero after conditioning on the PT. If the conditional covariance between any two items on the AT is greater than zero after conditioning on the PT, then a test is multidimensional. Note that testing whether the conditional covariance

between any two items is zero is analogous to testing the assumption of weak local independence (Gierl et al., 2005).

Based upon their PT score, each examinee is assigned to one of K subgroups. Two variance estimates, the total variance estimate and the "unidimensional" variance estimate, are computed using items on the AT (Nandakumar & Stout, 1993). Or in terms of the conditional covariances, the variance difference between the total test variability (σ^2_X) and item variability for examinees with the same score, *k*, on the PT (Gierl et al., 2005) is:

$$\sigma_X^2 = \sum_{i=1}^{N_I} p_i q_i + 2\sum_{i<l} Cov(U_i, U_l)$$

Rearranging terms,

$$\sum_{i<l} Cov(U_i, U_l) = \frac{\sigma_X^2 - \sum_{i=1}^{N_I} p_i q_i}{2}$$

By calculating the covariance for examinees with the same score, *k*, on the PT, the conditional covariances can be calculated as shown in the following equation:

$$T_{L,k} = \sum_{i<l,k} Cov(U_{i,k}, U_{l,k}) = \frac{\sigma_{X,k}^2 - \sum_{i=1}^{N_I} p_{i,k} q_{i,k}}{2}$$

The difference in these variance estimates is then normalized by an appropriate normalizing constant, S^2_k (the asymptotic variance of $T_{L,k}$), and summed over the subgroups to obtain the statistic, T_L. In other words, T_L is based on the sum of the estimated conditional covariances among the AT items for examinees that have obtained the same score, *k*, on the PT items. Specifically,

$$T_L = \frac{\sum_{k=1}^{K} T_{L,k}}{\sqrt{\sum_{k=1}^{K} S_k^2}}$$

The DIMTEST statistic is then defined as

$T = \frac{T_L - \overline{T}_B}{\sqrt{2}}$ where $\overline{T}_B$ corrects for bias introduced by a finite length test.

The test statistic, T, represents the degree of dimensional distinctiveness of the two clusters of items and is based on the fundamental principle that local independence should hold approximately when sampling from a subpopulation of examinees of approximately equal θ level (Hattie et al., 1996). Under typical conditions, the original DIMTEST statistic T (Stout, 1987) and the more powerful T' (Nandakumar & Stout, 1993) are distributed asymptotically standard normal when a test is unidimensional (van Abswoude et al., 2004). Therefore, given a significance level α and the upper 100(1-α) percentile of a standard normal distribution, Z_α, the null hypothesis of d_E=1is rejected if T'> Z_α.

Recall that DIMTEST requires two subtests, AT and PT. If DIMTEST is being used in a confirmatory analysis, the user selects items for the AT based on prior expectations such as test specifications or content strands. When operating DIMTEST in an exploratory mode, a method call ATFIND identifies items for the AT by using non-parametric conditional covariance dimensionality programs DETECT and HCA/CCPROX (The William Stout Institute for Measurement, 2005). ATFIND generates four output files: TEMP.OUT, PROX.OUT, HCA.OUT, and ATLIST.IN. ATLIST.IN is needed to tell DIMTEST which items are in the AT. The TEMP.OUT file contains a summary of the CCPROX analysis while the PROX.OUT file reports the conditional-covariance based proximity measures for

the item pairs as calculated by CCPROX. This output could be used as input for the HCA cluster analysis program. The HCA. OUT file is generated by the HCA cluster analysis program and contains a list of item clusters starting with the smaller cluster of items and ending with the largest cluster of items that are closer in dimensionality.

Once the ATLIST.IN file is either generated or user specified, DIMTEST can be run. There are two output files, DIMTEST.OUT and KERNPTS. DIMTEST.OUT contains a summary of the input, specifications, etc. and the final values for T_L, $T_{B\ bar}$, as well as the resulting T and its p-value. The KERNPTS provides estimated unidimensional item response function (IRF) for every item and can be used for diagnostic purposes to check whether the estimated IRFs seem reasonable in cases where the researcher suspects a problem.

DETECT

Another nonparametric approach to assessing dimensionality is the Dimensionality Evaluation to Enumerate Contributing Traits (DETECT) index and program (Kim, 1994; Stout et al., 1996; Zhang & Stout, 1999). The DETECT index was proposed by Kim (1994) to be data-driven index of dimensionality that would identify the number of distinct latent dimensions, estimate the amount of test multidimensionality and assign items to appropriate homogenous clusters when approximate simple structure exists. DETECT relies on the covariances of items conditioned on an estimate of the test composite ability. Test composite (θ_{TT}) is defined to be a particular linear combination of the test's complete latent trait variables (Zhang & Stout, 1999).

The DETECT method searches for a good or best (if it exists) choice of partitioning the test items into dimensionally homogenous clusters that maximize the DETECT index,

D(P). When the test exhibits approximate simple structure, the number of cluster resulting from the optimal partition will be equal to the number of dominant dimensions. The value of the index represents the magnitude of departure from being perfectly fitted by a unidimensional model. It is created by computing all item covariances after conditioning on the examinees' scores using the remaining items and can be computed:

$$D(P) = \frac{2}{n(n-1)} \sum_{1 \le i \le j \le N} \delta_{ij}(P) E\left[Cov(X_i, X_j \mid \Theta_{TT})\right],$$

where n is the number of dichotomous items on a test, P denotes the partitioning of n items into k clusters, Θ_{TT} is the test composite, X_i and X_j are scores on items i and j, and

$$\delta_{ij}(P) = \begin{cases} 1, if\ items\ X_i\ and\ X_j\ are\ in\ the\ same\ cluster\ P \\ -1, otherwise \end{cases}$$

The covariance is estimated using a contingency table approach that assumes no particular parametric form of the IRF (Finch & Habing, 2005). Checking of each possible partitioning of items would be computationally intensive so DETECT begins with the set of partitions generated using the HCA/CCPROX procedure and then uses a genetic algorithm to search for the maximum $D(P^*)$ or D_{Max} where P^* refers to the partition that maximizes *D(P)*. For unidimensional data, the conditional covariances of the homogeneous item clusters will be positive while the not particularly homogenous items will contribute negative values and thus, the resulting $D(P^*)$ index will be close to zero. If the underlying structure of the data is more multidimensional, the positive within-cluster conditional covariances and the negative between-cluster conditional covariances result in a $D(P^*)$ index that is greater than zero (Gierl et al., 2005). Kim (1994) suggests that a $D(P^*)$ index less than 0.10 indicates that the data can be considered unidimensional; an index between 0.10 and 0.50 suggests a weak amount of dimensionality; an index between 0.51 and 1.00 is considered a moderate amount

of dimensionality and an index greater than 1.00 would indicate a strong amount of dimensionality.

After reaching the search's stopping rule, DETECT reports the number of clusters in the final solution (equal to the number of dimensions), item membership for each cluster, the DETECT index and another index, *r*, which represents how well the underlying structure of the data approximates a simple structure. It is defined as,

$$r_{Max} = \frac{D(P^*)}{\tilde{D}(P^*)} \text{ where } \tilde{D}(P^*) = \frac{2}{n(n-1)} \sum_{1 \le i \le j \le N} \left| E\left[Cov(X_i, X_j \mid \Theta_{TT}) \right] \right|.$$

It compares the maximum value of the partition to the average absolute value for the conditional covariance across all item combinations (Gierl et al., 2005). An approximate simple structure will result in values of *r* greater than 0.80 and a complex structure is suggested by *r* values less than 0.80 (Kim, 1994). Simulations studies have shown that DETECT correctly identifies the correct partition when *r* is greater than 0.80 (i..e. the data display approximate simple structure) but when *r* is less that 0.80 (i.e. complex structure) the results and interpretation become unclear (Gierl et al., 2005; Zhang & Stout, 1999). van Abswoude, van der Ark and Sijtsma compared DETECT to several other methods for determining the dimensionality of item response data (2004). Their results indicated that DETECT was superior to Mokken Scale Analysis for Polytomous Items (MSP) and in most cases to HCA/CCPROX in retrieving simulated data structure . van Abswoude, van der Ark and Sijtsma also found that DETECT may be more effective in larger samples and is influenced by the number of items in a cluster assessing one trait.

APPENDIX B: RESULTS FROM DETECT

Table B.1 *Distribution of Strand-Designated Items by Cluster and Content Strand for Grade 4*

	Distribution of Strand-Designated Items by Cluster					
Content Strand	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Total
Numbers & Operations	3	3	1	0	na	7
Geometry	2	1	0	3	na	6
Algebra & Patterns	0	0	2	0	na	2
Data Analysis & Probability	3	0	2	0	na	5
Measurement	3	0	1	1	na	5
Total Number of Items in Cluster	11	4	6	4	0	25

Table B.2 *Distribution of Strand-Designated Items by Cluster and Content Strand for Grade 5*

	Distribution of Strand-Designated Items by Cluster					
Content Strand	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Total
Numbers & Operations	4	2	1	0	0	7
Geometry	0	1	0	3	0	4
Algebra & Patterns	1	1	2	1	0	5
Data Analysis & Probability	0	0	0	2	1	3
Measurement	3	0	0	2	0	5
Total Number of Items in Cluster	8	4	3	8	1	24

Table B.3 *Distribution of Strand-Designated Items by Cluster and Content Strand for Grade 6*

	Distribution of Strand-Designated Items by Cluster					
Content Strand	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Total
Numbers & Operations	1	2	2	1	2	8
Geometry	1	0	1	0	2	4
Algebra & Patterns	0	1	0	0	2	3
Data Analysis & Probability	0	0	2	1	1	4
Measurement	0	1	1	1	2	5
Total Number of Items in Cluster	2	4	6	3	9	24

Table B.4 *Distribution of Strand-Designated Items by Cluster and Content Strand for Grade 7*

	Distribution of Strand-Designated Items by Cluster					
Content Strand	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Total
Numbers & Operations	4	2	0	0	na	6
Geometry	2	1	1	0	na	4
Algebra & Patterns	1	2	0	2	na	5
Data Analysis & Probability	3	1	0	0	na	4
Measurement	2	2	1	0	na	5
Total Number of Items in Cluster	12	8	2	2	0	24

Table B.5 *Distribution of Strand-Designated Items by Cluster and Content Strand for Grade 8*

	Distribution of Strand-Designated Items by Cluster					
Content Strand	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Total
Numbers & Operations	3	1	0	0	0	4
Geometry	3	2	1	0	0	6
Algebra & Patterns	2	2	1	3	0	8
Data Analysis & Probability	0	0	1	0	1	2
Measurement	1	1	1	1	0	4
Total Number of Items in Cluster	9	6	4	4	1	24

APPENDIX C: NONLINEAR ITEM FACTOR ANALYSIS (NOHARM) FACTOR LOADINGS FOR ON-GRADE ITEMS (5-DIMENSIONS)

Table C.1 *Nonlinear Item Factor Analysis (NOHARM) Factor Loadings for Grade 4*

		Varimax Rotated Factor Loadings				
Strand	Item #	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
Numbers & Operations	1	**0.598**	0.126	0.097	-0.118	-0.003
	2	**0.525**	0.168	0.055	-0.102	-0.056
	3	**0.220**	-0.125	0.045	-0.102	0.177
	4	0.173	-0.048	**0.310**	-0.027	0.039
	5	**-0.309**	0.196	-0.142	-0.085	0.038
	6	0.159	-0.296	**0.520**	-0.133	0.101
	7	0.306	0.162	**-0.554**	0.072	0.342
Geometry	9	**0.497**	-0.033	-0.319	0.045	0.021
	10	-0.033	**-0.216**	0.050	0.033	-0.088
	12	0.016	0.061	**0.302**	0.049	-0.070
	13	**0.269**	0.258	0.128	0.111	-0.206
	14	0.190	-0.051	0.149	**0.966**	0.077
	15	0.042	-0.043	**-0.255**	-0.078	0.008
Algebra & Patterns	17	0.245	**0.931**	0.254	0.018	0.090
	18	0.170	0.188	-0.033	-0.053	**0.203**
Data Analysis & Probability	20	**0.710**	0.172	-0.142	0.075	0.302
	21	**0.887**	-0.069	-0.420	0.112	0.140
	22	0.116	-0.003	-0.054	0.068	**0.243**
	23	-0.143	0.056	0.173	-0.031	**0.227**
	24	0.087	0.002	**-0.159**	-0.125	0.023
Measurement	25	**0.446**	0.030	0.046	0.005	0.031
	26	**0.476**	0.217	-0.090	0.102	0.146
	27	-0.170	0.166	-0.007	0.066	**0.683**
	28	**0.560**	0.058	0.001	0.042	-0.040
	30	-0.162	-0.053	0.200	0.148	**-0.298**
Total	25	11	2	6	1	5

Table C.2 *Nonlinear Item Factor Analysis (NOHARM) Factor Loadings for Grade 5*

		Varimax Rotated Factor Loadings				
Strand	Item #	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
Numbers & Operations	1	**0.528**	0.021	0.073	0.064	0.009
	2	**0.506**	0.218	0.186	0.030	0.198
	3	0.302	0.154	-0.094	0.005	**0.416**
	4	0.103	0.002	0.188	**0.244**	0.050
	5	**0.345**	0.255	0.264	0.072	0.021
	7	0.141	0.247	-0.015	**0.340**	0.271
	8	0.199	0.013	-0.021	**0.743**	-0.067
Geometry	10	-0.221	0.202	0.011	**0.551**	0.254
	11	-0.069	**0.526**	0.152	0.393	0.070
	12	0.153	**0.751**	0.232	-0.060	-0.100
	14	0.038	0.330	**0.482**	0.222	0.006
Algebra & Patterns	17	-0.077	**0.494**	0.158	0.206	0.299
	18	0.074	**0.200**	0.014	0.004	0.068
	19	0.312	**0.374**	0.103	0.112	0.229
	20	0.108	0.151	0.294	**0.359**	0.069
	21	0.197	**0.291**	-0.013	0.116	0.077
Data Analysis & Probability	22	0.169	**0.282**	0.155	0.091	0.103
	24	0.001	**0.341**	0.333	0.097	0.243
	25	0.063	0.010	**0.403**	-0.074	0.073
Measurement	26	0.202	0.237	0.196	0.145	**0.282**
	27	0.239	0.281	0.089	**0.307**	0.271
	28	0.314	**0.357**	0.115	0.209	0.088
	29	0.004	0.045	0.461	0.160	**0.600**
	30	0.190	0.211	**0.626**	0.144	-0.032
Total	24	3	9	3	6	3

Table C.3 *Nonlinear Item Factor Analysis (NOHARM) Factor Loadings for Grade 6*

		Varimax Rotated Factor Loadings				
Strand	Item #	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
Numbers and Operations	1	**0.327**	-0.055	0.124	0.024	0.017
	2	-0.063	**0.741**	0.12	0.03	0.038
	3	**0.426**	0.008	-0.179	0.017	0.042
	4	0.198	**0.375**	-0.062	-0.158	0.221
	5	-0.077	0.414	-0.074	-0.098	**0.595**
	6	**0.296**	0.262	0.261	-0.127	0.182
	8	0.25	**0.332**	0.102	0.173	0.162
	10	-0.013	0.084	0.091	-0.106	**0.647**
Geometry	11	0.274	0.123	0.081	0.145	**0.527**
	12	**0.45**	0.02	-0.025	0.105	0.144
	13	**0.208**	0.028	0.31	-0.09	0.094
	14	-0.019	-0.054	-0.421	0.17	**0.446**
Algebra & Patterns	15	0.095	0.374	0.126	-0.075	**0.457**
	16	0.154	0.071	0.042	-0.026	**0.488**
	19	0.037	**0.506**	-0.055	0.088	0.259
Data Analysis & Probability	20	-0.103	0.057	**0.663**	0.177	0.462
	22	**0.454**	0.258	0.051	-0.245	0.067
	23	**0.339**	0.278	0.059	0.087	0.295
	24	0.319	0.268	0.106	-0.125	**0.358**
Measurement	25	0.134	0.083	0.09	0.014	**0.212**
	26	0.284	0.208	0.136	**0.799**	-0.026
	27	**0.512**	0.109	0.214	0.169	0.046
	28	0.003	0.062	0.106	**-0.516**	0.085
	30	0.112	0.135	0.074	-0.119	**0.46**
Total	24	8	4	1	2	9

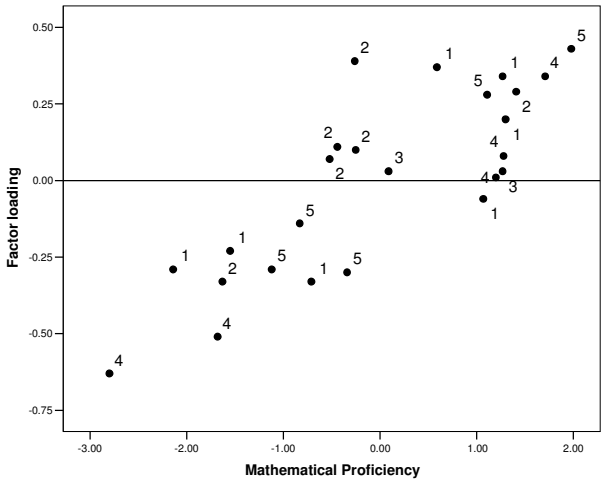
Table C.4 *Nonlinear Item Factor Analysis (NOHARM) Factor Loadings for Grade 7*

Strand	Item #	Varimax Rotated Factor Loadings				
		Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
Numbers and Operations	1	**0.467**	0.29	0.072	-0.216	0.053
	2	0.03	**0.194**	-0.012	0.045	-0.002
	3	0.222	**0.287**	-0.278	0.127	-0.007
	4	**0.483**	-0.127	0.205	0.239	0.394
	5	**0.389**	-0.083	0.373	0.023	0.275
	7	0.447	0.165	**0.499**	-0.365	0.185
Geometry	9	**0.641**	-0.214	0.212	0.083	0.11
	10	0.379	-0.106	**0.724**	0.081	-0.176
	11	0.106	0.166	0.097	**0.469**	0.33
	13	0.059	0.174	0.201	**0.226**	0.045
Algebra & Patterns	15	-0.035	**0.141**	0.013	-0.007	0.024
	16	-0.127	**0.498**	-0.234	-0.063	0.051
	17	0.124	-0.097	0.082	0.076	**-0.469**
	18	**0.448**	0.085	0.098	-0.008	-0.168
	19	**0.595**	0.098	0.188	0.037	-0.101
Data Analysis & Probability	20	0.206	-0.13	**0.457**	-0.215	-0.01
	21	**0.314**	-0.017	0.171	0.25	-0.099
	22	0.066	0.034	**0.601**	0.126	0.131
	23	0.241	**0.467**	0.256	0.326	-0.187
Measurement	24	0.179	0.072	**0.442**	0.29	-0.149
	25	**-0.311**	0.212	-0.058	0.029	0.037
	27	0.027	**0.364**	0.207	0.219	0.011
	29	0.304	0.1	**0.475**	0.063	-0.205
	30	-0.033	0.017	-0.02	**0.175**	-0.032
Total	24	8	6	6	3	1

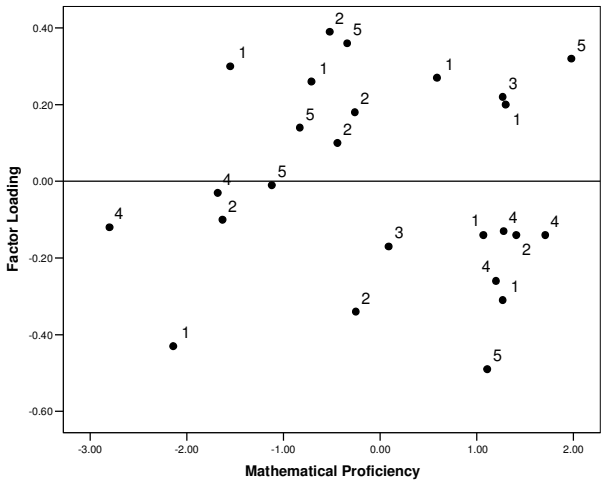
Table C.5 *Nonlinear Item Factor Analysis (NOHARM) Factor Loadings for Grade 8*

Strand	Item #	Varimax Rotated Factor Loadings				
		Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
Numbers & Operations	1	**0.556**	0.216	0.083	0.063	-0.032
	2	0.171	**0.615**	-0.104	0.164	0.014
	3	0.264	**0.542**	0.112	0.065	-0.050
	4	0.207	0.325	0.115	**0.630**	0.109
Geometry	7	0.189	-0.024	-0.070	**0.584**	-0.079
	8	0.273	**0.581**	-0.061	0.062	-0.132
	10	**0.443**	0.184	0.164	0.054	0.030
	11	**0.418**	0.121	-0.057	0.099	0.166
	12	**0.736**	0.081	-0.034	0.293	0.058
	13	0.239	**0.619**	0.310	0.050	0.017
Algebra & Patterns	14	0.284	0.055	**0.318**	0.137	-0.064
	15	0.297	0.241	**0.516**	0.093	0.199
	16	-0.034	-0.092	**-0.370**	0.109	0.043
	17	0.186	0.060	-0.027	**-0.289**	0.167
	18	-0.002	0.147	-0.120	**0.579**	0.429
	20	**0.528**	0.014	0.281	-0.048	0.307
	21	**0.316**	0.242	-0.244	-0.062	0.290
	22	**0.398**	0.213	0.201	-0.007	0.031
Data Analysis & Probability	24	-0.121	-0.036	0.194	-0.058	**0.338**
	25	0.188	-0.005	-0.131	0.132	**0.552**
Measurement	27	0.114	**0.562**	0.168	0.091	0.244
	28	0.181	0.025	0.007	-0.060	**0.506**
	29	0.027	**-0.454**	-0.185	0.191	-0.056
	30	**0.358**	0.123	0.027	-0.061	0.091
Total	24	8	6	3	4	3

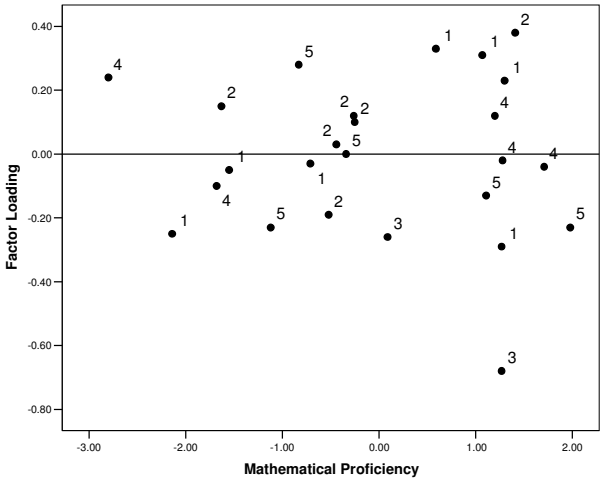
APPENDIX D: PRINCIPAL COMPONENTS (STANDARDIZED RESIDUAL) FACTOR PLOTS FOR ON-GRADE ITEMS



(a) First Factor

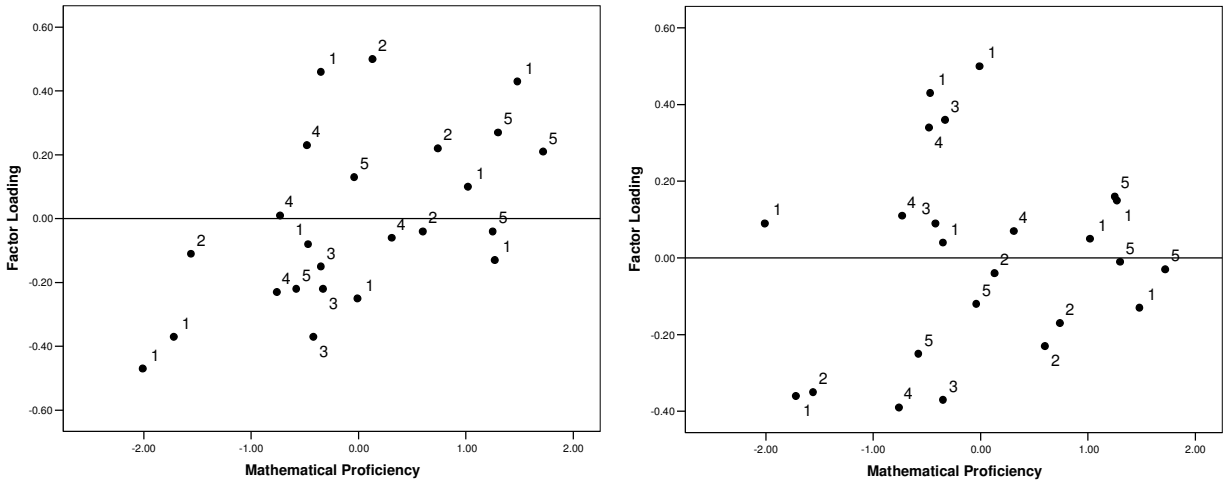


(b) Second Factor



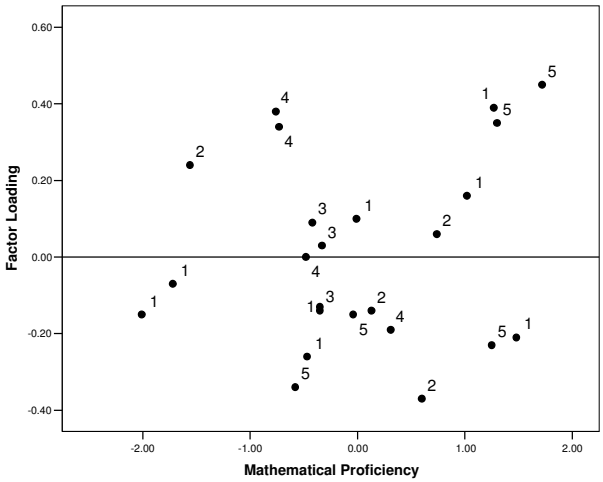
(c) Third Factor

Figure D.1. Principal Components (Standardized Residual) Factor Plots of Grade 4 On-Grade Items



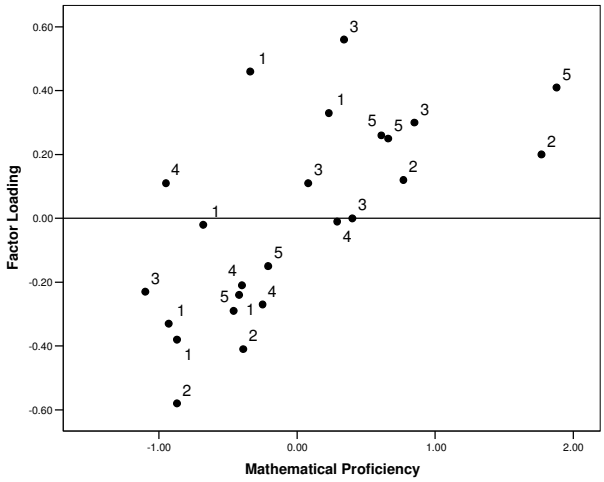
(a) First Factor

(b) Second Factor

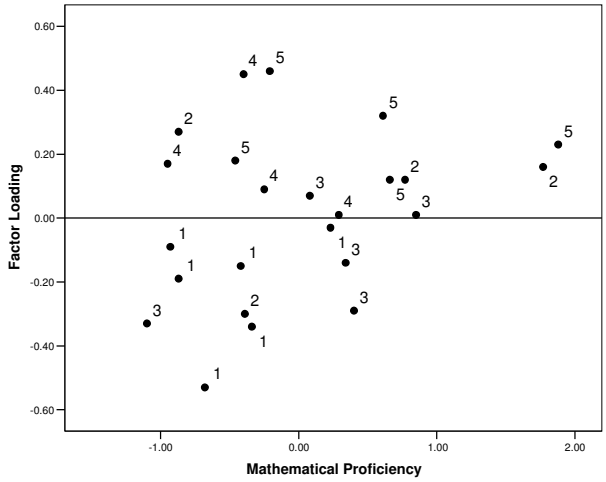


(c) Third Factor

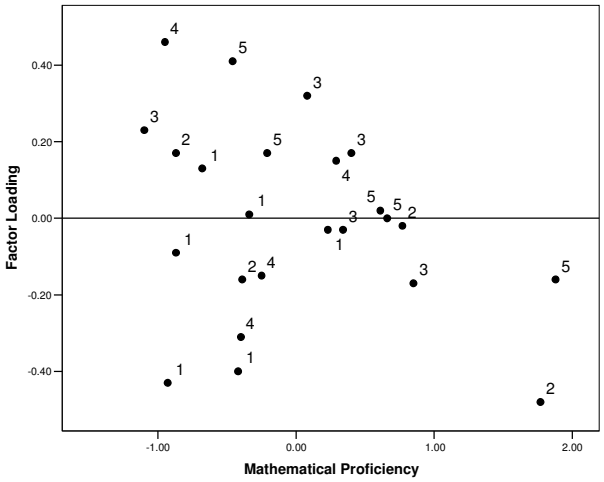
Figure D.2. Principal Components (Standardized Residual) Factor Plots of Grade 6 On-Grade Items



(a) First Factor

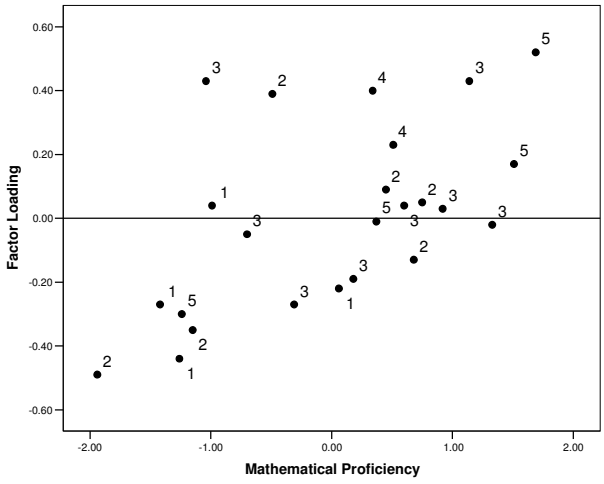


(b) Second Factor

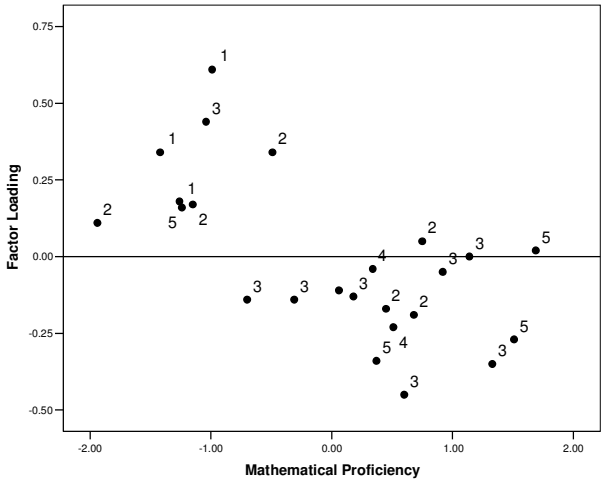


(c) Third Factor

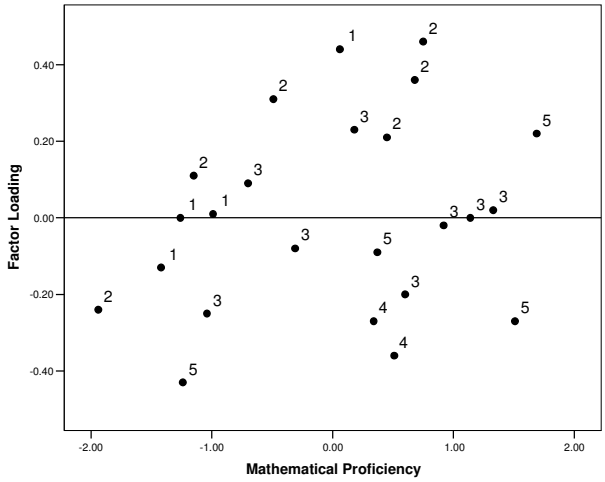
Figure D.3. Principal Components (Standardized Residual) Factor Plots of Grade 7 On-Grade Items



(a) First Factor



(b) Second Factor



(c) Third Factor

Figure D. 4. Principal Components (Standardized Residual) Factor Plots of Grade 8 On-Grade Items

APPENDIX E: CLUSTER RESULTS FOR INCLUDING ON-GRADE AND OFF-GRADE ITEMS USING CONDITIONAL ITEM COVARIANCES (DETECT)

Table E.1 *Cluster Results for Grade 3*

Item Level	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
Grade 3 items only	1, 2, 3, 9, 10, 14, 15, 17, 18, 19, 20, 26	4, 7, 8, 21, 22, 24, 25	6, 16	11, 23, 30	12, 28	
Grade 3: G2 & 3 Items	1, 2, 3, 5, 9, 10, 15, 18, 19, 20, 27	4, 11, 12, 22, 23, 30	6, 14, 16	7, 8, 21, 24, 25	26, 28	
Grade 3: G3 & 4 Items	2, 3, 9, 10, 11, 14, 15, 17, 18, 19, 20, 29, 30	1, 4, 7, 8, 21, 22, 24, 25	6, 12, 13, 16	23	26, 28	
Grade 3 Common Items (across clusters)	2, 3, 9,10,15, 18, 19, 20	4, 7, 22	6,16	na	28	
Theoretical Clusters by Content Strands	Items 1-9: Numbers & Operations	Items 10-13: Geometry	Items 14-19: Algebra & Patterns	Items 20-22: Data Analysis & Probability	Items 23-30: Measurement	

Note: Off-grade item numbers are underlined.

Table E.2 *Cluster Results for Grade 4*

Item Level	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Grade 4 items only	1, 2, 7, 9, 15, 20, 21, 24, 25, 26, 28	3, 4, 6, 10	5, 17, 18, 22, 23, 27	12, 13, 14, 30	na
Grade 4: G3 & 4 Items	1, 2, 3, 7, 9, 15, 19, 20, 21, 25, 26, 28	4, 6, 16, 23, 24	5, 18, 22, 27	10	12, 13, 14, 17, 29, 30
Grade 4: G4 & 5 Items	1, 2, 7, 8, 9, 15, 20, 21, 24, 25, 26, 28	3, 4, 6, 10	5, 18, 22, 27	11, 12, 13, 14, 17, 30	23
Grade 4 Common Items (across clusters)	1, 2, 7, 9 15, 20, 21,25, 28	4, 6	5, 18, 22, 27	na	na
Theoretical Clusters by Content Strands	Items 1-8: Numbers & Operations	Items 9-15: Geometry	Items 16-18: Algebra & Patterns	Items 19-24: Data Analysis & Probability	Items 25-30: Measurement

Note: Off-grade item numbers are underlined.

Table E.3 *Clusters Results for Grade 5*

Item Level	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Grade 5 items only	1, 2, 3, 6, 19, 26, 27, 28	4, 8, 10, 20	5, 18, 21	11, 12, 14, 17, 24, 25, 29, 30	22
Grade 5: G4 & 5 Items	1, 2, 3, 9, 19, 21, 22, 26, 27, 28	4, 6, 7, 8, 10, 11, 13	5, 21	12, 14, 17, 20, 23, 24, 25, 29, 30	na
Grade 5: G5 & 6 Items	1, 2, 3, 7, 19, 22, 26, 27, 28	4, 8, 10, 16, 20	5, 15, 18, 21	11, 12, 14, 17, 24, 25, 29, 30	na
Grade 5 Common Items (across clusters)	1, 2, 3, 19, 26, 27, 28	4, 8, 10	5, 21	11, 12, 14, 17, 24, 25, 29, 20	na
Theoretical Clusters by Content Strands	Items 1-9: Numbers & Operations	Items 10-15: Geometry	Items 16-21: Algebra & Patterns	Items 22-25: Data Analysis & Probability	Items 26-30: Measurement

Note: Off-grade item numbers are underlined.

Table E.4 *Clusters Results for Grade 6*

Item Level	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Grade 6 items only	1, 13	2, 8, 19, 26	3, 6, 12, 22, 23, 27	5, 10, 11, 14, 15, 16, 20, 25, 30	4, 24, 28
Grade 6: G5 & 6 Items	1, 13, 24, 28	2, 4, 18, 19	3, 6, 7, 8, 12, 22, 23, 26, 27, 29	5, 9, 10, 11, 14, 15, 16, 20, 25, 30	na
Grade 6: G6 & 7 Items	1, 3, 12, 13, 24	2, 8, 19, 26	6, 21, 22, 23, 27	5, 10, 11, 14, 15, 16, 20, 25, 30	4, 17, 28
Grade 6 Common Items (across clusters)	1, 13	2		5, 10, 11, 14, 15, 16, 20, 25, 30	na
Theoretical Clusters by Content Strands	Items 1-10: Numbers & Operations	Items 11-14: Geometry	Items 15-19: Algebra & Patterns	Items 20-24: Data Analysis & Probability	Items 25-30: Measurement

Note: Off-grade item numbers are underlined.

Table E.5 *Clusters Results for Grade 7*

Item Level	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Grade 7 items only	1, 4, 5, 7, 9, 10, 19, 20, 21, 22, 24, 29	2, 3, 11, 15, 16, 23, 25, 27	13, 30	15, 18	na
Grade 7: G6 & 7 Items	1, 4, 5, 7, 9, 10, 14, 18, 19, 20, 22, 28, 29	2, 13, 30	3, 8, 11, 16, 23, 24, 25, 26, 27	15, 17, 21	na
Grade 7: G7 & 8 Items	1, 3, 6, 11, 15, 16, 25	2, 12, 13, 23, 27, 30	4, 5, 7, 9, 10, 19, 20, 21, 22, 24, 29	17, 18	na
Grade 7 Common Items (across clusters)	1	2	na	na	na
Theoretical Clusters by Content Strands	Items 1-7: Numbers & Operations	Items 8-13: Geometry	Items 14-19: Algebra & Patterns	Items 20-23: Data Analysis & Probability	Items 24-30: Measurement

Note: Off-grade item numbers are underlined.

Table E.6 *Clusters Results for Grade 8*

Item Level	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
Grade 8 items only	1, 2, 3, 8, 10, 13, 15, 22, 27	4, 7, 12, 16, 18, 29	11, 21, 25, 28	14, 17, 20, 30	20	na
Grade 8: G7 & 8 Items	1, 10, 12, 22	4, 7, 16, 18, 29	11, 21, 25, 28	2, 3, 5, 8, 9, 13, 19, 27	14, 15, 20, 23, 24	17, 30
Grade 8: G8 & 9 Items	1, 2, 3, 8, 10, 13, 15, 22, 26, 27	4, 7, 11, 12, 16, 18, 21, 25, 29	6, 14, 17, 20, 30	24, 28	na	na
Grade 8 Common Items (across clusters)	1, 10, 22	4, 7, 16, 18, 29	na	na	na	na
Theoretical Clusters by Content Strands	Items 1-6: Numbers & Operations	Items 7-13: Geometry	Items 14-22: Algebra & Patterns	Items 23-26: Data Analysis & Probability	Items 27-30: Measurement	

Note: Off-grade item numbers are underlined.

APPENDIX F: NOHARM FACTOR LOADINGS FOR OFF GRADE ITEMS

Table F.1 *Grade 4: On- and Off-Grade Items*

Grade 3 and 4		
Item Number	Factor 1	Factor 2
1	**0.572**	0.008
2	**0.465**	0.057
3	**0.189**	0.055
4	0.175	**-0.267**
5	**-0.289**	0.167
6	0.079	**-0.298**
7	0.233	**0.774**
9	**0.444**	0.302
10	-0.072	**-0.139**
12	0.060	**-0.269**
13	**0.320**	-0.166
14	**0.257**	-0.148
15	0.030	**0.214**
16	***0.415***	*-0.132*
17	**0.411**	0.001
18	**0.231**	0.166
19	***0.849***	*0.079*
20	**0.658**	0.330
21	**0.827**	0.502
22	**0.205**	0.075
23	**-0.075**	-0.036
24	0.052	**0.147**
25	**0.515**	-0.049
26	**0.549**	0.168
27	-0.036	**0.181**
28	**0.581**	-0.028
29	***0.369***	*-0.133*
30	-0.101	**-0.400**

Grade 4 and 5		
Item Number	Factor 1	Factor 2
1	**0.571**	-0.019
2	**0.530**	-0.024
3	**0.178**	0.016
4	0.219	**-0.299**
5	**-0.320**	0.209
6	0.084	**-0.354**
7	0.310	**0.721**
8	***0.518***	*0.085*
9	**0.413**	0.267
10	-0.038	**-0.160**
11	***0.335***	*-0.006*
12	0.087	**-0.275**
13	**0.267**	-0.155
14	**0.231**	-0.094
15	0.012	**0.234**
17	**0.426**	0.020
18	**0.250**	0.164
20	**0.718**	0.279
21	**0.841**	0.436
22	**0.168**	0.140
23	**-0.102**	-0.024
24	0.059	**0.183**
25	**0.466**	-0.046
26	**0.548**	0.173
27	-0.010	**0.229**
28	**0.541**	-0.023
30	-0.147	**-0.322**

Table F.2 *Grade 5: On- and Off- Grade Items*

G4 and 5		
Item Number	Factor 1	Factor 2
1	**0.230**	0.140
2	**0.423**	0.198
3	0.117	**0.333**
4	**0.192**	0.140
5	**0.434**	0.168
6	0.037	**0.591**
7	0.140	**0.537**
8	0.119	**0.391**
9	0.217	**0.391**
10	0.049	**0.541**
11	0.322	**0.481**
12	**0.458**	0.292
13	0.193	**0.430**
14	**0.520**	0.218
17	**0.382**	0.356
18	0.101	**0.187**
19	**0.423**	0.271
20	**0.382**	0.190
21	0.173	**0.281**
22	**0.293**	0.215
23	**0.479**	0.199
24	**0.469**	0.206
25	**0.401**	-0.142
26	**0.355**	0.295
27	0.293	**0.469**
28	0.324	**0.409**
29	**0.429**	0.223
30	**0.639**	0.036

G5 and 6		
Item Number	Factor 1	Factor 2
1	**0.221**	0.135
2	**0.426**	0.270
3	**0.352**	0.086
4	0.156	**0.183**
5	0.292	**0.334**
7	**0.340**	0.248
8	**0.377**	0.103
10	**0.438**	0.125
11	0.331	**0.436**
12	0.241	**0.488**
14	0.162	**0.563**
15	**0.225**	0.019
16	**0.567**	0.138
17	0.278	**0.457**
18	**0.256**	0.049
19	**0.442**	0.315
20	0.269	**0.346**
21	**0.453**	0.071
22	0.259	**0.276**
24	0.088	**0.561**
25	-0.040	**0.320**
26	0.294	**0.352**
27	**0.474**	0.280
28	**0.409**	0.317
29	0.212	**0.433**
30	0.073	**0.616**

Table F.3 *Grade 6: On- and Off- Grade Items*

G5 and 6		
Item Number	Factor 1	Factor 2
1	**0.291**	-0.045
2	0.114	**0.384**
3	**0.253**	0.023
4	0.286	**0.364**
5	-0.032	**0.754**
6	0.270	**0.361**
7	0.360	**0.404**
8	**0.378**	0.279
9	0.230	**0.751**
10	0.014	**0.597**
11	0.263	**0.525**
12	**0.524**	0.045
13	**0.157**	0.140
14	-0.084	**0.227**
15	0.180	**0.584**
16	0.084	**0.449**
18	0.010	**0.263**
19	0.179	**0.463**
20	0.117	**0.389**
22	**0.421**	0.200
23	**0.427**	0.349
24	0.301	**0.444**
25	0.131	**0.218**
26	**0.395**	-0.009
27	**0.594**	0.056
28	0.003	**0.125**
29	**0.645**	0.319
30	0.143	**0.441**

G6 and 7		
Item Number	Factor 1	Factor 2
1	**0.294**	-0.019
2	0.109	**0.370**
3	**0.329**	0.002
4	0.246	**0.375**
5	-0.024	**0.735**
6	**0.346**	0.337
8	**0.344**	0.268
10	-0.054	**0.613**
11	0.230	**0.488**
12	**0.442**	0.082
13	**0.162**	0.135
14	-0.014	**0.213**
15	0.165	**0.605**
16	0.092	**0.442**
17	**0.159**	-0.002
19	0.143	**0.451**
20	0.033	**0.415**
21	**0.380**	0.369
22	**0.434**	0.216
23	**0.415**	0.380
24	0.317	**0.453**
25	0.091	**0.237**
26	**0.445**	0.005
27	**0.601**	0.061
28	-0.009	**0.176**
30	0.079	**0.470**

Table F.4 *Grade 7: On- and Off-Grade Items*

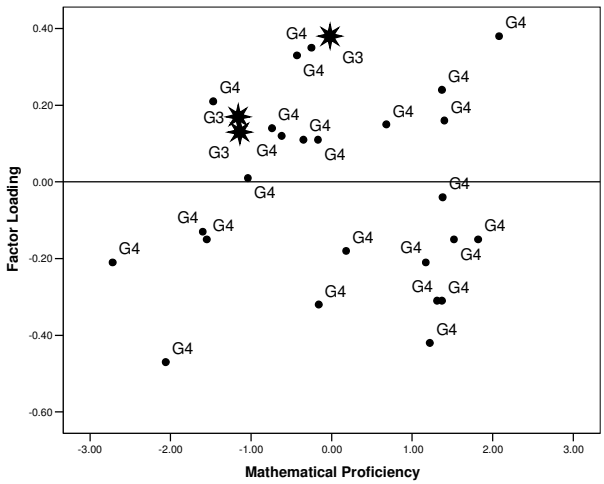
G6 and 7		
Item Number	Factor 1	Factor 2
1	**0.302**	0.266
2	-0.036	**0.164**
3	-0.117	**0.326**
4	**0.478**	0.039
5	**0.529**	0.040
7	**0.582**	0.108
8	0.083	**0.267**
9	**0.614**	-0.019
10	**0.744**	0.127
11	0.169	**0.170**
13	0.141	**0.249**
14	**0.678**	-0.175
15	-0.020	**0.051**
16	**-0.391**	0.335
17	**0.144**	0.102
18	**0.369**	0.160
19	**0.468**	0.271
20	**0.506**	-0.134
21	**0.325**	0.206
22	**0.470**	0.081
23	0.240	**0.612**
24	**0.416**	0.283
25	**-0.288**	0.109
26	0.151	**0.561**
27	0.090	**0.403**
28	**0.428**	0.072
29	**0.516**	0.297
30	0.003	**0.018**

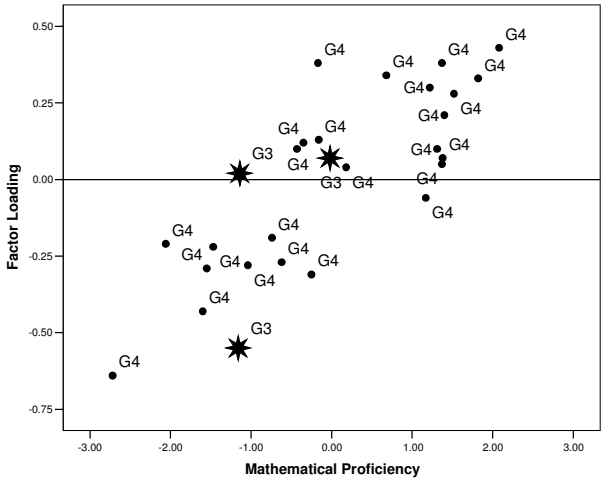
G7 and 8		
Item Number	Factor 1	Factor 2
1	**0.447**	0.076
2	**0.155**	-0.103
3	**0.452**	-0.410
4	**0.417**	0.293
5	0.378	**0.398**
6	**0.686**	0.089
7	**0.458**	0.418
9	**0.411**	0.405
10	0.297	**0.773**
11	**0.351**	-0.037
12	-0.027	**0.082**
13	**0.194**	0.108
15	**0.094**	-0.088
16	0.101	**-0.417**
17	0.034	**0.154**
18	**0.402**	0.152
19	**0.459**	0.303
20	0.151	**0.464**
21	**0.329**	0.211
22	0.242	**0.437**
23	**0.361**	0.149
24	**0.228**	0.406
25	-0.058	**-0.264**
27	0.255	0.035
29	0.336	**0.449**
30	-0.008	**-0.022**

Table F.5 *Grade 8: On- and Off- Grade Items*

G7 and 8			G8 and 9		
Item Number	Factor 1	Factor 2	Item Number	Factor 1	Factor 2
1	**0.489**	0.133	1	**0.386**	0.376
2	**0.479**	0.209	2	0.106	**0.565**
3	**0.553**	0.121	3	0.098	**0.603**
4	0.313	**0.671**	4	**0.431**	0.363
5	**0.485**	0.404	6	**0.513**	-0.254
7	0.027	**0.496**	7	**0.310**	0.072
8	**0.530**	0.107	8	0.126	**0.559**
9	**0.375**	0.251	10	**0.358**	0.316
10	**0.491**	0.115	11	**0.311**	0.271
11	**0.330**	0.221	12	**0.581**	0.343
12	**0.465**	0.391	13	0.112	**0.704**
13	**0.729**	0.100	14	**0.223**	0.207
14	**0.292**	0.107	15	0.308	**0.400**
15	**0.518**	0.121	16	0.031	**-0.159**
16	**-0.204**	0.134	17	**0.167**	0.024
17	0.182	**-0.194**	18	**0.453**	0.043
18	0.036	**0.659**	20	**0.568**	0.172
19	**0.741**	0.135	21	**0.303**	0.183
20	**0.505**	0.087	22	0.264	**0.370**
21	**0.277**	0.118	24	**0.138**	-0.089
22	**0.449**	0.023	25	**0.375**	0.015
23	**0.403**	0.093	26	0.303	**0.380**
24	**0.050**	-0.001	27	0.116	**0.593**
25	0.111	**0.300**	28	**0.374**	0.005
27	**0.513**	0.208	29	0.198	**-0.470**
28	**0.183**	0.110	30	**0.264**	0.203
29	**-0.423**	0.188			
30	**0.303**	0.005			

APPENDIX G: PRINCIPAL COMPONENTS (STANDARDIZED RESIDUAL) FACTOR PLOTS OF ON- AND OFF-GRADE ITEMS

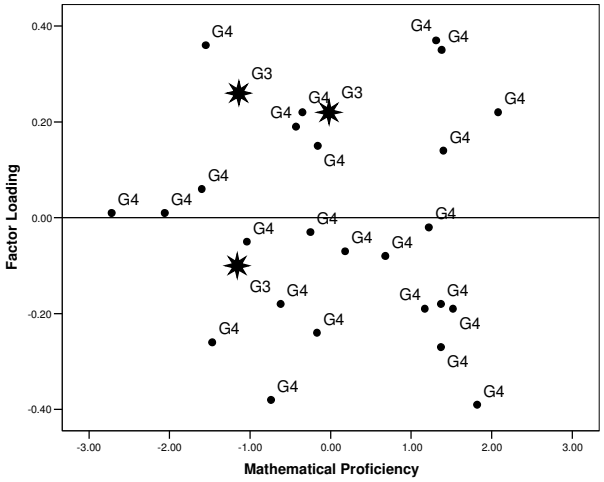


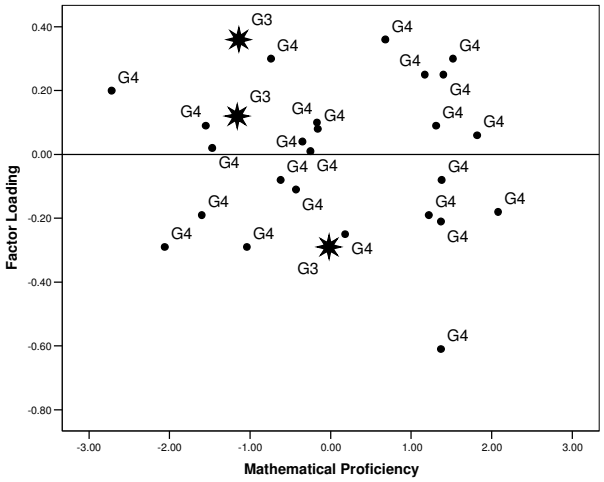


Note: Off-grade items are designated with a ” symbol.

(a) First Factor

(b) Second Factor

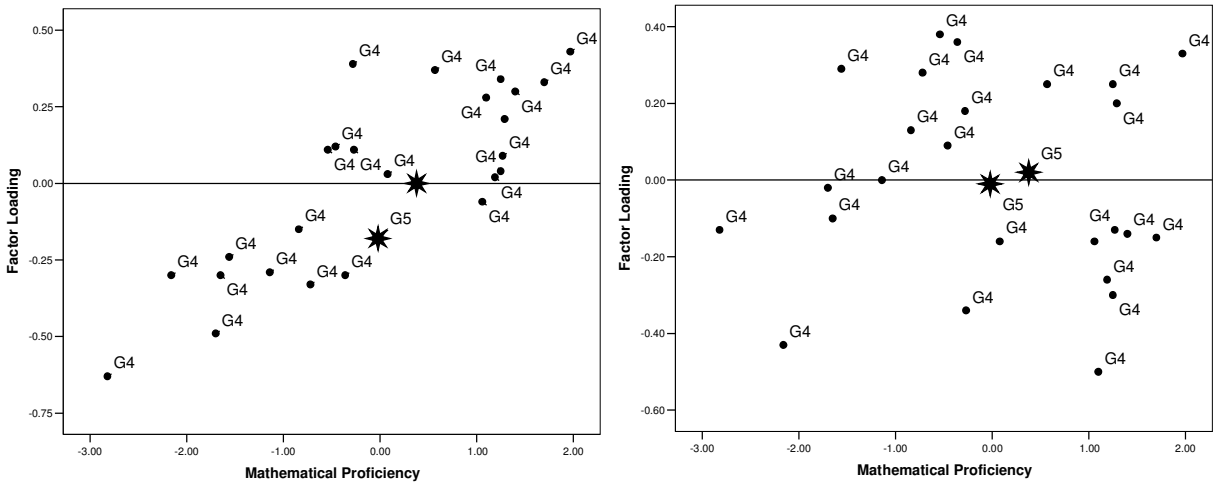




(c) Third Factor

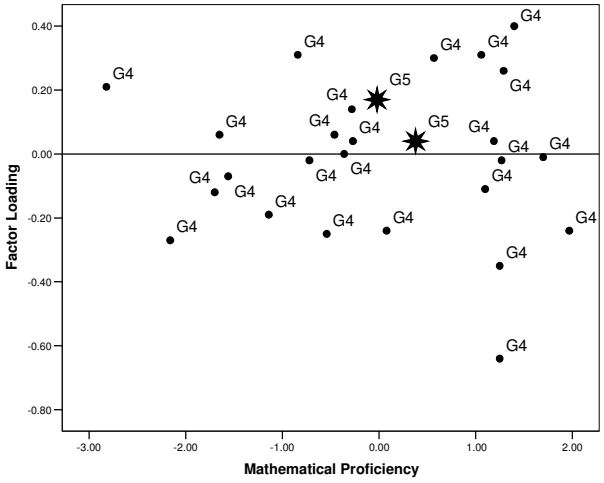
(d) Fourth Factor

Figure G.1. Grade 4: Grade 3 and 4 Items



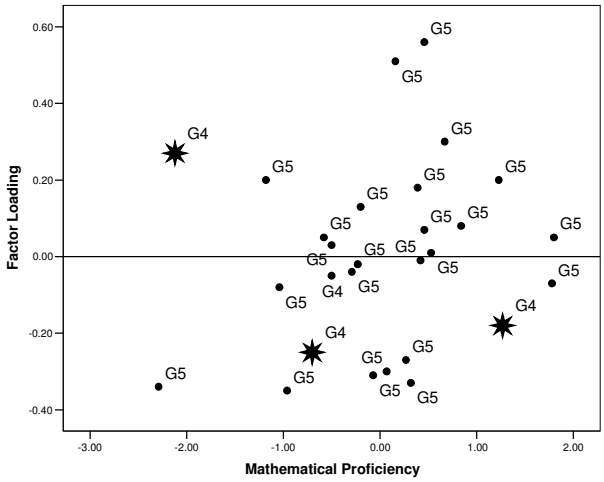
Note: Off-grade items are designated with a ” symbol.

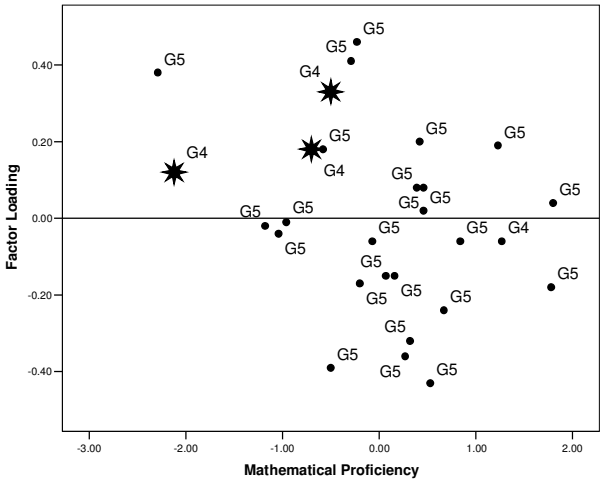
(a) First Factor *(b) Second Factor*



(c) Third Factor

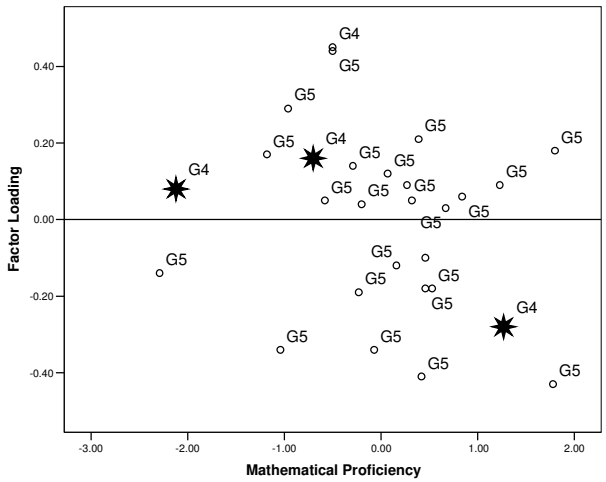
Figure G.2. Grade 4: Grade 4 and 5 Items





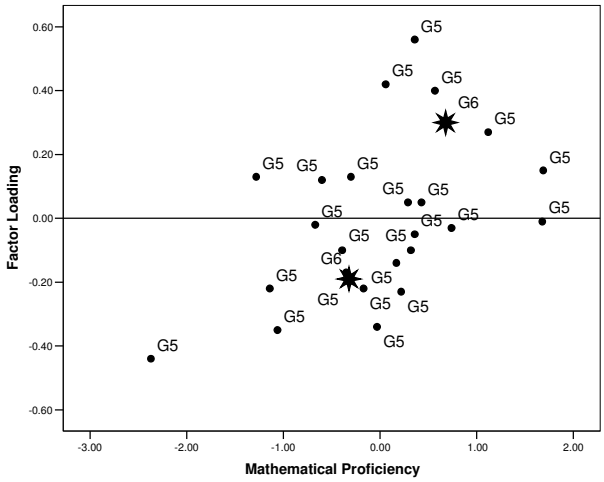
Note: Off-grade items are designated with a ” symbol.

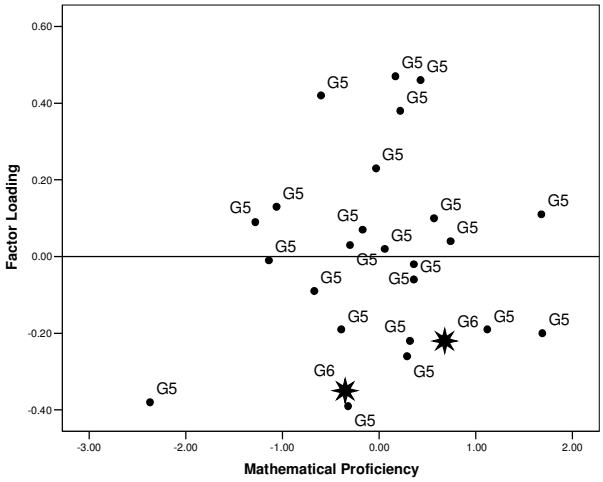
(a) First Factor *(b) Second Factor*



(c) Third Factor

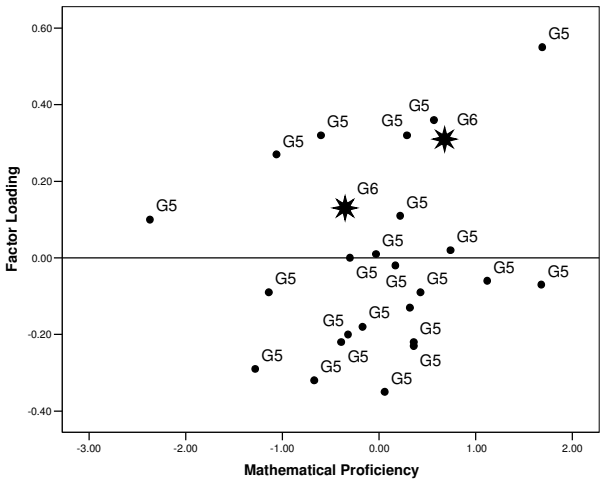
Figure G.3. Grade 5: Grade 4 and 5 Items





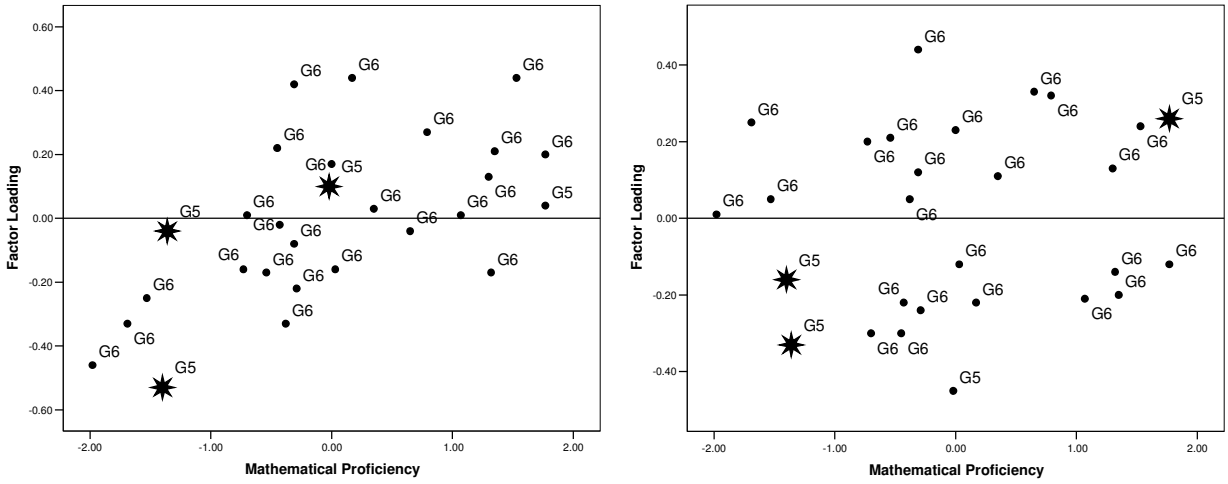
Note: Off-grade items are designated with a ” symbol.

(a) First Factor *(b) Second Factor*



(c) Third Factor

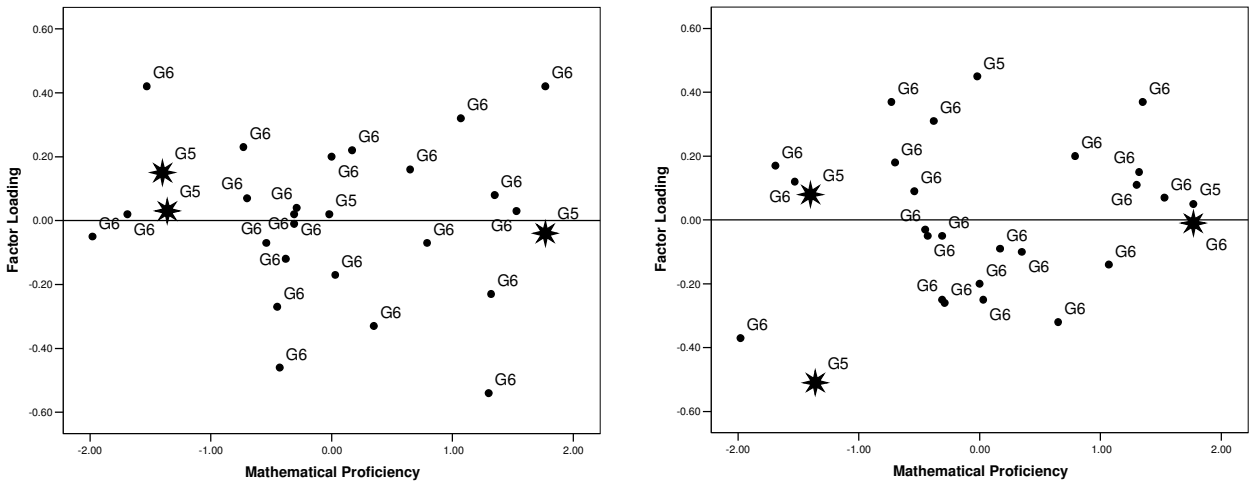
Figure G.4. Grade 5: Grade 5 and 6 Items



Note: Off-grade items are designated with a ” symbol.

(a) First Factor

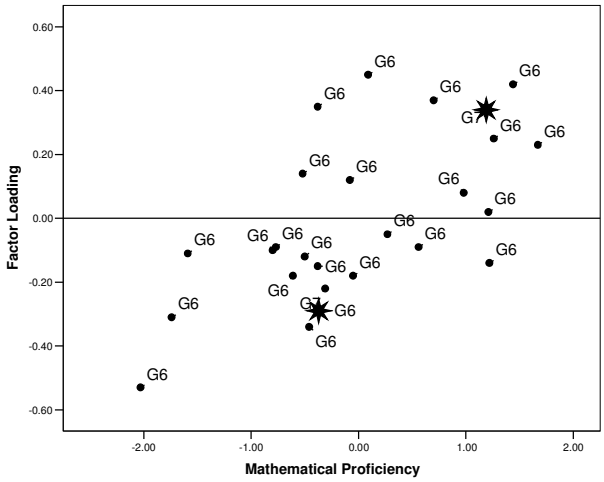
(b) Second Factor

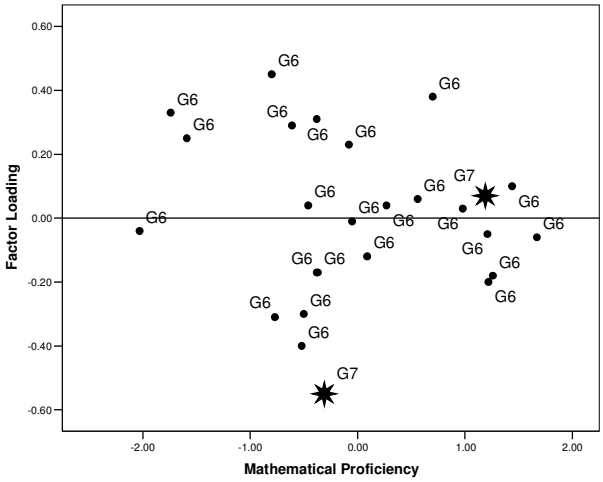


(c) Third Factor

(d) Fourth Factor

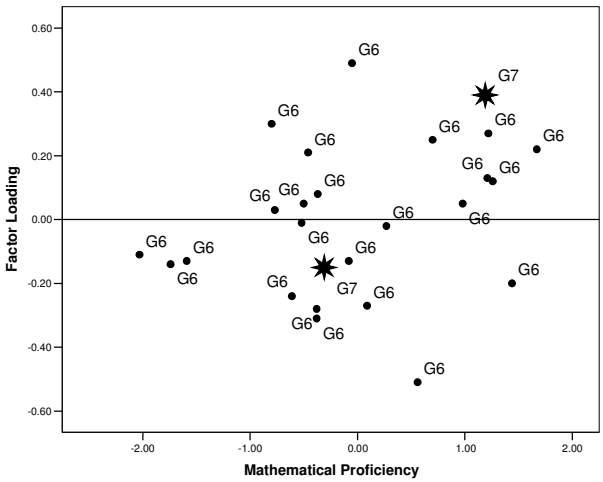
Figure G.5. Grade 6: Grade 5 and 6 Items





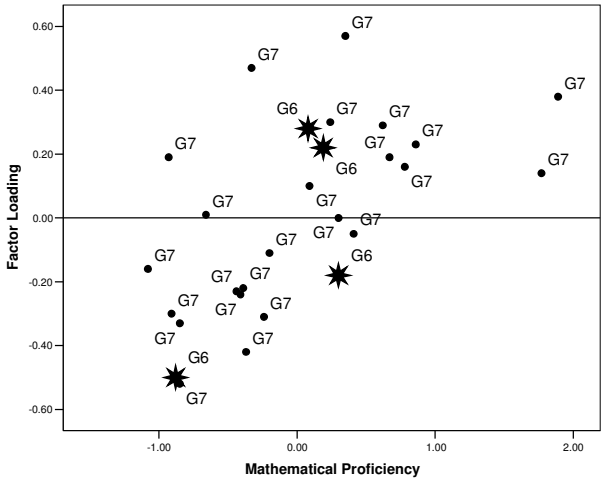
Note: Off-grade items are designated with a ” symbol.

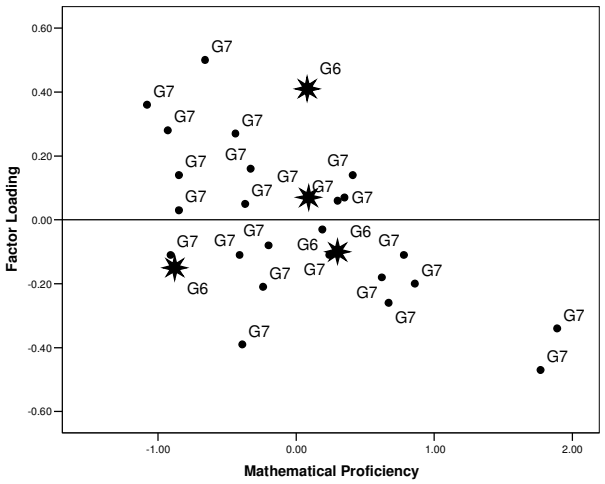
(a) First Factor *(b) Second Factor*



(c) Third Factor

Figure G.6. Grade 6: Grade 6 and 7 Items

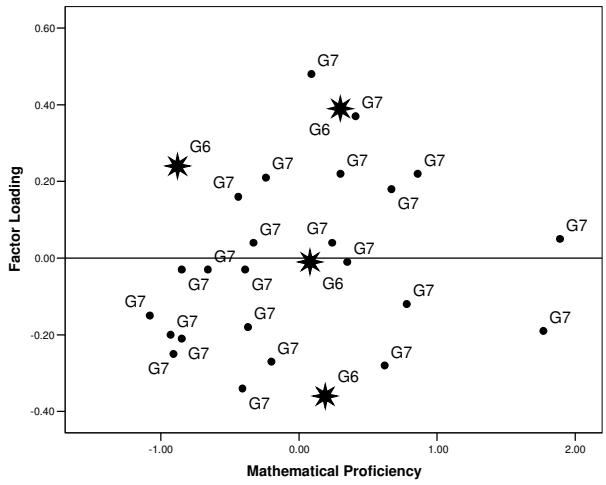


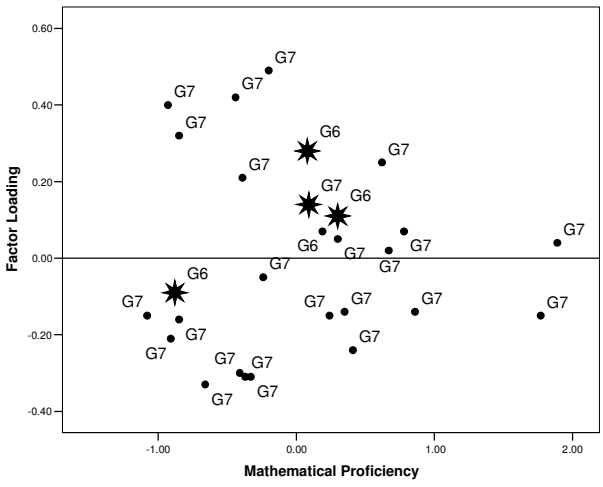


Note: Off-grade items are designated with a ” symbol.

(a) First Factor

(b) Second Factor

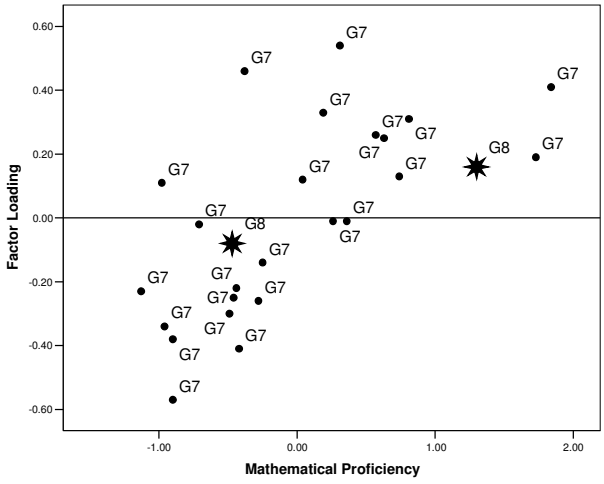


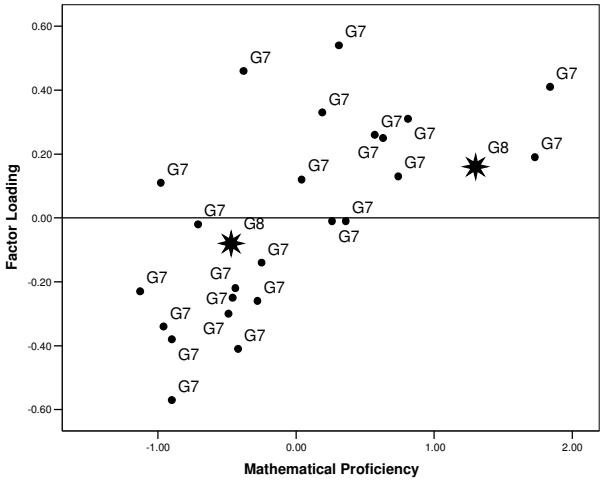


(c) Third Factor

(d) Fourth Factor

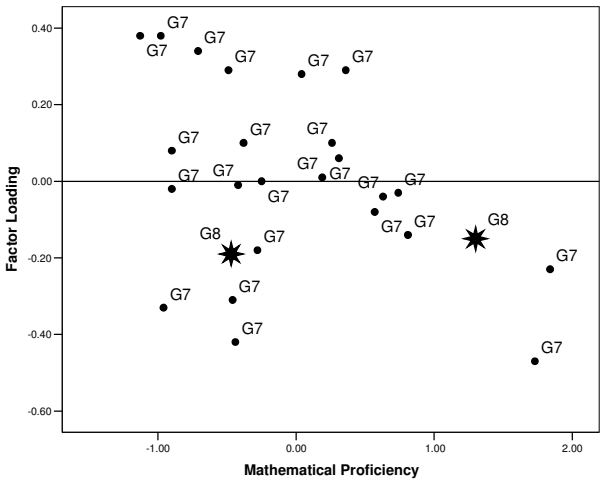
Figure G.7. Grade 7: Grade 6 and 7 Items





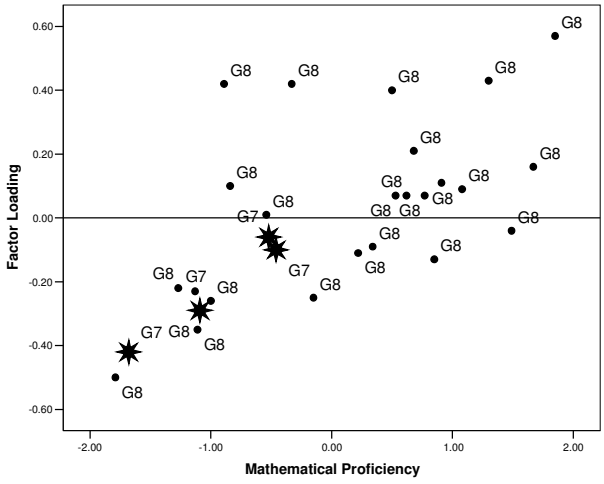
Note: Off-grade items are designated with a ” symbol.

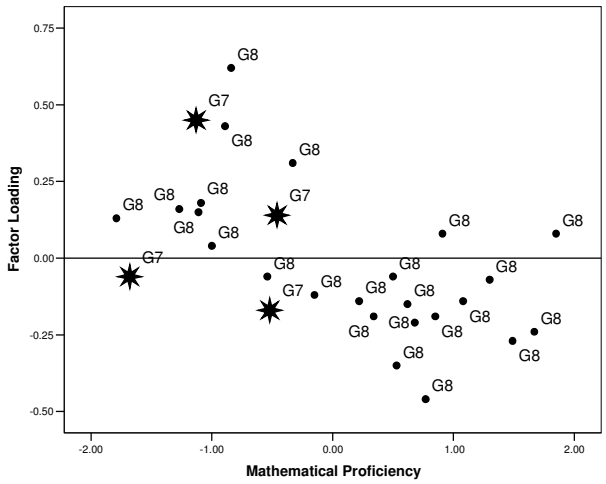
(a) First Factor *(b) Second Factor*



(c) Third Factor

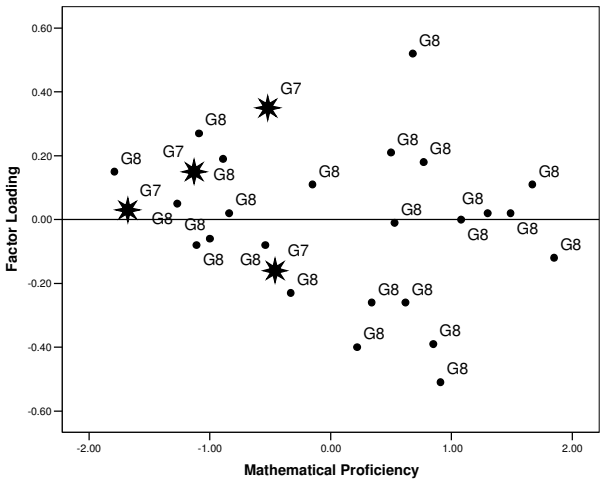
Figure G.8. Grade 7: Grade 7 and 8 Items





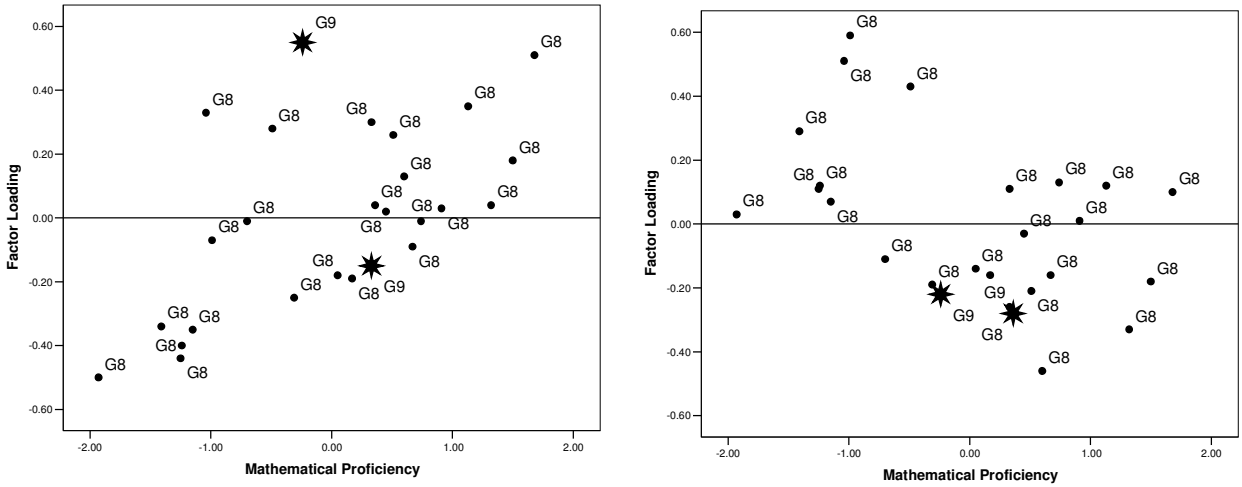
Note: Off-grade items are designated with a ” symbol.

(a) First Factor *(b) Second Factor*



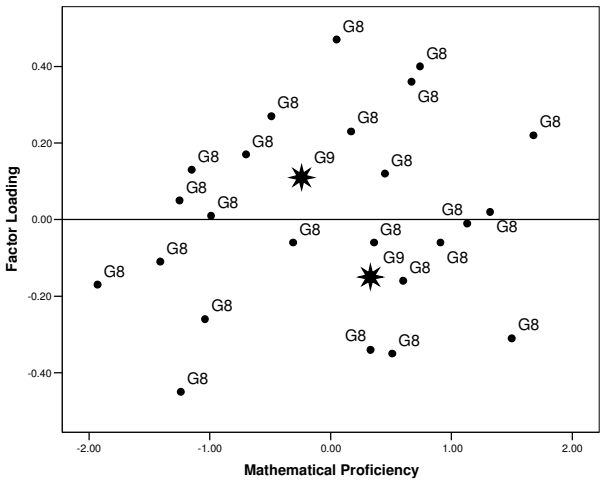
(c) Third Factor

Figure G.9. Grade 8: Grade 7 and 8 Items



Note: Off-grade items are designated with a ” symbol.

(a) First Factor *(b) Second Factor*



(c) Third Factor

Figure G.10. Grade 8: Grade 8 and 9 Items

APPENDIX H: COMPARISON OF EXPLORATORY RESULTS OF ON-GRADE ITEMS BY SOFTWARE PROGRAM

Table H.1 *Comparison of Exploratory Results from Grade 3 On-Grade Items by Software Program*

Software Program	Results				
DETECT	**Cluster 1**	**Cluster 2**	**Cluster 3**	**Cluster 4**	**Cluster 5**
	1, 2, 3, 9, 10, 14, 15, 17, 18, 19, 20, 26	4, 7, 8, 21, 22, 24, 25	6, 16	11, 23, 30	12, 28
NOHARM (3 Factors)	**Factor 1**	**Factor 2**	**Factor 3**		
	1, 2, 3, 4, 6, 9, 10, 14, 15, 17, 18, 19, 20	4, 8, 11, 12, 16, 22, 23, 24, 30	7, 21, 25, 26, 28		
NOHARM (2 Factors)	**Factor 1**	**Factor 2**			
	1, 2, 3, 6, 9, 10, 11, 12, 14, 15, 16, 17, 18, 19, 20, 26, 30	4, 7, 8, 21, 22, 23, 24, 25, 28			
DIMTEST	**AT Subtest**				
	7, 8, 21, 24, 25, 26, 28				

Table H.2 *Comparison of Exploratory Results from Grade 4 On-Grade Items by Software Program*

Software Program	Results				
DETECT	**Cluster 1**	**Cluster 2**	**Cluster 3**	**Cluster 4**	**Cluster 5**
	1, 2, 7, 9, 15, 20, 21, 24, 25, 26, 28	3, 4, 6, 10	5, 17, 18, 22, 23, 27	12, 13, 14, 30	
NOHARM (3 Factors)	**Factor 1**	**Factor 2**	**Factor 3**		
	1, 2, 3, 5, 9, 13, 14, 17, 20, 21, 25, 26, 28	4, 6, 7, 12, 15, 22, 24, 30	10, 18, 23, 27		
NOHARM (2 Factors)	**Factor 1**	**Factor 2**			
	1, 2, 3, 5, 9, 13, 14, 17, 18, 20, 21, 23, 25, 26, 28	4, 6, 7, 10, 12, 15, 22, 24, 27, 30			
DIMTEST	**AT Subtest**				
	3, 4, 6, 10, 12, 14, 30				

Table H.3 *Comparison of Exploratory Results from Grade 5 On-Grade Items by Software Program*

Software Program	Results				
	Cluster 1	**Cluster 2**	**Cluster 3**	**Cluster 4**	**Cluster 5**
DETECT	1, 2, 3, 6, 19, 26, 27, 28	4, 8, 10, 20	5, 18, 21	11, 12, 14, 17, 24, 25, 29, 30	22
	Factor 1				
NOHARM	1, 2, 3, 4, 5, 7, 8, 10, 11, 12, 14, 17, 18, 19, 20, 21, 22, 24, 25, 26, 27, 28, 29, 30				
	AT Subtest				
DIMTEST	1, 2, 3, 7, 19, 26, 27, 28				

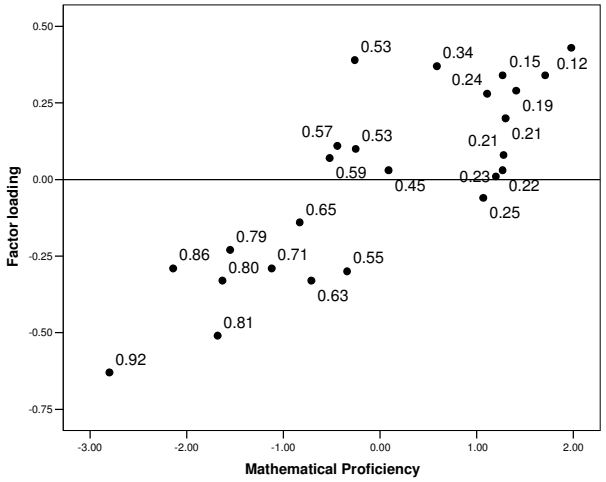
Table H.4 *Comparison of Exploratory Results from Grade 6 On-Grade Items by Software Program*

Software Program	Results				
DETECT	**Cluster 1**	**Cluster 2**	**Cluster 3**	**Cluster 4**	**Cluster 5**
	1, 13	2, 8, 19, 26	3, 6, 12, 22, 23, 27	5, 10, 11, 14, 15, 16, 20, 25, 30	4, 24, 28
NOHARM (3 Factors)	**Factor 1**	**Factor 2**	**Factor 3**		
	5, 10, 11, 14, 15, 16, 20, 24, 25, 28, 30	1, 3, 6, 8, 12, 13, 22, 23, 26, 27	2, 4, 8 ,9		
NOHARM (2 Factors)	**Factor 1**	**Factor 2**			
	2, 4, 5, 6, 10, 11, 14, 15, 16, 19, 20, 23, 24, 25, 28, 30	1, 3, 8, 12, 13, 22, 26, 27			
DIMTEST	**AT Subtest**				
	2, 4, 6, 8, 19, 22, 23, 24, 28				

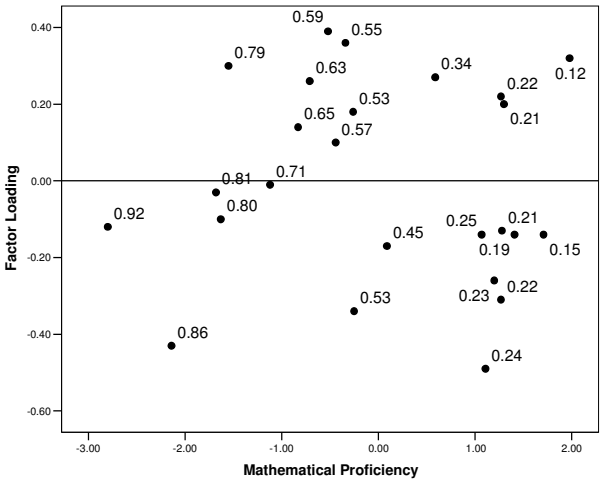
Table H.5 *Comparison of Exploratory Results from Grade 7 On-Grade Items by Software Program*

Software Program	Results			
DETECT	**Cluster 1**	**Cluster 2**	**Cluster 3**	**Cluster 4**
	1, 4, 5, 7, 9, 10, 19, 20, 21, 22, 24, 29	2, 3, 11, 15, 16, 23, 25, 27	13, 30	15, 18
NOHARM	**Factor 1**	**Factor 2**	**Factor 3**	**Factor 4**
	11, 13, 17, 21, 23, 24, 29, 30	1, 4, 5, 9, 18, 19, 25	3, 7, 10, 20, 22	2, 15, 16, 27
DIMTEST	**AT Subtest**			
	3, 11, 15, 16, 25			

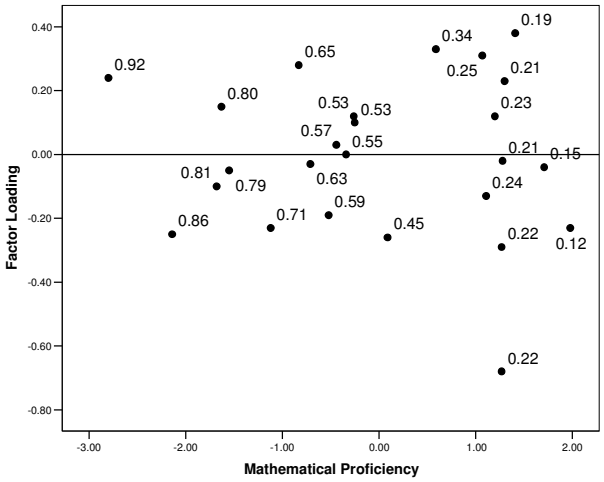
APPENDIX I: WINSTEPS RESIDUAL PLOTS OF ON-GRADE ITEMS BY P-VALUES



(a) First Factor

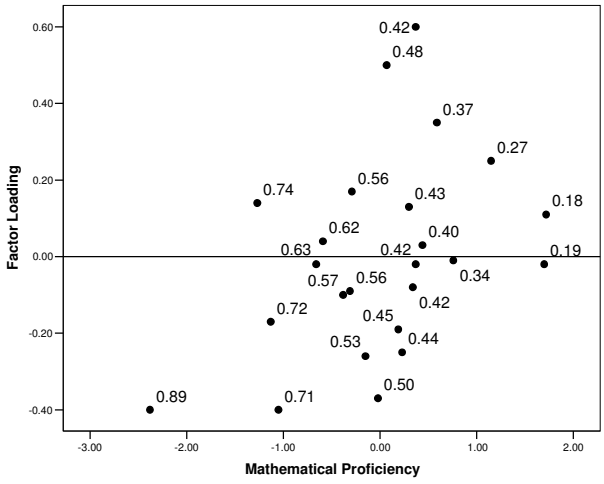


(b) Second Factor

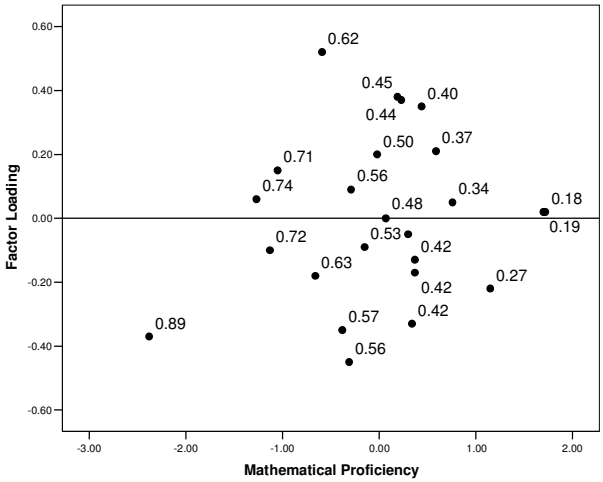


(c) Third Factor

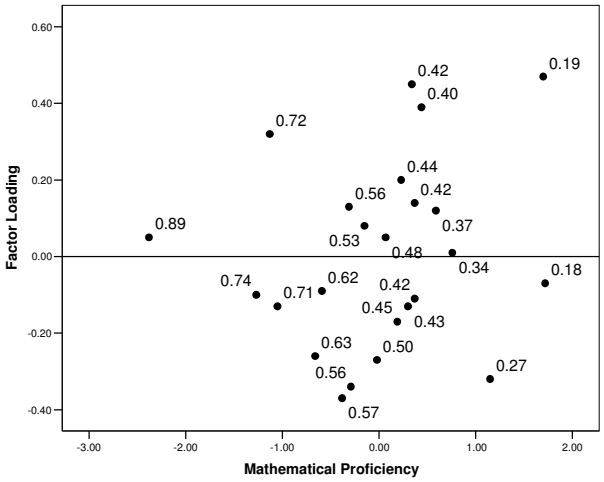
Figure I.1. Grade 4 WINSTEPS Residual Plots of On-Grade Items By P-Values



(a) First Factor

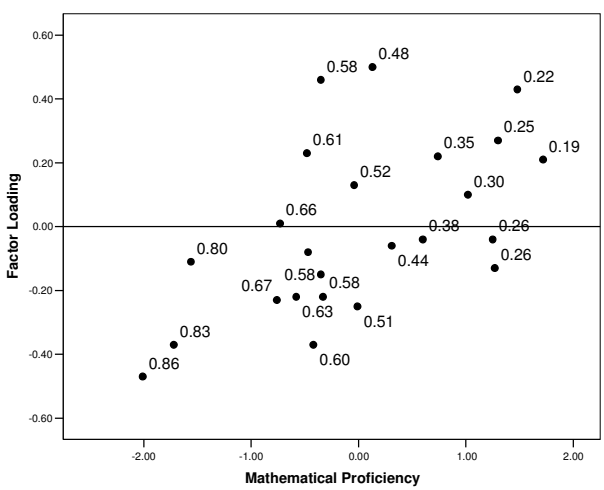


(b) Second Factor

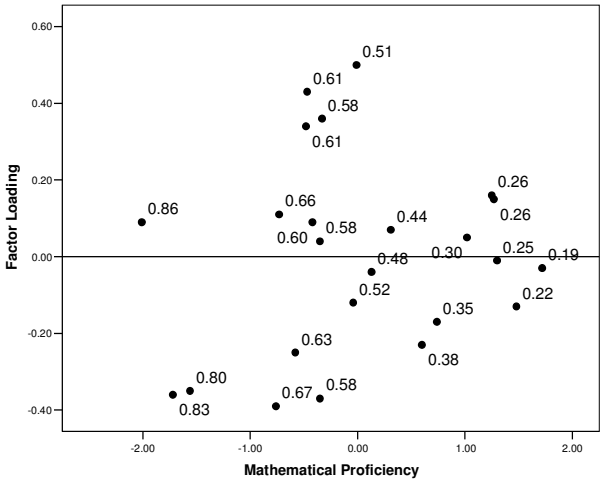


(c) Third Factor

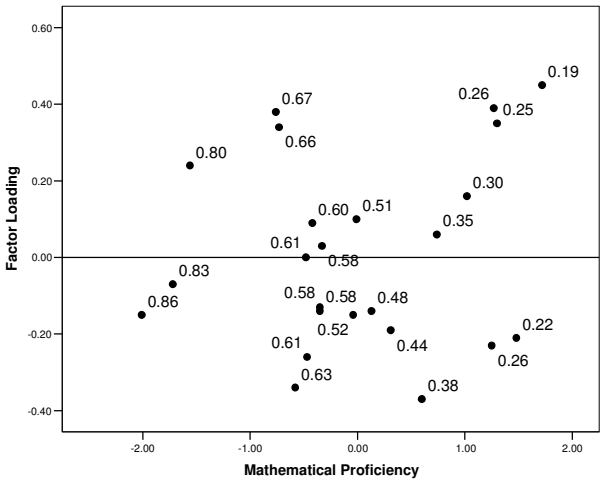
Figure I.2. Grade 5 WINSTEPS Residual Plots of On-Grade Items by P-Values



(a) First Factor

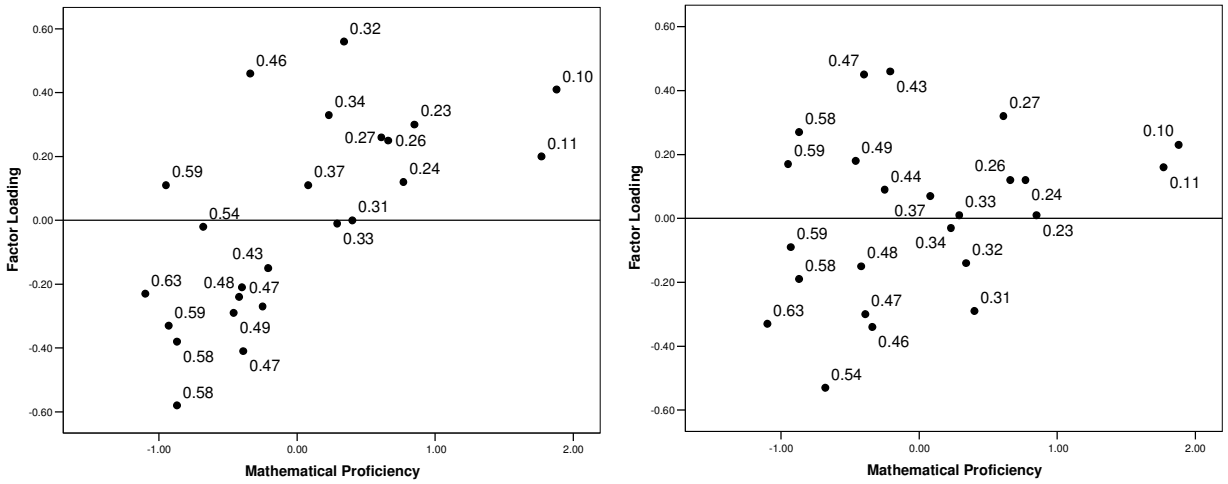


(b) Second Factor

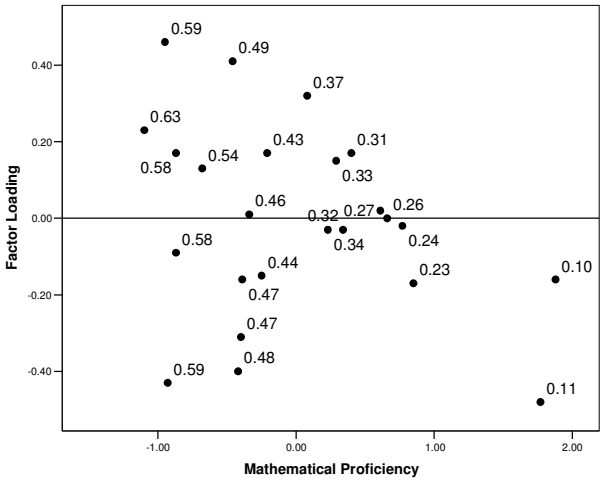


(c) Third Factor

Figure I. 3. Grade 6 WINSTEPS Residual Plots of On-Grade Items by P-Values

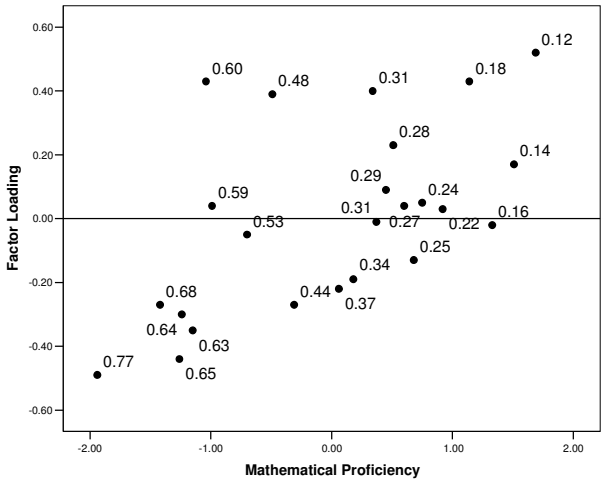


(a) First Factor *(b) Second Factor*

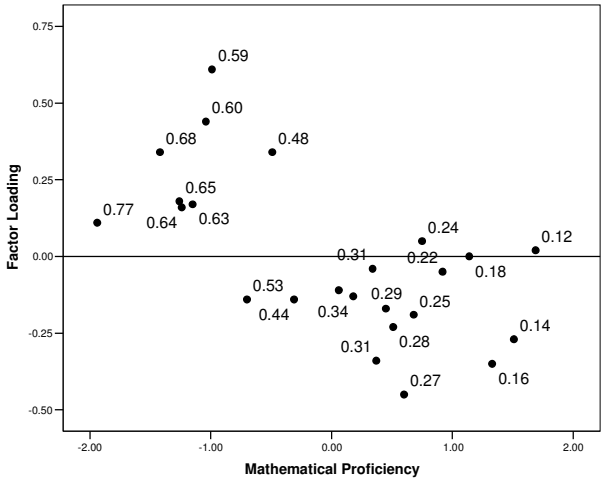


(c) Third Factor

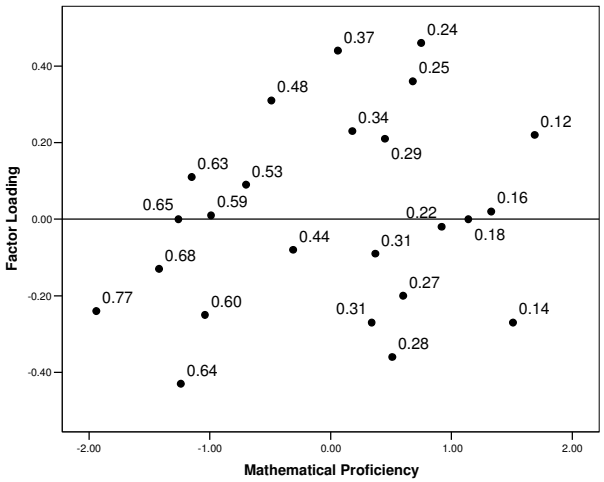
Figure I.4. Grade 7 WINSTEPS Residual Plots of On-Grade Items by P-Values



(a) First Factor



(b) Second Factor



(c) Third Factor

Figure I.5. Grade 8 WINSTEPS Residual Plots of On-Grade Items by P-Values

APPENDIX J: P-VALUES FOR DETECT CLUSTERS

Table J.1 *DETECT Cluster P-Values for Grade 4*

Item	P-value	Item	P-value	Item	P-value	Item	P-value
Cluster 1		*Cluster 2*		*Cluster 3*		*Cluster 4*	
1	0.79	5	0.22	3	0.25	12	0.53
2	0.63	17	0.22	4	0.21	13	0.59
7	0.86	18	0.45	6	0.34	14	0.57
8	0.86	22	0.21	10	0.19	30	0.12
13	0.59	23	0.15	*Mean:*	0.25	*Mean:*	0.45
16	0.53	27	0.24				
17	0.22	*Mean:*	0.25				
20	0.81						
21	0.92						
22	0.21						
24	0.23						
Mean:	0.60						

Table J.2 *DETECT Cluster P-Values for Grade 5*

Item	P-value	Item	P-value	Item	P-value	Item	P-value	Item	P-value
Cluster 1		*Cluster 2*		*Cluster 3*		*Cluster 4*		*Cluster 5*	
28	0.74	11	0.71	10	0.89	21	0.43	22	0.34
27	0.63	12	0.62	20	0.72	5	0.37		0.34
19	0.56	29	0.53	8	0.56	18	0.18		
2	0.48	17	0.5	4	0.42	*Mean:*	0.33		
1	0.42	24	0.45	*Mean:*	0.65				
26	0.42	14	0.44						
6	0.37	30	0.4						
3	0.27	25	0.19						
Mean:	0.49	*Mean:*	0.48						

Table J.3 *DETECT Cluster P-Values for Grade 6*

Item	P-value	Item	P-value	Item	P-value	Item	P-value	Item	P-value
Cluster 1		*Cluster 2*		*Cluster 3*		*Cluster 4*		*Cluster 5*	
5	0.86	3	0.58	2	0.51	4	0.61	1	0.22
10	0.83	6	0.26	8	0.30	24	0.44	13	0.35
11	0.80	12	0.48	19	0.58	28	0.26	*Mean:*	0.29
14	0.38	22	0.61	26	0.19	*Mean:*	0.44		
15	0.60	23	0.66	*Mean:*	0.40				
16	0.58	27	0.25						
20	0.67	*Mean:*	0.47						
25	0.52								
30	0.63								
Mean:	0.65								

Table J.4 DETECT Cluster P-Values for Grade 7

Item	P-value	Item	P-value	Item	P-value	Item	P-value
Cluster 1		*Cluster 2*		*Cluster 3*		*Cluster 4*	
19	0.63	23	0.59	18	0.31	30	0.26
5	0.59	3	0.46	15	0.23	13	0.24
7	0.58	2	0.34	*Mean:*	0.27	*Mean:*	0.25
10	0.58	16	0.32				
1	0.54	27	0.27				
29	0.49	15	0.23				
4	0.48	11	0.11				
9	0.47	25	0.10				
22	0.47	*Mean:*	0.30				
20	0.44						
24	0.43						
21	0.33						
Mean:	0.50						

Table J.5 DETECT Cluster P-Values for Grade 8

Item	P-value	Item	P-value	Item	P-value	Item	P-value	Item	P-value
Cluster 1		*Cluster 2*		*Cluster 3*		*Cluster 4*		*Cluster 5*	
1	0.37	4	0.59	11	0.29	14	0.53	20	0.16
2	0.68	7	0.48	21	0.22	17	0.27	*Mean:*	0.16
3	0.65	12	0.24	25	0.31	20	0.16		
8	0.63	16	0.18	28	0.14	30	0.31		
10	0.25	18	0.6	*Mean:*	0.24	*Mean:*	0.32		
13	0.77	29	0.12						
15	0.44	*Mean:*	0.37						
22	0.34								
27	0.64								
Mean:	0.53								

APPENDIX K: NOHARM FACTOR LOADINGS AND P-VALUES FOR ON-GRADE ITEMS

Table K.1 *Grade 4 NOHARM Two Factor Solution*

First Factor			Second Factor		
Item	Factor Loading	P-value	Item	Factor Loading	P-value
21	0.855	0.92	7	0.724	0.86
20	0.736	0.81	15	0.230	0.53
9	0.451	0.80	27	0.229	0.24
1	0.599	0.79	24	0.169	0.23
26	0.524	0.71	22	0.142	0.21
25	0.451	0.65	10	-0.152	0.19
2	0.523	0.63	12	-0.274	0.53
13	0.324	0.59	4	-0.288	0.21
14	0.209	0.57	30	-0.321	0.12
28	0.560	0.55	6	-0.367	0.34
18	0.215	0.45		*Mean p-value*	0.35
3	0.178	0.25			
17	0.431	0.22			
5	-0.270	0.22			
23	-0.096	0.15			
	Mean p-value	0.55			

Table K.2 *Grade 4 NOHARM Three Factor Solutions*

First Factor			Second Factor			Third Factor		
Item	Factor Loading	P-value	Item	Factor Loading	P-value	Item	Factor Loading	P-value
21	0.817	0.92	7	0.671	0.86	18	0.237	0.45
20	0.692	0.81	15	0.263	0.53	27	0.300	0.24
9	0.448	0.80	12	-0.310	0.53	10	-0.244	0.19
1	0.595	0.79	6	-0.282	0.34	23	0.128	0.15
26	0.484	0.71	24	0.188	0.23	*Mean p-value:*		0.26
25	0.462	0.65	22	0.136	0.21			
2	0.513	0.63	4	-0.246	0.21			
13	0.315	0.59	30	-0.300	0.12			
14	0.209	0.57	*Mean p-value:*		0.38			
28	0.577	0.55						
3	0.196	0.25						
5	-0.323	0.22						
17	0.355	0.22						
Mean p-value:		0.59						

Table K.3 *Grade 5 NOHARM One Factor Solution*

	Factor	
Item	Factor Loading	P-value
1	0.258	0.42
2	0.472	0.48
3	0.297	0.27
4	0.239	0.42
5	0.443	0.37
7	0.424	0.57
8	0.313	0.56
10	0.364	0.89
11	0.548	0.71
12	0.534	0.62
14	0.524	0.44
17	0.524	0.50
18	0.184	0.18
19	0.517	0.56
20	0.435	0.72
21	0.320	0.43
22	0.384	0.34
24	0.486	0.45
25	0.211	0.19
26	0.469	0.42
27	0.520	0.63
28	0.506	0.74
29	0.468	0.53
30	0.516	0.40

Table K.4 *Grade 6 NOHARM Two Factor Solution*

First Factor			Second Factor		
Item	Factor Loading	P-value	Item	Factor Loading	P-value
5	0.716	0.86	22	0.376	0.61
10	0.600	0.83	3	0.373	0.58
11	0.496	0.80	12	0.455	0.48
20	0.419	0.67	13	0.194	0.35
23	0.401	0.66	8	0.339	0.30
30	0.474	0.63	27	0.565	0.25
4	0.395	0.61	1	0.329	0.22
15	0.620	0.60	26	0.433	0.19
16	0.440	0.58		*Mean p-value:*	0.37
19	0.456	0.58			
25	0.232	0.52			
2	0.406	0.51			
24	0.478	0.44			
14	0.184	0.38			
6	0.370	0.26			
28	0.185	0.26			
	Mean p-value:	0.57			

Table K.5 *Grade 6 NOHARM Three Factor Solution*

First Factor			Second Factor			Third Factor		
Item	Factor Loading	P-value	Item	Factor Loading	P-value	Item	Factor Loading	P-value
5	0.613	0.86	23	0.383	0.66	4	0.309	0.61
10	0.688	0.83	22	0.381	0.61	19	0.437	0.58
11	0.502	0.80	3	0.379	0.58	2	0.833	0.51
20	0.379	0.67	12	0.483	0.48	8	0.333	0.30
30	0.487	0.63	13	0.203	0.35	*Mean p-value:*		0.50
15	0.484	0.60	8	0.333	0.30			
16	0.491	0.58	6	0.297	0.26			
25	0.206	0.52	27	0.569	0.25			
24	0.383	0.44	1	0.333	0.22			
14	0.289	0.38	26	0.405	0.19			
28	0.182	0.26	*Mean p-value:*		0.39			
Mean p-value:		0.60						

Table K.6 *Grade 7 NOHARM Four Factor Solution*

First Factor		
Item	Factor Loading	P-value
11	0.241	0.11
13	0.255	0.24
17	0.275	0.37
21	0.334	0.33
23	0.540	0.59
24	0.498	0.43
29	0.387	0.49
30	0.149	0.26
Mean p-value:		0.35

Second Factor		
Item	Factor Loading	P-value
1	0.431	0.54
4	0.518	0.48
5	0.447	0.59
9	0.706	0.47
18	0.374	0.31
19	0.525	0.63
25	-0.324	0.10
Mean p-value:		0.45

Third Factor		
Item	Factor Loading	P-value
3	-0.319	0.46
7	0.617	0.58
10	0.608	0.58
20	0.508	0.44
22	0.477	0.47
Mean p-value:		0.51

Fourth Factor		
Item	Factor Loading	P-value
2	0.193	0.34
15	0.140	0.23
16	0.516	0.32
27	0.298	0.27
Mean p-value:		0.29

Table K.7 *Grade 8 NOHARM Five Factor Solution*

First Factor		
Item	Factor Loading	P-value
1	0.556	0.37
22	0.398	0.34
30	0.358	0.31
11	0.418	0.29
10	0.443	0.25
12	0.736	0.24
21	0.316	0.22
20	0.528	0.16
Mean p-value:		0.27

Second Factor		
Item	Factor Loading	P-value
13	0.619	0.77
2	0.615	0.68
3	0.542	0.65
27	0.562	0.64
8	0.581	0.63
29	-0.454	0.12
Mean p-value:		0.58

Third Factor		
Item	Factor Loading	P-value
14	0.318	0.53
15	0.516	0.44
16	-0.370	0.18
Mean p-value:		0.38

Fourth Factor		
Item	Factor Loading	P-value
18	0.579	0.60
4	0.630	0.59
7	0.584	0.48
17	-0.289	0.27
Mean p-value:		0.49

Fifth Factor		
Item	Factor Loading	P-value
25	0.552	0.31
24	0.338	0.28
28	0.506	0.14
Mean p-value:		0.24

REFERENCES

Abedi, J., & Lord, C. (2001). The language factor in mathematics tests. *Applied Measurement in Education, 14*, 219-234.

Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement, 29*, 67-91.

Ackerman, T. A. (1994). Using multidimensional item response theory to understand what items and tests are measuring. *Applied Measurement in Education, 7*, 255-278.

Ackerman, T. A. (1996). Graphical representation of multidimensional item response theory analyses. *Applied Psychological Measurement, 20*, 311-329.

Ackerman, T. A., Gierl, M. J., & Walker, C. M. (2003). Using multidimensional item response theory to evaluate educational and psychological tests. *Educational Measurement: Issues and Practice, 22*(3), 37-53.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Andrich, D. (2004). Controversy and the Rasch model: A characteristic of incompatible paradigms? In E. V. Smith Jr. & R. M. Smith (Eds.), *Introduction to Rasch measurement: Theory, models and applications*. Maple Grove, MN: JAM Press.

Ansley, T. N., & Forsyth, R. A. (1985). An examination of the characteristics of unidimensional IRT parameter estimates derived from two-dimensional data. *Applied Psychological Measurement, 9*, 37-48.

Assessment Systems Corporation. (1996). XCALIBRE: Marginal maximum likelihood estimation program, Version 1.10. St. Paul, MN: Author.

Baker, F. B. (1987). Methodology review: Item parameter estimation under the one-, two- and three-parameter logistic models. *Applied Psychological Measurement, 11*, 111-141.

Bejar, I. J. (1980). A procedure for investigating the unidimensionality of achievement tests based on item parameter estimates. *Journal of Educational Measurement, 17*, 283-296.

Bejar, I. J. (1983). *Achievement testing: Recent advances*. Beverly Hills, CA: Sage.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate:A practical powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B, 57*, 289-300.

Beretvas, S. N., & Williams, N. J. (2004). The use of hierarchical generalized linear model for item dimensionality assessment. *Journal of Educational Measurement, 41*, 379-395.

Berger, M. P. F., & Knol, D. L. (1990). *On the assessment of dimensionality in multidimensional item response theory models* (No. 90-8). Enschede, Netherlands: Twente University.

Bock, R. D., Gibbons, R., & Muraki, E. (1988). Full-information item factor analysis. *Applied Psychological Measurement, 12*, 261-280.

Bock, R. D., Gibbons, R., Schilling, S. G., & Muraki, E. (1999). *TESTFACT 3: Test scoring, item statistics, and full-information item factor analysis*. Chicago, IL: Scientific Software International.

Bogan, E. D., & Yen, W. M. (1983, April). *Detecting multidimensionality and examining its effects on vertical equating with the three parameter logistic model.* Paper presented at the annual meeting of the American Education Research Association, Montreal, Canada.

Briggs, D. C., & Wilson, M. (2004). An introduction to multidimensional measurement using Rasch models. In E. V. Smith Jr. & R. M. Smith (Eds.), *Introduction to Rasch measurement: Theory, models and applications* (pp. 322-341). Maple Grove, MN: JAM Press.

Burdick, D., Stenner, A. J., & Kyngdon, A. (2007). From model to measurement with dichotomous items: Unpublished manuscript.

California Department of Education. (1985). *Mathematics framework for California public schools: Kindergarten through Grade Twelve*. Sacremento, CA: Author.

Camilli, G., Wang, M., & Fesq, J. (1995). The effects of dimensionality on equating the law school admission test. *Journal of Educational Measurement, 32*, 79-96.

Carpenter, T. P., Corbitt, M. K., Kepner, H. S., Linquist Jr., M. M., & Reys, R. E. (1980). Solving verbal problems: Results and implications from national assessment. *Arithmetic Teacher, 28*(1), 8-12.

Carpenter, T. P., & Lehrer, R. (1999). Teaching and learning mathematics with understanding. In E. Fennema & T. A. Romberg (Eds.), *Mathematics classrooms that promote understanding* (pp. 19-32). Mahwah,NJ: Lawrence Erlbaum.

Carroll, J. B. (1945). The effect of difficulty and chance success on correlations between items or between tests. *Psychometrika, 10*, 1-19.

Chen, W. H. (1993). *IRT-LD: A computer program for the detection of pairwise local dependence between test items*. Chapel Hill, NC: University of North Carolina at Chapel Hill: Thurstone Psychometric Laboratory.

Chen, W. H., & Thissen, D. (1997). Local dependence for item pairs using item response theory. *Journal for Educational and Behavioral Statistics, 22*(3), 265-289.

Chin, T., Kim, W., & Nering, M. L. (2006, April). *Five statistical factors that influence IRT vertical scaling.* Paper presented at the annual meeting of National Council of Measurement in Education, San Francisco, CA.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Belmont, CA: Wadsworth.

De Champlain, A. F., & Tang, L. (1997). CHIDIM: A Fortran program for assessing the dimensionality of binary item responses based on McDonald's nonlinear factor analytic model. *Educational and Psychological Measurement, 57*, 174-178.

Doody, E. (1985, April). *Examining the effects of multidimensional data on ability and item parameter estimation using the three parameter logistic model.* Paper presented at the American Educational Research Association, Chicago, IL.

Dorans, N. J., & Lawrence, I. M. (1999). *The role of the unit of analysis in dimensionality of assessment*. Princeton, NJ: Educational Testing Service.

Drasgow, F., & Parsons, C. K. (1983). Application of unidimensional item response theory models to multidimensional data. *Applied Psychological Measurement, 7*, 189-199.

Edwards, M. C., & Vevea, J. L. (2006). An empirical Bayes approach to subscore augmentation: How much strength can we borrow? *Journal for Educational and Behavioral Statistics, 31*, 241-260.

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

Ferrini-Mundy, J., & Martin, W. G. (2003). Using research in policy development: The case of the National Council of Teachers of Mathematics' *Principles and Standards for School Mathematics*. In J. Kilpatrick, W. G. Martin & D. Schifter (Eds.), *A Research Companion to Principles and Standards for School Mathematics* (pp. 395-413). Reston, VA: NCTM.

Feuer, M. J., Holland, P. W., B.F., G., Bertenthal, M. W., & Hemphill, F. C. (Eds.). (1999). *Uncommon measures: Equivalence and linkage among educational tests*. Washington, DC: National Academy Press.

Finch, H. (2006). Comparison of the performance of Varimax and Promax rotations: Factor structure recovery for dichotomous items. *Journal of Educational Measurement, 43*, 39-52.

Finch, H., & Habing, B. (2005). Comparison of NOHARM and DETECT in item cluster recovery: Counting dimensions and allocating items. *Journal of Educational Measurement, 42*, 149-169.

Fraser, C., & McDonald, R. P. (1988). *NOHARM: An IBM PC computer program for fitting both unidimensional and multidimensional normal ogive models of latent trait theory*. Armidale, Australia: The University of New England.

Froelich, A., & Habing, B. (2001, April). *Refinements of the DIMTEST methodology for testing unidimensionality and local independence.* Paper presented at the annual meeting of the National Council for Measurement in Education, Seattle, WA.

Gessaroli, M. E., & De Champlain, A. F. (1996). Using the approximate Chi-Square statistic to test the number of dimensions underlying the repsonses to a set of items. *Journal of Educational Measurement, 33*, 157-179.

Gierl, M. J., Tan, X., & Wang, C. (2005). *Identifying content and cognitive dimensions on the SAT* (No. College Board Research Report 2005-11). New York: The College Board.

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

Habing, B., & Roussos, L. A. (2003). On the need for negative local item dependence. *Psychometrika, 68*, 435-451.

Hambleton, R. K. (1993). Principles and selected applications of item repsonse theory. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed.). Washington, DC: American Council on Education.

Hambleton, R. K., & Rovinelli, R. J. (1986). Assessing the dimensionality of a set of test items. *Applied Psychological Measurement, 10*, 287-302.

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

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

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

Hattie, J. (1984). An empirical study of various indices for determining unidimensionality. *Multivariate Behavioral Research, 19*, 49-78.

Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement, 9*, 139-164.

Hattie, J., Krakowski, K., Rogers, H. J., & Swaminathan, H. (1996). An assessment of Stout's index of essential unidimensionality. *Applied Psychological Measurement, 20*, 1-14.

Hirsh, T. M. (1989). Multidimensional equating. *Journal of Educational Measurement, 26*, 337-349.

Hofstetter, C. H. (2003). Contextual and mathematics accomodation test effects for English-language learners. *Applied Measurement in Education, 16*, 159-188.

Holmes, S. E. (1982). Unidimensionality and vertical equating with the Rasch model. *Journal of Educational Measurement, 19*, 139-147.

Jang, E. E., & Roussos, L. A. (in press). Nonparametric dimensionality analysis of TOEFL. *Journal of Educational Measurement*.

Kilpatrick, J., Swafford, J., & Findell, B. (Eds.). (2001). *Adding it up: Helping children learn mathematics*. Washington, DC: National Academy Press.

Kim, H. R. (1994). New technique for the dimensionality assessment of standardized test data. *Dissertation Abstracts International, 55-12B*, 5598.

Kirisci, L., Hsu, T. C., & Yu, L. (2001). Robustness of item parameter estimation programs to assumptions of unidimensionality and normality. *Applied Psychological Measurement, 25*, 146-162.

Kline, R. B. (2005). *Principles and practice of structural equation modeling* (2nd ed.). New York: Guilford.

Knol, D. L., & Berger, M. P. F. (1991). Empirical comparison between factor analysis and multidimensional item response models. *Applied Behavioral Research, 26*, 457-477.

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

Kolen, M. J., & Whitney, D. R. (1982). Comparison of four procedures for equating the tests of general educational development. *Journal of Educational Measurement, 19*, 279-293.

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

Lawrence, I. M., & Dorans, N. J. (1987, April). *An assessment of the dimensionality of the SAT-Mathematical.* Paper presented at the annual meeting of the National Council on Measurement in Education, Washington, DC.

Linacre, J. M. (1998). Structure in Rasch models: Why principal component analysis? *Rasch Measurement Transactions, 12*, 266-283.

Linacre, J. M. (2005). A user's guide to Winsteps Ministep Rasch-Model computer programs [Computer software and manual]. Chicago, IL: Winsteps.

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

Lissitz, R. W., & Huynh, H. (2003). *Vertical equating for state assessments: Issues and solutions in determination of adequate yearly progress and school accountability*. Retrieved September 5, 2006 from http://PAREonline.net/getvn.asp?v=8&n=10

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

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

Luecht, R. M., & Miller, T. R. (1992). Unidimensional calibrations and interpretations of compositive traits for multidimensional tests. *Applied Psychological Measurement, 16*, 279-293.

MacCallum, R. C. (2004). *Course pack for Factor Analysis PSYC 236*. Chapel Hill: University of North Carolina.

Martineau, J., Subedi, D., Ward, K., Li, T., & Diao, Q. (2006, October). *Non-linear unidimensional scale trajectories through multidimensional content spaces.* Paper presented at the Assessing and modeling cognitive development in school: Intellectual growth and standard setting, College Park, MD.

McDonald, R. P. (1967). Nonlinear factor analysis, *Psychometrika Psychometric Monographs* (Vol. 15).

McDonald, R. P. (1981). The dimesionality of tests and items. *British Journal of Mathematical and Statistical Psychology, 34*, 100-117.

McDonald, R. P., & Mok, M. C. (1995). Goodness of fit in item response models. *Multivariate Behavioral Research, 30*, 23-40.

McLeod, D. B., Stake, R. E., Schappelle, B. P., Mellissinos, M., & Gierl, M. J. (1996). Setting the standards: NCTM's role in the reform of mathematics education. In S. A. Raizen & E. D. Britton (Eds.), *Bold ventures: Case studies of US innovations in mathematics education* (Vol. 3). Boston: Kluwer.

McLeod, L. D., Swygert, K. A., & Thissen, D. (2001). Factor analysis for items scored in two categories. In D. Thissen & H. Wainer (Eds.), *Test scoring* (pp. 189-216). Mahwah, NJ: Lawrence Erlbaum.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed.). Phoeniz, AZ: Oryx Press.

Messick, S. (1993). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 13-104). Washington, DC: American Council on Education.

Miller, T. R., & Hirsh, T. M. (1992). Cluster analysis of angular data in applications of multidimensional item response theory. *Applied Measurement in Education, 5*, 193-212.

Millman, J., & Greene, J. (1989). The specification and development of tests of achievement and ability. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed.). Washington, DC: American Council on Education.

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

Mislevy, R. J., & Bock, R. D. (1984). BILOG. Moooresville, IN: Scientifici Software, Inc.

Muthen, B. (1993). Goodness of fit with categorical and other non-normal variable. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 205-243). Newbury Park, CA: Sage.

Muthen, L. K., & Muthen, B. (1998). *Mplus: The comprehensive modeling program for applied researchers: User's guide*. Los Angeles, CA: Muthen & Muthen.

Nandakumar, R. (1991). Traditional dimensionality versus essential dimensionality. *Journal of Educational Measurement, 28*, 99-117.

Nandakumar, R. (1993). Assessing essential unidimensionality of real data. *Applied Psychological Measurement, 17*, 29-38.

Nandakumar, R. (1994). Assessing dimensionality of a set of item responses: Comparison of different approaches. *Journal of Educational Measurement, 31*, 17-35.

Nandakumar, R., & Stout, W. (1993). Refinements of Stout's procedure for assessing latent trait dimensionality. *Journal of Educational Statistics, 18*(1), 41-68.

National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics*. Reston,VA: Author.

National Council of Teachers of Mathematics. (2006). *NCTM Statement of Beliefs*. Retrieved June 18, 2006

Pedhazur, E. J., & Schmelkin, L. P. (1991). *Measurement, analysis, design and analysis: An integrated approach*. Hillsdale, NJ: Lawrence Erlbaum.

Perkhounkova, Y., & Dunbar, S. B. (1999, April). *Influences of item content and format on the dimensionality of tests combining multiple-choice and open-response items: An application of the Poly-DIMTEST procedure.* Paper presented at the annual meeting of the American Educational Research Association, Montreal, Canada.

Phillips, S. E., & Mehrens, W. A. (1987). Curricular differences and unidimensionality of achievement test data: An exploratory analysis. *Journal of Educational Measurement, 24*, 1-16.

Preacher, K. J., & MacCallum, R. C. (2003). Repairing Tom Swift's electric factor analysis machine. *Understanding Statistics, 2*(1), 13-43.

Raiche, G. (2005). Critical eigenvalue sizes in standardized residual principal components analysis. *Rasch Measurement Transactions, 19*, 1012.

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal for Educational Statistics, 4*(3), 207-230.

Reckase, M. D. (1987, April). *A comparison of the results of applying several different unidimensional IRT estimation procedures to multidimensional item response data.* Paper presented at the annual meeting of the American Educational Research Association, Washington, DC.

Reckase, M. D. (1990, April). *Unidimensional data from multidimensional tests and multidimensional data from unidimensional tests.* Paper presented at the annual meeting of the American Educational Research Association, Boston, MA.

Reckase, M. D. (1997). The past and future of multidimensional item response theory. *Applied Psychological Measurement, 21*, 25-36.

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

Reckase, M. D., & Ackerman, T. A. (1986, April). *Building a test using items that require more than one skill to determine a correct answer.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Reckase, M. D., Ackerman, T. A., & Carlson, J. E. (1988). Building a unidimensional test using multidimensional items. *Journal of Educational Measurement, 25*, 193-203.

Reckase, M. D., Carlson, J. E., & Ackerman, T. A. (1985, June). *When unidimensional data are not unidimensional.* Paper presented at the annual meeting of the Psychometric Society, Nashville, TN.

Reckase, M. D., Davey, T., & Ackerman, T. A. (1989, April). *Similarity of the multidimensional space defined by parallel forms of a mathematics test.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Reckase, M. D., & Li, T. (2006, October). *Estimating growth when content specifications change: A multidimensional IRT approach.* Paper presented at the Maryland Assessment Research Center for Education Success (MARCES) conference on

"Assesssing and modeling cognitive development in school: Intellectual growth and standard setting" College Park, MD.

Reckase, M. D., & Martineau, J. (2004). *The vertical scaling of science achievement tests*. Commissioned paper prepared for the National Research Council's Committee on Test Design for K-12 Science Achievement, Washington, DC.

Rittle-Johnson, B., Siegler, R. S., & Alibali, M. W. (2001). Developing conceptual understanding and procedural skill in mathematics: An iterative process. *Journal of Educational Psychology, 93*, 346-362.

Romberg, T. A. (Ed.). (1995). *Reform in school mathematics and authentic assessment*. Albany, NY: State University of New York Press.

Roussos, L. (1992). Hierarchical agglomerative clustering computer programs manual: Urbana-Champaign: University of Illinois, Department of Statistics

Roussos, L., Stout, W., & Marden, J. L. (1998). Using new proximity measures with hierarchical cluster analysis to detect multidimensionality. *Journal of Educational Measurement, 35*, 1-30.

Sass, D. A., & Walker, C. M. (2006, April). *Item parameter's influence on multidimensionality detection using DIMTEST.* Paper presented at the annual meeting of the National Council of Measurement in Education, San Francisco, CA.

Schaeffer, G. A., & Kingston, N. M. (1988). *Strength of the analytical factor of the GRE general test in several subgroups: A full-information factor analysis approach* (No. GRE Board Professional Report No. 86-7P; ETS Research Report 88-5). Princeton, NJ: Educational Testing Service.

Seraphine, A. E. (2000). The performance of DIMTEST when latent trait and item difficulty distributions differ. *Applied Psychological Measurement, 24*, 82-94.

Skaggs, G., & Lissitz, R. W. (1981, April). *Test equating: Relevant issues and a review of recent research.* Paper presented at the annual meeting of the American Educational Research Association, Los Angeles, CA.

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

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

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

Smith Jr., E. V. (2004). Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. In E. V. Smith Jr. & R. M. Smith (Eds.), *Introduction to Rasch measurement: Theory, models and applications* (pp. 575-599). Maple Grove, MN: JAM Press.

Smith Jr., E. V., & Smith, R. M. (Eds.). (2004). *Introduction to Rasch measurement: Theory, models and applications*. Maple Grove, MN: JAM Press.

Smith, R. M., & Miao, C. Y. (1994, April). *Assessing unidimensionality for Rasch measurement.* Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Steele, C. M. (1997). A threat in the air: How stereotypes shape intellectual identity and performance. *American Psychologist, 52*, 613-629.

Stipek, D. J. (1987). *Motivation to learn: From theory to practice* (2nd ed.). Boston: xxx.

Stocking, M., & Eignor, D. R. (1986). *The impact of different ability distributions on IRT pre-equating* (ETS Report RR-86-49). Princeton, NJ: Educational Testing Service.

Stone, C. A., & Yeh, C. (2006). Assessing the dimensionality and factor structure of multiple-choice exams: An empirical comparison of methods using the multistate bar examination. *Educational and Psychological Measurement, 66*, 193-214.

Stout, W. (1987). A nonparametric approach for assessing latent trait dimensionality. *Psychometrika, 52*, 589-617.

Stout, W. (1990). A new item response theory modeling approach with applications to unidimensionality assessment and ability estimates. *Psychometrika, 55*, 293-325.

Stout, W. (2002). Psychometrics: From practice to theory and back. *Psychometrika, 67*, 485-518.

Stout, W., Habing, B., Douglas, J., Kim, H. R., Roussos, L., & Zhang, J. (1996). Conditional covariances-based nonparametric multidimensionality assessment. *Applied Psychological Measurement, 20*, 331-354.

Tabachnick, B. G., & Fidell, L. S. (2001). *Using multivariate statistics* (4th ed.). Boston, MA: Allyn and Bacon.

Tanaka, J. S. (1993). Multifacted concepts of fit in structural equation models. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 10-39). Newbury Park, CA: Sage.

Tate, R. (2002). Test dimensionality. In D. Tindal & T. M. Haladyna (Eds.), *Large-scale assessment programs for all students: Validity, technical adequacy, and implementation* (pp. 181-211). Mahwah, NJ: Lawrence Erlbaum.

Tate, R. (2003). A comparison of selected empirical methods for assessing the structure of responses to test items. *Applied Psychological Measurement, 27*, 159-203.

The William Stout Institute for Measurement. (2005). DIMTEST (Version 2.0) [Computer software and manual]. St. Paul, MN: Assessment Systems Corporation.

Thissen, D. (1991). MULTILOG Version 6.0 user's guide. Chicago, IL: Scientific Software.

Thissen, D., & Wainer, H. (Eds.). (2001). *Test scoring*. Mahwah, NJ: Lawrence Erlbaum.

Thompson, S., Johnstone, C. J., & Thurlow, M. L. (2002). *Universal design applied to large scale assessments* (Synthesis Report 44). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Traub, R. E. (1983). A priori considerations in choosing an item response model. In R. K. Hambleton (Ed.), *Applications of item response theory* (pp. 57-70). Vancouver, Canada: Educational Research Institute of British Columbia.

Traub, R. E., & Lam, Y. R. (1985). Latent structure and item sampling models for testing. *Annual Review of Psychology, 36*, 19-48.

van Abswoude, A. A. H., van der Ark, L. A., & Sijtsma, K. (2004). A comparative study of test data dimensionality assessment procedures under non parametric IRT models. *Applied Psychological Measurement, 28*, 3-24.

van der Linden, W. J., & Hambleton, R. K. (Eds.). (1997). *Handbook of modern item response theory*. New York: Springer.

Wainer, H., & Thissen, D. (1996). How is reliability related to the quality of test scores? What is the effect of local dependence on reliability? *Educational Measurement: Issues and Practice, 15*(2), 22-29.

Wainer, H., & Wang, X. (2001). *Using a new statistical model for testlets to score TOEFL.* Princeton, NJ: Educational Testing Service.

Walker, C. M., & Beretvas, S. N. (2000, April). *Using multidimensional versus unidimensional ability estimated to determine student proficiency in mathematics.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Way, W. D., Ansley, T. N., & Forsyth, R. A. (1988). The comparative effects of compensatory and noncompensatory two-dimensional data on unidimensional IRT estimates. *Applied Psychological Measurement, 12*, 239-252.

Wingersky, M. S., Barton, M. A., & Lord, F. M. (1982). LOGIST user's guide. Princeton, NJ: Educational Testing Service.

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

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

Zhang, J. (1996). Some fundamental issues in item response theory with applications: Urbana-Champaign: University of Illinois, Department of Statistics.

Zhang, J., & Stout, W. (1999). The theoretical DETECT index of dimensionality and its application to approximate simple structure. *Psychometrika, 34*, 213-249.