ROXBURY, TIESE L., Ph.D. A Psychometric Evaluation of a State Testing Program: Accommodated Versus Non-accommodated Students. (2010)
Directed by Dr. Terry A. Ackerman. 92pp.

Federal legislation such as *No Child Left Behind* mandated that students with disabilities be included in accountability standards, creating an important responsibility to fairly assess all students, even those with disabilities. Consequently, a sense of urgency was placed on the entire educational system to ensure that these students had a fair chance at being adequately tested, to ensure equity and access. Alternate assessments and accommodations are a part of access for these students in the testing realm. This study, which centered around fairness, focused on psychometrically comparing three subject exams administered to eighth grade students who received accommodations as opposed to those who did not. Covariance matrices, dimensionality, factor structure, and item functioning were all compared across groups to examine invariance and show that the use of accommodations did not affect the validity or fairness of the testing program.

Analyses revealed that for each of the tests, there was a significant difference in the covariance matrices, but unidimensionality held across groups implying that the dimensionality structure was the same for both groups. The factor structure was tested only for the math exam and the one-factor structure held for accommodated and non-accommodated students, again confirming unidimensionality. Despite consistent dimensions being measured for both groups, DIF did manifest but without actual test items it could not be attributed to bias.

A PSYCHOMETRIC EVALUATION OF A STATE TESTING PROGRAM: ACCOMMODATED VERSUS NON-ACCOMMODATED STUDENTS

by

Tiese L. Roxbury

A Dissertation Submitted to
the Faculty of The Graduate School at
The University of North Carolina at Greensboro
in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Greensboro
2010

Approved by

_____
Committee Chair

To everyone who listened to me during my doctoral journey and showed unwavering confidence in me and my ability to make it through to the light at the end of the tunnel. To all who offered their prayers and words of encouragement that motivated me to push even harder through this long process to make you all proud.

APPROVAL PAGE

This dissertation has been approved by the following committee of the Faculty of The Graduate School at The University of North Carolina at Greensboro.

Committee Chair _____

Committee Members _____

_____

_____

_____
Date of Acceptance by Committee

_____
Date of Final Oral Examination

ACKNOWLEDGEMENTS

I would like to acknowledge the brilliant guidance of Dr. Terry Ackerman, my advisor and dissertation committee chair. A special thanks to my other committee members, Dr. Robert Henson, Dr. Jewell Cooper, and Dr. Luz Bay. Each of you brought a very unique background and expertise to this study which allowed it to flourish and change. Your input has been valued more than you will ever know. I would also like to thank my former classmate, Dr. Joshua Goodman, who has always been just an email away. Each and every one of you has graced me with your genius, flexibility, and encouragement.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER I

INTRODUCTION

The implementation of *No Child Left Behind* has placed more responsibility on the major facets of the educational system—schools, teachers, and students. Regulations have made improvement and growth a requirement creating ambivalent feelings toward testing and accountability. Needless to say the demands of accountability are viewed both positively and negatively by many. Testing has been continually dissected to find its worth and its flaws. Standardized tests are sometimes considered high stake because results are used to make important decisions about graduation exit requirements, college admissions, certification, and licensure, thus making validity a central issue.

Assessing students to see what they know and what they have learned is logical because it informs teachers about the effectiveness and success of educating students, yet people still contest that testing is unfair and ineffective. In turn, testing companies feel the heavy burden of responsibility when it comes to standardized tests because test makers have the duty of making sure that tests prove to be fair for all students. Fairness is one of the biggest issues concerning testing. Fairness should extend across gender, race, socioeconomic status (SES), and exceptionalities implying that only ability should be the deciding factor in test scores. Several testing opponents such as Berlak (2001) often refer to standardized tests as "culturally and racially biased" declaring that the bias is "lodged in the content or language of individual test items." He places heavy emphasis

on the racial achievement gap, which should undoubtedly be of great concern. But since *No Child Left Behind* mandates accountability, eliminating testing is not a favorable option, but rather a solution that would eliminate the tool which has the power to reveal the achievement gaps that exist across different subgroups.

In addition to research on race and testing, there has been much gender research pertaining to standardized testing because the topic of differential performance between males and females has been a long standing controversial issue. Research in this area continues as the achievement gap between males and females still exists. However, this gap is not consistent across subjects. For years males have been said to be better at math and science whereas females outperform males in reading and writing. There is no quick solution to this gap since a compound of factors contributes to this gap—motivation, teacher expectations, SES, etc. SES is also a variable by which testing populations are often grouped and compared. This study will use data that contains demographic information such as ethnicity, gender, and SES, but the main subgroup of interest here is exceptionality. Exceptionalities range from academically gifted to severely disabled. However, this study will specifically focus those students who are learning disabled but still take traditional assessments as opposed to an alternate assessment.

When looking at race and gender of students with disabilities, the trend of disproportionate representation of minority students in special education emerges and is a long standing issue (Beratan, 2008). More specifically, African American male students are overrepresented in the learning disabled population and underrepresented in those classified as gifted (Hosp & Reschly, 2004). This problem has been in existence for the

past four decades and much research has been done on the issue (Hosp & Reschly, 2004; Salend, Duhaney, & Montgomery, 2002). Research to analyze this concern has stemmed from the disproportionate ratios of African Americans, American Indians, and Latinos to Caucasians. Such patterns should prompt research on the identification process of special education students. Hosp & Reschly (2004) discussed how the disproportionate number of minority students in special education was a result of minority students consistently performing significantly lower on achievement measures. Given that poverty and academic achievement have been identified as one of the most important predictors of identification for special education services, minorities are overidentified as needing these services more so than Whites, who tend to be overrepresented in gifted programs. As a result, minorities are a large portion of the students who are labeled with disabilities. The particular notion of proportionality was examined briefly within the data sample of this study.

Improvement of testing practices warrants extensive innovative research in the areas of psychometrics, validity, and test taking populations. Furthering research in this area will undoubtedly lead to more critics, but more importantly, to more plausible solutions as well. Psychometric research has already led to the development of several DIF detection procedures, which flag potentially biased test items. These procedures aim to identify and eliminate malfunctioning items from tests. So even though testing is a controversial topic, procedures are in place to make sure that items remain invariant across subgroups with respect to irrelevant aspects of the testing population. Addressing

issues such as fairness and test validity is a popular way to substantiate that test scores are valid and can be interpreted accurately.

When it comes to students who are classified as exceptional, the task of creating a set of fair standardized tests becomes more complex as they must be appropriate for the targeted population. Because of accountability standards, mandated end-of-course (EOC) and end-of-grade (EOG) tests are now standard. As the accountability standards were raised, states have had to adopt alternate ways of assessing students who were not able to take traditional assessments. However only up to 2% of a state's student population is allowed to achieve proficiency for Adequate Yearly Progress (AYP) by taking an alternate assessment (U.S. Department of Education, 2007). Alternatives such as projects and portfolios are still popular and widely used, even though they are subject to inter-rater reliability issues. However, all students with exceptionalities are not qualified to take alternate assessments and are sometimes mainstreamed and end up taking traditional assessments with accommodations, reinforcing the need for these tests to be well-designed, reliable, and fair. Just as alternate assessments must evolve, the standardized traditional assessments also have to be continually researched and improved as students with disabilities are no longer exempt from the accountability system. It is very important that when students with disabilities are tested, they are given the fair opportunity to demonstrate concept mastery, whether this be through a traditional or alternate assessment. Promoting equity in this way demonstrates that students with mild learning disabilities to those with severe cognitive disabilities are capable of being tested and that we are able to adequately measure their abilities.

What exactly makes testing students with disabilities fair and equitable? First is access to a universally designed education so that they are properly taught before being tested. As described by the Center for Universal Design (CUD), universal design (UD) is a term that originated in architecture by Ron Mace in order to emphasize creating buildings that are usable by all people (King-Sears, 2009). Burgstahler (2009) notes that UD in education stresses diversity and inclusion. The guiding principles of UD that have transferred from architecture to education include: equitable use, flexibility in use, perceptible information, simple and intuitive use, tolerance for error, low physical effort, and, size and space for approach and use (Burgstahler, 2009; King-Sears, 2009). Putting these principles to use in a classroom setting allow for all students, including students with learning disabilities, to have a more natural acclimation to the educational environment (King-Sears, 2009). Although these changes can occur to the physical classroom setting, King-Sears (2009) and Passman and Green (2009) reinforce that the principles should be applied to pedagogy and syllabi as well, so that delivery of instruction is considered universally designed and suitable for all students.

Another way to ensure fairness is the concept of testing individualization, which is an attempt to cater to the learning and testing needs of particular students. As students with exceptionalities often learn differently from those without, they are consequently tested differently. Alternate assessments are an attempt to fairly measure the knowledge base of students with severe cognitive disabilities. Standard achievement tests paired with accommodations are another attempt to equalize the testing process and assist students with exceptionalities. Accommodations normally include options such as

extended test time, modified test items, a different testing environment, or read-aloud test administration. These options are chosen based on the examinee's particular learning needs, which help to articulate the direct link between how a student learns and how the student is assessed. It must also be noted that adding these options should not create an unfair advantage for those who receive the accommodations.

When students are assessed, they have to take what they have learned and apply that knowledge in a testing situation. Tests are designed to measure a specific set of academic skills and all other skills are seen as "invalid" and therefore are not constructs of interest (Ackerman, 1994). For example, consider a math test designed to measure if a student has mastered multiplication. If the multiplication problem is hidden in a word problem, a student with a learning disability may have trouble reading and sifting through the language to find the appropriate mathematical information necessary to multiply; or, this student may just take a little longer to read through the problem to fully comprehend the objective. If this student is provided with the appropriate accommodations (e.g., sheltered English or extended time) then he/she will be better able to demonstrate if he/she has mastered multiplication. With multiplication being the "valid" skill intended to be measured, reading ability should not confound the student's ability to correctly answer the test items. Therefore, it is necessary to be clear on what a test is intended to measure in order for scores to be valid and appropriately interpreted.

Psychometric techniques allow for validation of achievement tests. Item response theory (IRT) is a probabilistic way of describing how an examinee interacts with an item, assuming that one skill or the same composite of skills is being used to answer each item

on the test (Ackerman, 1994). However, this notion of unidimensionality is improbable, especially when looking at students who have disabilities versus those who do not. These groups of students learn differently, making it plausible that they may interact with test items differently. Students with exceptionalities may use different sets of skills to compensate for a lack of one skill, which leads to the notion of multidimensionality. Under the assumptions of multidimensional item response theory (MIRT), items on a given assessment may actually measure different composite of abilities. This is typically not problematic as long as the assessment is primarily measuring the same or a similar composite for all students. On some assessments the student-item interaction could result in different composites of ability being measured for different groups of students—in this case students with exceptionalities and those without.

The purpose of this research is to extend the literature on the fairness of testing of students with disabilities and exceptionalities by using a framework that incorporates differential item functioning (DIF) and MIRT dimensionality techniques. Although a lot of research has been done on special education (Bachor, 1990; Byrnes, 2008; Salvia, Ysseldyke, & Bolt, 2006), alternate assessments, and accommodations, this study intends to add to the literature on students with disabilities who take regular assessments with accommodations. The substantial amount of research previously done on these topics does not show consistent findings (Abedi, Leon, and Kao, 2008; Willingham et al., 1988), meaning more work is still needed in the area. Emerging literature has generally used DIF alone and does not combine psychometric techniques as this study will attempt to do. Reckase (1997) succinctly summarized the connection between MIRT and DIF:

> MIRT has been used to help understand the skills required to successfully respond to test items, the extraneous examinee characteristics that affect the probability of response to items (DIF), and the complexities behind equating test forms, among other applications.

For this reason, MIRT software will be used to examine differential dimensionality to reveal the latent abilities used by the different groups, though possibly shedding light on how to better educate them. If the dimensionality is found to be different, then research explaining the possible causes of this difference can be done. In addition to dimensional analyses, a DIF analysis will be performed to further differentiate between these two populations and how they interact with specific items on a standardized assessment. These methodologies were selected because they were most appropriate for answering the research questions.

Research Questions

The current study has its basis around the essential topic of testing fairness and validity for one state's eighth grade educational assessment program. A technical report has used the complete data to evaluate the validity, dimensionality, and other areas of the assessment program for grades 3-8 in State A. However this study focuses on a specific grade level and a specific split of the students. Only data from grade eight was used because this grade level is known as a transition period that is extremely pivotal (Gutman, 2006; Isakson & Jarvis, 1999; Neild, Stoner-Eby, & Furstenburg, F, 2008; Reyes, Gillock, Kobus, & Sanchez, 2000). Through the use of this data this study will address the following research questions:

1. Is there a statistically significant difference in the covariance structure for students who were accommodated and for those who were not?
2. Does the assessment measure the same dimensional construct(s) for those students who have disabilities and use accommodations versus those who do not? Is there differential dimensionality across the groups?
3. Is there measurement invariance across the two groups (accommodated versus non-accommodated students)? Does the same factor structure hold?
4. Is there a difference in item functioning for those students with disabilities who received accommodations versus those without?

The results are intended for test makers, psychometricians, state officials, teachers, and anyone else involved with a standardized testing system. Besides emphasizing fair testing, this study intends to place an emphasis on identifying students who are disabled and need to take alternate assessments or receive accommodations. This front end process is also very critical in establishing the validity of testing programs for these students. If students are incorrectly identified or given the wrong accommodations, test results will be flawed and could result in a loss of internal and external validity.

CHAPTER II

REVIEW OF THE LITERATURE

Recent legislature has placed heavy demands on the field of educational testing. With the 1997 amendments to the Individuals with Disabilities Education Act (IDEA) and the implementation of *No Child Left Behind* in 2001, equity and accountability have become mandated for all students. Across each state, standardized achievement tests in reading, mathematics, and science are mandated for students in all districts. States have a certain level of autonomy on the specific assessment program they choose to administer. However, federal accountability unites the states by holding them all to one common metric—Adequate Yearly Progress (AYP), even though specific computations used by the states are not the same.

In order to comply with federal mandates that address equity education for these students, having alternate ways of assessing students with learning disabilities and severe cognitive disabilities is required for school districts (Goh, 2004; Schafer & Lissitz, 2009; Sireci, 2009). Because of the current legislation, all students with disabilities—mild, significant, or severe—must be accounted for, so different procedures have to be implemented to create a fair testing situation depending on the nature of the student's disability. These approaches can be summed up into three categories—alternate assessments, enhanced assessments, or accommodations to traditional assessments. It is important to know the difference between the three and each will be discussed.

To include students with significant disabilities in Adequate Yearly Progress (AYP), the U.S Department of Education announced in 2003 that states may be held to new alternate achievement standards (Yovanoff & Tindal, 2007). The goal of an alternate assessment is to assist educators in evaluating a more direct and authentic measure of student learning and progress as opposed to a traditional standardized achievement test (Goh, 2004). Thus, when is it appropriate to use alternate assessments? According to previous research, alternate assessments are used when students with disabilities or English Language Learners (ELLs) cannot take tests under standard administration and must be assessed by alternative procedures (Lam, 1993). Students typically needing an alternate assessment are students with severe cognitive disabilities and/or students with minimum English language proficiency who are unable to take tests under standard administration, even if the tests are accompanied by accommodations (Goh, 2004).

Some alternate assessments are very similar to traditional tests and allow students to choose an answer, but others permit a student to demonstrate his or her knowledge learned in the classroom through oral examinations, hands-on assessments, or projects. Bintz & Harste (1994) noted that an authentic assessment, performance assessment, and portfolio assessment are all types of alternate assessments that can be used to measure the achievement of students. According to Goh (2004), performance assessment and authentic assessment are sometimes used interchangeably. A performance assessment is a type of assessment that allows students to perform, demonstrate, and/or develop a product or solution based on the subject material learned. There were five important

elements of the performance assessment identified by Elliott and Fuchs (1997): (1) linking assessment tasks that are clearly aligned or connected to what has been taught to the students, (2) presenting the scoring criteria of the assessment task to students prior to their working on the task, (3) sharing with students standards and modes acceptable and exemplary performance prior to assessment, (4) encouraging students to complete self-assessments of their performances, and (5) comparing students' performance to predetermined standards as well as to the performances of other students. Authentic assessments are similar to performance assessments in which there is a direct focus on the examination of a student's performance and/or work product. In addition to the performance aspect, an authentic assessment measures the behaviors and skills that are required to survive in the real world. Moreover, others also seem to view a portfolio assessment as a subcategory of performance assessment. Portfolio assessments are used to reflect a student's accomplishments by using procedures to plan, collect, and analyze multiple sources of data such as work samples, journal entries, performance, exhibitions, teacher ratings, student assessments, and so on. Darling-Hammond (1994) argued that portfolio assessments take into account the unique characteristics and backgrounds of diverse learners which are tailored to the assessment needs of these students.

For students with disabilities who have Individualized Education Programs (IEPs), alternate testing options with modifications are necessary to promote fairness and equal opportunity. Goh (2004) noted that the traditional psychometric methods of evaluating standardized tests may not apply to alternative assessment strategies and suggests that rethinking the issues of reliability and validity of alternate assessments may

be necessary. It is important that these alternate methods be deemed just as reliable and valid as traditional standardized tests. Addressing validity, Elliot and Roach (2007) referred to Wilson (2004) who discussed the role a teacher's control plays in validating alternate assessments. From Wilson's control chart for assessment, it was concluded that the less control a teacher has in the task specification and evaluative judgments, then there is perceptively more psychometric validity and comparability (Wilson, 2004). Performance and portfolio assessments, along with rating scales of achievement were examined by Elliot and Roach (2007) to analyze the features that influenced the psychometric quality and validity. However, Bintz and Harste (1994) suggested that educators should focus less on the method and more on the attainable outcome of the method and ask what do we want these methods to tell us about learning? Their concentration on the relationship between alternate assessments and the curriculum emphasized that the two are not separate entities and must be properly aligned to equally coexist in order to demonstrate validity (Bintz & Harste, 1994). There is a limited source of information regarding the technical validity of alternate assessments. Goh (2004) noted that alternate assessments have relatively high face validity, but face validity does not assure accurate measurement and test interpretation. It is suggested that an abundant of research is needed to provide other types of validity evidence. Empirical data is needed to make instructional decisions or educational placement decisions. It is also needed to assure that alternate assessments can actually measure high-order thinking.

Creating alternate assessments that demonstrate accommodation validity takes planning and careful alignment to appropriate curriculum standards, hence why the first

study on validating an alternate assessment reported several inadequacies (Johnson and Arnold, 2004). These assessments are meant to be extensions of standard assessments which test the general population, therefore should be designed with the same measurement precision so results will be meaningful. There is also a limited source of information regarding the technical validity of alternate assessments. Goh (2004) noted that alternate assessments have relatively high face validity, but face validity does not assure accurate measurement and test interpretation. It is suggested that an abundant of research is needed to provide other types of validity evidence. Empirical data is needed to make instructional decisions or educational placement decisions. It is also needed to assure that alternate assessments can actually measure high-order thinking.

Not quite considered alternate assessments, enhanced assessments are another type of inclusive assessment that target identified gap students (Salvia, Ysseldyke, & Bolt, 2006). The method is similar to accommodations but a little more involved and state programs are just beginning to explore this alternative. This type of assessment is seemingly more holistic and comprehensive. Grants have been funded specially for enhanced assessment initiatives to improve the quality, reliability and validity of state assessment programs so that all students, even students that fall in the gap, can be fairly tested.

Accommodations are another way testing can be modified to suit the special needs of students with disabilities. They are somewhat less complex because the process does not involve creating an entire new assessment form, but that does not lessen the value of previous and future research done on them. Accommodations are usually given

to students with mild learning disabilities or ELLs and according to Camara (2009), and vary with individual circumstances. Accommodations should be handed out selectively after careful consideration due to the volume of requests for these accommodations (Ofiesh & Bisagno, 2008). They are designed to make testing more accessible to these students and reduce the impact of "invalid" skills to the construct being tested. Byrnes (2008) referred to accommodations as a means of removing the barriers caused by disabilities so that a student with a disability can have the same access as a student without. In other words accommodations create a level playing field because they do not equate to instant success, but rather help for students who need them in order to effectively demonstrate what they have learned (Byrnes, 2008). It is important that accommodations are properly chosen by a qualified individual. Most often they are recommended by a teacher, learning specialist, or through Individualized Educational Program (IEP) teams (Gibson, Haeberli, Glover, & Witter, 2005). Fuchs, Fuchs, and Capizzi (2005) provided a thorough overview of accommodations that addressed validity, but the study emphasized a system that extensively described how to identify accommodations appropriately as the task is complex because there is no one accommodation that is suitable for every student. They presented a tool called DATA, The Dynamic Assessment of Test Accommodations, which presents teachers with a way of identifying which accommodations will result in differential improvements for specified individual students in grades 2-7 for the subjects of reading and math (Fuchs & Fuchs, 2001).

There are several different types of accommodations which can be administered alone or in combination with others and some research suggests that accommodations should be used in "packages" (Elliott, Kratochwill, & McKevitt, 2001). First, changes can simply be made to the content of test items or the format of the test can be altered. This may sometimes be considered a modification, but nevertheless modifications are an attempt to make the test more accessible. Modifications and accommodations are often used interchangeably, but some might note a slight technical difference. However, some modifications are indeed accommodations. Kettler, Elliott, and Beddow (2009) examined how to modify test items in order to get a better sense of what students with disabilities know. The study introduced a new tool that provides educators with a systematic way of modifying grade-level tests so that items would consistently be fair and appropriate (Kettler, Elliott, & Beddow, 2009). Students who qualify for accommodations can also be given an extended testing period, the most frequently used accommodation (Fuchs & Fuchs, 2001). However, this accommodation has proved to increase scores for general and special education students (Fuchs, Fuchs, Eaton, Hamlett, Binkley, & Crouch, 2000) and research shows there has been little evidence to prove that students with disabilities are the sole beneficiary of the testing boosts from having extended time (Tindal & Fuchs, 2000). Students with disabilities who had trouble reading did show large test score gains from extended time (Fuchs, Fuchs, Eaton, Hamlett, & Karns, 2000). Lewandowski, Lovett, and Rogers (2008) also studied the effects of extended time on a reading test and found that the extra time benefited the

students with disabilities and allowed them to attempt as many questions as students without disabilities who took the exam in the standard amount of time allotted.

Another type of accommodation that can occur is a change in administration. There are two instances of this type of accommodation with the first being that an examinee can receive an oral administration of the exam. Orally presenting the test to a student means a teacher reads the test directions and items to a student. Bolt and Thurlow (2006) conducted a longitudinal study across different grade levels that examined the item-level effects of this accommodation on math and reading test items. They found that performance on some items was positively affected, although not so much on others. Other research studies on the effects of this particular accommodation also showed mixed results, but students with disabilities have shown an increase in their math scores when items are read aloud to them (Tindal, Heath, Hollenbeck, Almond, & Harniss, (1998). Fuchs et al. (2000) attempted to prove the validity of this accommodation by showing that oral administration resulted in more score gains for students with disabilities than those without. Huynh and Barton (2006) studied the effects of oral administration on the South Carolina High School Exit Exam (HSEE) by administering the test with the accommodation and without the accommodation to students with disabilities. Results showed that the factorial structure of the test remained stable across the groups and that the accommodation did create equitable testing conditions for those who fittingly qualified (Huynh & Barton, 2006). In a meta-analysis, Elbaum (2007) found that oral accommodations resulted in higher math scores for

students with disabilities than for those without, but found the converse true for secondary students.

The other common administration change is setting. That is, the exam can be administered in an alternate location so that students can be tested in an environment that has fewer students than a regular classroom. The purpose of this is to limit the amount of distractions. This accommodation is often paired with others such as oral administration and extended time. There is little to no empirical research on the effect of an alternate location used as an accommodation. However, if tests are delivered using the proper protocol under equitable standardized conditions as in regular classrooms, then this change should not prove to be an unfair accommodation by falsely driving up scores (Fuchs, Fuchs, & Capizzi, 2005).

Accommodations allow inclusion in the accountability system for those students who were excluded for a long time until recent legislation. For this reason accommodations are directly linked to testing fairness and validity. A fair and valid accommodation has been defined as providing help to students with disabilities so that they can demonstrate what they know on tests that produce meaningful scores and can be evaluated on the same constructs that are measured in the testing of nondisabled peers (Fuchs & Fuchs, 2001; Fuchs, Fuchs, & Capizzi, 2005; Phillips, 1994). Elliott, Kratochwill, McKevitt, and Malecki (2009) allowed accommodations to be rated and deemed fair by testing experts.

Sireci, Scarpati, and Li (2005) performed a thorough review on testing accommodations and the previous research that has been done on them. More

specifically, Lang, Elliott, Bolt, and Kratochwill (2008) examined the effects of testing accommodations on a group of fourth and eighth-grade students, with and without disabilities. In the study they gave all of the students a reading and math test with and without accommodations and also administered a questionnaire to find out the students' perceptions of test accommodations. Results revealed that students perceived accommodations as fair for all students but even more fair for students with disabilities. Also, the accommodations had an overall positive impact on test scores (Lang, Elliot, Bolt, & Kratochwill, 2008).

Willingham (1986) investigated admissions testing and how the procedures could negatively affect those with disabilities if not modified properly. Whereas admissions tests are for the realm of higher education, tests with such high stakes particularly emphasize the importance of fairness for the disabled population. According to Geisinger (1994), most laws concerning educational testing affected tests associated with college admissions. Most standardized assessments are usually attached to high stakes whether it is at the examinee, teacher, school, or district level. Willingham et al. (1988) took his research on "testing handicapped people" a step further by implementing the measurement technique of differential item functioning (DIF) to compare items across students with and without disabilities and there was ultimately no DIF found. The premise behind DIF is that examinees of the same achievement level (latent ability) should have the same probability of correctly responding to an item. Therefore, DIF is said to be present when a certain group of students performs lower on a test item than a comparable reference group of students after controlling for the ability being measured.

These subgroups are commonly divided by conventional splits such as Race/Ethnicity or gender. Ackerman (1992) reasoned that standardized tests must have a level of discrimination to differentiate between examinees of different ability levels. However, this type of discrimination should not be an artifact of culture, race, or gender. Other Reference and Focal group splits have often been researched as well, like the aforementioned split on disability or a split on native language to see whether items show DIF for those students whose primary language is not English (Haugers & Sireci, 2008; Mahoney, 2008; Martiniello, 2008; Sinharay, Dorans, & Liang, 2009).

Abedi, Leon, and Kao (2008) examined DIF in reading assessments for students with disabilities. In this study a logistic regression approach to DIF was used in order to answer the primary research question—do items on a reading assessment exhibit DIF for students with disabilities? They found that a number of items did indeed show signs of uniform and non-uniform DIF for students with disabilities more so in the latter parts of the subscales in the reading exam (Abedi, Leon, & Kao, 2008).

More recent research using DIF with this topic has compared various groups of students with disabilities on exams that range across different content areas (Barton & Finch, 2004; Bolt, 2004; Bolt & Bielinski, 2002; Cahalan-Laitusis, Cook, Cope, & Rizavi, 2004; Camara, 2009; Finch, Barton, & Meyer, 2009). For example Finch, Barton, & Meyer (2009) used the split of accommodated versus non-accommodated students on math and language items and found both uniform and non-uniform DIF. Clearly to this point there have been mixed findings as to whether accommodations create a bias or unfair testing advantage. Although some studies provide evidence to

support the finding, others do not deem this to be the case. A number of the studies using this DIF framework concluded that the DIF was related to a difference in skill and not to the accommodation factor, as the direction of the DIF did not favor any group consistently (Finch, Barton, & Meyer, 2009). This is where DIF becomes intertwined with assessing skills and specified constructs. Camara (2009) reiterated the common and important notion that accommodations should not change the intended construct being measured. This study will examine that notion through the use of dimensionality software, which is rooted in MIRT, to attempt to use a different methodology than in previous research. If constructs change across groups, then this further affects the validity of the assessments and dimensionality is not population invariant.

Over the past years, there has been a lot of research done on the theoretical framework of MIRT. Reckase (1997) summarized the historical antecedents of MIRT with factor analysis and unidimensional item response theory (IRT) being the main predecessors. He followed up describing the works of contributors such as Spearman, Thurstone, Lord and Novick, and Samejima (Reckase, 1997). Ackerman, Gierl, and Walker (2003) looked at the applicability of MIRT and discussed using the method to evaluate educational tests. The underlying assumption is that tests are naturally multidimensional, meaning they often measure more than one construct. A valid construct is what a tests purports to measure and is described by tests developers in detailed test blueprints. Consequently, there are valid and invalid constructs. Items on a given measure usually measure a composite of abilities, all of which may not be intended to be measured as specified by the test blueprint. If an item is not sensitive enough to

measure more than one skill or if examinees vary on the same skills, then the examinee-item interaction will behave unidimensionally (Ackerman, 1992). Thus, dimensionality is not of great importance when the same composite is being measured for all students. However, when studying students with disabilities, this variance in dimensionality could very well be present due to the fact that these students may use different levels of compensation which causes them to interact with the items differently than nondisabled students. This method is informative because it shows if the same constructs for both populations are being measured or it would reveal differential dimensionality and possibly non-valid dimensions. If the same constructs are not being measured consistently then more emphasis needs to be placed on what exactly we are trying to accomplish when testing alternate populations (Bintz and Harste, 1994).

According to the work of Stout, Habing, Douglas, Kim, Roussos, & Zhang (1996), doing a test of multidimensionality serves the purpose of refuting unidimensionality and identifying a test's multidimensional structure. Simple structure is important in detecting multidimensionality and requires that items can be assigned to distinct homogeneous clusters showing that they measure only one dimension. The number of item clusters equals the number of dominant dimensions (Stout et al., 1996). This can be done using factor analytic techniques or by substantive expert review. Determining if a test is factorially simple or complex is one of the purposes of using a multidimensional framework. In MIRT a two-dimensional plane represents a two-dimensional latent space. Analysis of response data and item parameters reveals the item location with respect to the coordinate axes, explaining what items are measuring and

how. The angle of the vector represents the composite, the length of the vector represents the discrimination and the location of the item denotes the item's difficulty.

Stout et al. (1996) discussed three multidimensionality assessment procedures. The first procedure, HCA/CCPROX (Roussos, 1995) is a method referred to as a sorting algorithm that gives several groups of items that could potentially be differ in dimensionality, but does not do a formal statistical test of the distinctness of the items from the rest of the test items (Stout et al., 1996). Unlike HCA/CCPROX, DIMTEST is a statistical hypothesis test that assesses the conditional covariance relationship between two previously identified clusters of items on an exam and determines if they are dimensionally distinct. Moreover, Stout et al. (1996) discussed DETECT (Zhang & Stout, 1999b) as the third procedure which is a "specialized estimation procedure" which is different from the previous two because it measures the amount of multidimensionality. It too uses conditional covariances to partition items into dimensionally homogeneous and distinct clusters. The reported DETECT index is estimated by a complex algorithm and is fairly subjective and can be evaluated by specified criteria (Stout et al., 1996).

Stout's work in the field has allowed for continuous research on the topic of dimensionality using the discussed procedures (Douglas, Kim, Roussos, Stout, & Zhang, 1999; Froelich & Habing, 2008; Nandakumar, 1994). Reckase (1997) suggested that because most educational tests likely assess multiple skills, more research needs to be done in the area of MIRT to establish a solid base for a methodology that is still in its early stages compared to other psychometric techniques. However, MIRT research has

been the gateway that examines the importance of dimensionality research with the expressed purpose of confirming or refuting unidimensionality. If a test is unidimensional for one group, it may or may not be for another group. Thus, the issue of invariant dimensionality is so important for the valid interpretation of test results when multiple subpopulations are involved.

CHAPTER III

METHODOLOGY

Data

Preexisting test data allowed for the development and completion of this study. The data source is a small Northeastern state that is predominantly Caucasian. For purposes of anonymity, this study will refer to the state as State A. The data were obtained from the 2007-08 academic year of State A's traditional educational assessment program and only includes grade eight. The entire sample included 15,274 participants in each of the required subjects—math, reading, and science.

To get a snapshot of the students in this sample, there was a number of demographic variables collected including gender, ethnicity, Title I status, and national school lunch program participation. Of the 15,274 examinees, 7,338 (48.0%) were female and 7,936 (52.0%) were male. A majority of the students were Caucasian (94.7%), whereas the remaining students were Asian/Pacific Islander (N=186; 1.2%), African American (N=368; 2.4%), Hispanic (N=139; 0.9%), or American Indian/Native American (N=120; 0.8%). Only 829 (5.4%) had Title I status and 14,445 (94.6%) did not. Of the total 15,274 students, 5,420 (35.5%) participated in the national school lunch program and the remaining 9,854 (64.5%) did not. In the data 2.1% of the students were listed as currently as having a Limited English Proficiency (LEP).

Per *No Child Left Behind,* each state is required to asses each student in grade levels 3 through 8 and in high school in math, science, and reading. Usually one of three avenues is used: standard administration, standard administration with accommodations, or through alternate assessment. The subgroup of interest was composed of those students who were administered the standard eighth grade subject exams with accommodations. This group was compared with those who did not use accommodations. In all, there were 12,766 (83.6%) students who were not identified as having a disability and 2,508 (16.4%) who were. Some of the specific types of disabilities are attention deficit hyperactivity disorder (ADHD), articulation disorder, auditory processing disorders, dyscalculia, dyslexia, expressive language disorder, and visual processing disorder. Of the 2,508 who were identified as having a disability, 5.6% were minorities. The specific ethnic/racial breakdown is as follows: 0.6% Asian/Pacific Islander; 2.5% African American; 1.2% Hispanic; 1.3% American Indian/Native American; 94.4% Caucasian; 17.3% and 16.4% of all minorities and non-minorities, respectively, in this sample are identified as having a disability. Of the minority population, 64.5% are male and 35.5% are female. These values are slightly different from national percentages which report the Race/Ethnicity breakdown of those with disabilities as 1.3% American Indian/Native American; 1.8% Asian/Pacific Islander; 20.3% African American; 13.7% Hispanic; 62.9% Caucasian (U.S. Department of Education, 2001). Concerning the issue of proportionality, a chi-square was performed to test if the Race/Ethnicity demographics were different across the two groups—students identified as having a disability versus those who were not. A chi-square value of 22.306 with four degrees of freedom resulted

in a significant value less than 0.05 denoted significant differences between the two groups.

Accommodation types and rules vary across states, but in State A accommodations must be approved by a team and be a part of the student's daily instructional program. In this sample, on the science subject test 2,197 (14.4%) students tested with accommodations and 12,710 (83.2%) tested without. In the math subject test 2,227 (14.6%) students used accommodations and 12,694 (83.1%) did not. For the reading exam, 2,221 (14.5%) students tested with accommodations and 12,703 (83.2%) did not. The specific types of accommodation given to each student are listed in Appendix A on Table 12 along with user frequencies. The accommodations range across four different areas of testing—timing, setting, how the test is presented, and how the students may respond. The most frequently used accommodations were extended time (same day), different testing site other than the student's regular classroom, small group administration, and verification that the student understood the directions. It should also be noted that specific accommodation types were not mutually exclusive, as some students received more than one. For the purposes of this study, all accommodations were grouped together. There were other special circumstances on the participation status of some students, but for the purposes of this study these students were excluded because the interest is the effect of an examinee having or not having access to an accommodation. This exclusion decreased the sample size for each subject test to N=14,921, N=14,924, and N=14,907 for math, reading, and science, respectively.

Each subject's exam included multiple choice (MC) and constructed response (CR) items and the math exam also included short answer (SA) items. The items used in this study were only items released by State A. Because of the chosen analyses used in this study, only the multiple choice items and responses from each exam were used. These items were dichotomously scored as 0 or 1, for incorrect and correct responses respectively. Correct responses were consequently summed for each examinee to create a revised total score which reflected only items used.

On the complete exam which used all item types, each item had a maximum score value which resulted in total raw score that determined the examinee's performance level. Multiple choice items received a maximum of one point for a correct answer. Constructed response items were scored 0, 1, 2, 3, or 4, with 4 being the maximum point value. Short answer items appeared only on the math exam and 2 points represented the maximum value. With this scoring methodology, the maximum number of points was 56 for each subject exam. The performance levels ranged from the Level 1 to Level 4, lowest to highest. Table 1 shows the distribution of students across the performance levels and is based on scaled scores that took sets of items into consideration, not all of which are included in computing the raw scores averaged in Table 1. It is clear that the group who tested with accommodations have higher percentages of students in Levels 1 and 2 than those students who tested without accommodations. Table 2 shows the average scores of the students across all three subjects and includes all item types previously mentioned with a maximum score value of 56. There is a notable score differential between those who received accommodations and those who did not.

Average scores for both students with and without accommodations proved to be highest on the reading subject test and the lowest on the math.

Table 1. Distribution of Students Across Performance Levels

	Level 1	Level 2	Level 3	Level 4
	Students tested <u>with</u> accommodations			
Math	62.5%	21.6%	14.5%	1.3%
Reading	42.4%	31.1%	23.6%	2.9%
Science	37.9%	30.1%	28.6%	3.3%
	Students tested <u>without</u> accommodations			
Math	17.2%	25.6%	44.4%	12.8%
Reading	5.3%	15.9%	52.4%	26.5%
Science	7.1%	19.8%	55.1%	18.1%

Table 2. Mean Scores on Subject Tests

	Students tested w/ accommodations	Students tested w/o accommodations
Math	17.78 (8.86)	29.64 (10.89)
Reading	26.23 (9.37)	38.73 (8.19)
Science	23.35 (9.19)	33.07 (8.76)

Note: Standard deviation in ( ); Includes all item types; Maximum score = 56.

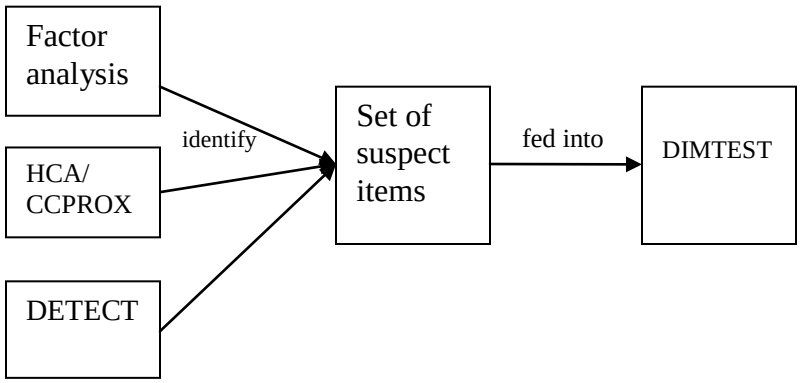
The mean differences displayed in Table 2 were statistically significant (p-value = .000; power =1.0) for the math, reading, and science exams, with small effect sizes of .137, .220, and .132, respectively (Cohen, 1988).

Procedures

In order to answer the research questions of interests, specific analyses were employed. Initially, to test the homogeneity of the covariance matrices of the dependent variables (test items) across groups (accommodated versus non-accommodated), a straightforward Box's M statistic tested at p = 0.001 was used because of the test's sensitivity to normality assumptions. This method tests the null hypothesis that the observed covariance matrices of the dependent variables are equal across groups. If a p-value less than 0.001 is found, then the covariance matrices are significantly different and the null is rejected, which is to be expected in this study. The Box's M test was done for each subject exam—math, reading, and science.

To answer the remaining research questions, more complex analyses were necessary. In determining if there is differential dimensionality for those who have disabilities and use accommodations versus those who do not, select dimensional analyses will be used. The logical order of some of the procedures can be tracked in the flow chart provided in Figure 1.

Figure 1. Flow Chart of DIMPACK Procedures



If the test is not unidimensional, the first task is to attempt to locate a dimensionally distinct subset of items from the original test in question. To increase power, it is suggested that this previously identified of subset of items be homogenous (Stout et al., 1996). This task can be accomplished through the use of techniques such as a substantive expert content review, HCA/CCPROX (Roussos, 1995; Roussos, Stout, & Marden, 1998) or factor analysis and are often used in combination with other MIRT programs such as DIMTEST (Zhang & Stout, 1999a) or DETECT (Kim, 1996; Zhang & Stout, 1999b). Hierarchical item cluster analysis, HCA, was used to produce dendrograms to cross validate the dimensionality structure. This analysis, done in SPSS, was the agglomerative method which progressively merges all individual clusters until one cluster remains. The options chosen for this analysis were the average linkage method, which was selected because of its ability to deal with small variances, along with squared Euclidean distance as the distance measure.

Additionally, an exploratory factor analysis (EFA) with scree plots was also used in selecting the appropriate assessment subtest for DIMTEST since the main purpose of an EFA, as defined by Harman (1976), is to resolve a set of variables into a smaller

number of “factors” or categories. Through the use of SPSS Statistics 18 a principal components analysis was done to identify a group of items that cluster together. Using dichotomous item data for factor analytic procedures has been controversial in research because results may be based solely on item distribution similarity. Research studies such as Shapiro, Lasarev, and McCauley (2002) observed that the use of dichotomous data in factor analysis procedures could be problematic because continuous multivariate normal data is expected and often suggest the use of tetrachoric correlations. Since item-level data rarely meets these criteria, this study followed the suggestion of Kim and Mueller (1978) which deemed the use of dichotomous data permissible in EFA if correlations between the variables were less than 0.7.

For the math exam only, the invariance of factor structure was formally tested across groups through the use of a multi-group structural equation model also using the LISREL program (Joreskog & Sorbom, 1993b). To take into account the use of dichotomous data in structural equation modeling, Joreskog and Sorbom (1993a) suggested obtaining the necessary covariance and asymptotic covariance matrices through the program PRELIS to obtain appropriate tetrachoric correlation values from the raw item response data since product- moment correlations are not recommended. Use of the asymptotic covariance matrix in the input file prompts LISREL to use the Weighted Least Squares (WLS) estimation method, which is appropriate for dichotomous or ordinal data. In determining if the factor structure holds across the accommodated and non-accommodated subjects, the hypotheses being tested are:

H_0: The factor structure is the same for each group

or

H_1: The factor structure is different for each group

Loosening certain constraints in the measurement models allowed different factor scenarios to be tested and then compared using the fit statistics produced by LISREL.

The DETECT program was used as a specialized estimation procedure that measures the amount of multidimensionality. It uses conditional covariances to partition items into widely spread dimensionally homogeneous clusters. This program also requires a data file that contains examinee responses. The reported DETECT index is estimated by a complex algorithm that takes into account the number of latent abilities required to make the inter-item conditional covariances approach zero. The DETECT index is somewhat subjective and ranges from 0 to values well over 1. Values near 0.1 indicated unidimensionality and greater than 1 indicate sizeable multidimensionality. In this study the index will be evaluated by the following criteria per Stout et al., 1996.

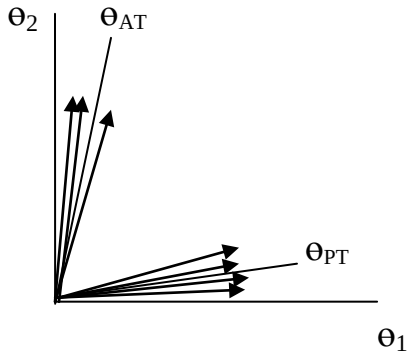
- Values around 0.1-0.5 indicate unidimensionality
- Values between 0.6-0.8 indicate moderate multidimensionality
- Values near and above 0.9 indicate sizeable multidimensionality

DETECT output gives the maximum DETECT value which will be compared across the subgroups and different subject assessments.

After assessing the structure, confidently identifying a subset of suspect items using a cluster analytic approach is an important step before continuing to the DIMTEST which makes use of two distinct subsets of items—AT1 (assessment subtest) and PT

(partitioning subtest). Froelich and Habing (2008) noted that it is favorable if the item vectors of AT1 are widely spread from the item vectors of the PT subtest as shown in Figure 2. The recommended minimum for the AT1 subset is three items and the minimum for PT is fifteen (Stout et al., 1996). In this study, a group of suspect items was to be identified for all three subject tests.

Figure 2. Example of a Good AT1 Selection for the DIMTEST Program



After identifying a distinct set of suspect items, the next step is to statistically test if these items are dimensionally distinct from the remaining items on the test through the DIMTEST procedure, which tests the hypothesis:

H_0: AT1 ∪ PT satisfies $d = 1$, where d is the number of test dimensions

or

H_0: AT1 ∪ PT is one-dimensional and there is no dimensional difference

The DIMTEST procedure requires the input of an item level data file with examinee responses. The specific option in DIMTEST used in this study is to hand enter pre-selected items for AT1. Once items are entered, AT1 is tested against PT. The result is

the DIMTEST statistic, T, which is normally distributed and can be interpreted like a z-statistic. A p-value less than 0.05 indicates that the two item clusters are significantly different in dimensionality—or dimensionally distinct. The T-statistics and p-values will be compared across subject tests for those students who tested with accommodations and those who did not in order to get a comprehensive view of the differences in dimensionality.

Following the dimensionality analyses, differential item functioning (DIF) was used to assess the issue of item fairness as a function of receiving an accommodation or not. The DIF-detection was performed through the use of the non-parametric Simultaneous Item Bias Test, SIBTEST (Stout & Roussos, 1995). This approach was chosen because the method uses a multidimensional perspective to calculate the size of DIF. The statistical test that assesses this degree of DIF in SIBTEST is $\hat{B}_{uni}$. The statistical null hypothesis for SIBTEST is:

$$H_0: B(T) = P_R(T) - P_F(T) = 0$$

where B(T) is the difference in the probability of a correct response on an item for the Reference and Focal groups; $P_R(T)$ is the probability of a correct response on an item for the Reference group; $P_F(T)$ is the probability a correct response on an item for the Focal group; and T is the match criteria, true score. Simply stated, the hypothesis posits that there is no difference in the probability of correctly answering an item for the Reference and Focal groups after controlling for ability, hence why the difference B(T) = 0. The alternative hypothesis is:

$$H_1: B(T) = P_R(T) - P_F(T) \neq 0.$$

Similar to DIMTEST, SIBTEST also uses two subsets of items—a suspect set of items which are the items believed to be biased and a valid set of matching items.

To run the program, SIBTEST first requires the use of SIBIN in order to create a control file to serve as the input file for SIBTEST. Several options must be chosen in order for SIBTEST to be run properly for a specific study. The specific options that this study employed were to analyze all items, use pooled weighting and Bonferroni p-values for all runs, and to test bias against either group. SIBTEST also needs separate item response files for the Reference and Focal groups. For this study, the split between the groups was based on testing with or without an accommodation. The Reference group included the non-accommodated students and the Focal group consisted of the accommodated students. The procedure was run three times, once for each of the subject tests to flag items that are potentially biased according to this split of the sample. SIBTEST outputs both a SIB-uni z-statistic and a Mantel-Haenszel chi-square, both with p-values and indication of whether the direction of the DIF is against the Reference or Focal group. Along with the items that are flagged as exhibiting DIF, these statistics were examined across the different subject tests in the results section. Also, path coefficients from the best fitting SEM model were compared with the DIF flagged items from the math exam to cross-validate the results.

CHAPTER IV

RESULTS

The purpose of this study was to psychometrically evaluate a set of standardized subject tests taken by eighth grade students—some who used accommodations and some who did not. This research revolved around the issue of test fairness because of the fact that some students with disabilities were appropriately accommodated whereas traditional students were not. It is important to determine if the tests fairly served both groups and to do this covariance structure, measurement invariance, dimensionality, and item functioning were examined. This chapter will provide an overview of the classical item statistics and reveal the statistical results for the psychometric analyses.

Item Analyses

Classical item and test statistics were obtained for each test and group (non-accommodated and accommodated) through the use of the program ITEMAL 6 (Ackerman, 2010) using only the dichotomously-scored multiple choice items. Table 3 is a summary table including averages and the complete item statistics are reported in Tables 13 through 18 in Appendix A. Looking at the overall test distributions graphically for each subject exam highlights the differences in then scores of the two groups (see Figures 3, 4 and 5 in Appendix A).

Table 3. Average Item and Test Statistics

	Avg. Item Diff.	Avg. Item Discrimination	Avg. Point Biserial	Avg. Test Score	Reliability	Skewness
Math						
Accommodated	0.363	0.322	0.329	11.6(4.9)	.738	1.057
Non-accommodated	0.551	0.469	0.408	15.6(6.1)	.841	0.166
Reading						
Accommodated	0.525	0.428	0.407	16.8(6.4)	.837	0.207
Non-accommodated	0.768	0.358	0.389	24.6(5.1)	.815	-0.931
Science						
Accommodated	0.511	0.443	0.379	15.9(5.7)	.803	0.221
Non-accommodated	0.691	0.389	0.377	21.4(5.1)	.796	-0.493

Notes: Item difficultly = % correctly answering item; Standard deviation in ( )

On the math exam, the accommodated group averaged a score of 11.6 compared to a 17.6 by the non-accommodated group. The distribution of the accommodated group was positively skewed. The average item difficulty was 0.363 for the accommodated group as opposed to 0.551 for the non-accommodated group. This item difficulty pattern held across each subject exam. The non-accommodated group on average consistently had a higher proportion of examinees who correctly responded to the test items. This resulted in higher mean test scores for the non-accommodated group across each test as well. The mean differences between the two groups were tested for each subject exam and results showed p-values of .000, indicating a statistically significant mean difference for accommodated and non-accommodated students. Power was equal to 1.0 for each

analysis and small effect sizes of .113, .211, and .125 were reported for the math, reading, and science exams respectively.

The reliabilities ranged from 0.738 to 0.841, with the lowest and highest both occurring on the math test. Discrimination values were also reported for each item and were calculated using the following formula:

$$d = p_u - p_l$$

where p_u is the proportion of the upper group that answered the item correctly and p_l is the proportion of the lower group that correctly answered the item (Crocker & Algina, 1986). The average discrimination, ranged from 0.322 to 0.469. As noted in Classical Test Theory (CTT), the reported point biserials are also a measure of discrimination. In order to properly obtain averages of the point biserial correlations, the Fisher's z' transformation was used to convert the correlations to a normally distributed z', and numbers were then converted back to correlations. Note that conversion calculators for this transformation exist, but the formula for transformation to z' is:

$$z' = .5[\ln(1+r) - \ln(1-r)]$$

The average point biserials for the items ranged from 0.329 to 0.408. The non-accommodated groups had negatively skewed distributions on both the reading and science exams (see Figures 4 & 5).

Analysis 1: Box's M

First, it was important to have a hypothesis test that tested the covariance structure of the two separate groups. To address this, the Box's M statistic was used. The null

hypothesis for this statistic tests the equality of the covariance matrices for the accommodated versus non-accommodated students. The procedure was repeated for the math, reading, and science exam.

The results in Table 4 show that for each of the three subject exams, there was a statistically significant difference in the covariance matrices across the groups. The p-value was tested at the 0.001 level.

Table 4. Box's M Results

	Box's M	**Significance**
Math	3355.48	.000*
Reading	8740.39	.000*
Science	5532.38	.000*

Note: * denotes a significant p-value at the p = .001 level

The statistical significance across all three tests showed that accommodated students and non-accommodated students did not have homogenous covariance matrices. However, this test is quite often statistically significant due to its sensitivity to the number of covariances and large sample sizes.

Analysis 2: Dimensionality

The Box's M statistic tested whether the dimensionality and the structure were the same. Although the results provided evidence that there may be some inherent

differences between the groups, dimensionality was then further investigated alone through a set of dimensionality techniques. Dimensionality was examined to be compared across groups to see if there was differential dimensionality.

First, an exploratory factor analysis (EFA) was run on the correlation matrices of the dichotomous data, and as mentioned earlier this method may be better employed through the use of tetrachoric correlations. An EFA was done separately for each group (accommodated and non-accommodated) to determine if the items on each subject exam could be condensed into smaller groups. This dimension reduction technique indicated that the correlational structure suggested that the subject exams were written as unidimensional and this was the case for both accommodated and non-accommodated students. Although the exploratory factor analysis extracted several components, factor loadings and scree plots (see Figures 6-8) confirmed the unidimensional nature of the tests. The scree plots were evaluated by determining the number of components above the break from linearity, which seemingly revealed one. For math, there was a slight difference between the scree plot for the accommodated and non-accommodated group.

In the resulting component matrices, the factor loadings showed that the correlation between the observed items and the factors were highest on one main component even though few items did have a stronger correlation with other extracted components. The amount of explained variance was also reported (see Table 5). In the EFA done with the accommodated students on the math subject test, the main component explained 11.7% of the 42.9% explained by all ten extracted components. For the non-accommodated group, over half of the variance was explained by the use of one

component. For the reading test with the accommodated group, the dominant component extracted explained nearly half (17.3%) of the variance given that the seven components explained a total of 37.7% of the variance. For the EFA performed with the non-accommodated student data, the main component explained well over half of the total cumulative variance explained by the four components extracted. In the science accommodated group, the EFA extracted eight components which explained a total variance of nearly 40%, but the major component alone explained 15.2%. For the non-accommodated group, five components explained 28.7% of the total variance and the major component explained half of that variance at 14.7%.

Table 5. Principal Components Analysis Results

Test & Group	**# of Extracted Components**	**# of items loading on major component**	**Explained Variance***
Math			
Accommodated	10	21 of 32	11.7%
Non-accommodated	5	29 of 32	17.2%
Reading			
Accommodated	7	22 of 32	17.3%
Non-accommodated	4	30 of 32	15.9%
Science			
Accommodated	8	21 of 31	15.2%
Non-accommodated	5	23 of 31	14.7%

Note: * Explained variance of major component

To further analyze the clustering of the items on each of the subject tests, dendrograms (see Figures 9 to 11 in Appendix A) and HCA/CCPROX were used. The dendrograms varied in appearance for each groups on all three exams suggesting different

clustering patterns of the items, so another clustering technique was used for more validation because it is expected that dendrograms produced from unidimensional data have long stems of equal length. The order of the items listed on the dendrograms reveal which items are most similar. The closer items are the more similar they are and the further apart the items are listed the less similar they are. For each subject exam, the item location on the vertical axis of the dendrograms varied for accommodated versus non-accommodated students. As determined by examining their respective p-values, the items seemingly clustered according the difficulty level, which could also indicate unidimensionality.

The output from HCA/CCPROX showed that starting with the one-cluster solution and working backwards, items only began forming one to two item clusters, varying across groups and subject tests. Looking at the various cluster solutions, it was clear that there was no distinct clustering of items that looked to form a complete second factor or dimension. Consequently, the use of DIMTEST became moot. The purpose of DIMTEST was to statistically test a group of suspect items, but due to the unidimensional nature of the data, this group of items could not be precisely identified and the DIMTEST procedure is most powerful when there is a wide clear-cut partition between the groups of tested items.

Running the DETECT program to measure the amount of multidimensionality further confirmed that each subject exam was unidimensional. Table 6 summarizes the results.

Table 6. DETECT Results

Test & Group	# of Dimensions that maximized DETECT	DETECT value	Ratio r
Math			
Accommodated	5	0.271	0.588
Non-accommodated	5	0.236	0.699
Reading			
Accommodated	5	0.260	0.540
Non-accommodated	5	0.158	0.630
Science			
Accommodated	5	0.241	0.536
Non-accommodated	4	0.153	0.574

Performing an exploratory DETECT analysis for each test resulted in a number of dimensions that maximized the DETECT value. The closer a DETECT value is to the zero minimum, the more likely the data is to be unidimensional and values near and above 1 indicate multidimensionality. Even though four to five dimensions were found, the maximal DETECT value never went above 0.3, with the non-accommodated group on the math exam having the highest value of 0.271. The lowest value was 0.153 for the non-accommodated group on the science exam. DETECT values of this low magnitude do not implicate multidimensionality. Ratio r is a quantitative measure of the notion of simple structure. Values greater than 0.8 generally indicate simple structure, but in this study the maximum ratio r was only 0.699. Thus, again DIMTEST was not employed to evaluate the dimensionally distinctness of the DETECT clusters given their low DETECT values.

Analysis 3: Measurement Invariance

After using Box's M test to examine the covariance structure and after confirming the dimensionality of the tests, the next research question involved measurement invariance across the multiple groups. This analysis was only performed on the math exam since previous analyses seemed to provide the most evidence that this exam is most likely to show differential dimensionality.

First, to confirm the unidimensional structure of the math exam, a confirmatory factor analysis (CFA) in LISREL was used and showed that a one-factor model structure fit both the accommodated and non-accommodated data. Again, because the data were dichotomous, asymptotic covariance matrices and the WLS estimation method were used. RMSEA values of 0.021 and 0.020 showed that this structure was a good fit to both groups. To further investigate the measurement invariance across the two groups, four different factorial structures were modeled and tested in LISREL with the math data, using a multi-group approach. The Chi-Square, Root Mean Square Error of Approximation (RMSEA), and Comparative Fit Index (CFI) are summarized in Table 7 for comparison purposes. The table includes four models that range from completely restrictive (Model A) to a model that allows both factor loadings and error variances to be freely estimated (Model D). The factor correlation is not used as a parameter because there is only one factor. Model B allowed factor loadings to be estimated for each group independently, whereas Model C allowed the same for error variances.

Table 7. Fit Statistics Summary of Measurement Invariance Models

	Global Goodness of Fit Statistics		
Model	Chi-square	RMSEA	CFI
Model A: **Factor loadings & Error variances Invariant**	3430.37* (992)	0.023	0.99
Model B: **Error variances Invariant (Factor loadings free)**	3189.58* (960)	0.022	0.99
Model C: **Factor loadings Invariant (Error variances free)**	3361.32* (960)	0.023	0.99
Model D: **Factor loadings & Error variances Free**	3116.47* (928)	0.023	0.99

Note: degrees of freedom in ( )

Each of the four models showed good fit by the normal standards of RMSEA and CFI. An RMSEA less than 0.05 and a CFI greater than 0.9 indicates acceptable fit implicating no measurement invariance across the groups. However, here the chi-square differences are important because this measure takes the chi-square and degrees of freedom of the most restrictive model and subtracts the next restrictive model (see Table 8). After computing the differences and p-values, the objective was to see if a less restrictive model had better fit. Model D, the least restrictive model, seemed to be most appropriate.

Table 8. Chi-square Difference

	Chi-Square $\chi 2$	**Model difference**	**$\chi 2$ difference**
Model A: **Factor loadings & Error variances Invariant**	3430.37* (992)	Model A - Model B	240.79* (32)
Model B: **Error variances Invariant (Factor loadings free)**	3189.58* (960)	Model A - Model C	69.05* (32)
Model C: **Factor loadings Invariant (Error variances free)**	3361.32* (960)	Model B - Model D	73.41* (32)
Model D: **Factor loadings & Error variances Free**	3116.47* (928)	Model C - Model D	244.85* (32)

Note: degrees of freedom in ( ); * denotes significant p-value

To further analyze Model D its item path coefficients for both groups were compared to see how different they were across groups. Table 19 in the appendix displays the path coefficients. Large differences in these coefficients could possibly hint at which items on the math exam are more likely to exhibit differential item functioning when using the accommodation split.

Analysis 4: Differential Item Functioning

When looking at the differential item functioning across accommodated and non-accommodated students, the program SIBTEST was used to see if there were any indications of differential performance due to the use of accommodations. SIBTEST outputs the statistic of interest, SIB-uni, as well as the Mantel-Haenszel (MH) statistic which is also a nonparametric approach for determining the degree of DIF. Beta-uni, the average p-value difference between the Reference and Focal group is also reported. In testing the significance of these statistics, a p-value = 0.01 was used. Tables 20, 21, and 22 show the complete results of the SIBTEST runs for each of the three content areas.

Table 9 shows the summary of items flagged with significant p-values ($p<0.01$) by both the SIBTEST SIB-uni and MH statistic. The table also explains the number of items that favor the Reference and Focal groups, which was determined by the sign of the SIB-uni statistic. A positive value indicates the Reference group is being favored while a negative value indicates the Focal group is being favored.

Table 9. DIF Results Summary

	# of flagged items		**# of items favoring**	
	Sib-uni	MH	Reference group	Focal group
Math	11	13	4	7
Reading	10	12	6	4
Science	15	16	9	6

Note: Reference group = non-accommodated students
Focal group = accommodated students

Overall, the total number of items flagged by the DIF detection procedure varied for each subject exam and ranged from 10-15 when referring to the SIB-uni p-values. The number of flagged items slightly increased when using the MH statistic. On the math exam, 11 significant items were identified with seven of those items favoring the focal group. For the reading and science exams, most of the items flagged favored the Reference group. On the reading exam, 10 items were flagged with only four favoring the Focal group and six favoring the Reference group. On the science exam, a total of 15 items were flagged and of that total nine favored the Reference group.

Using published item information for the math and reading exams, flagged DIF items were linked back to their original content specifications. These results are summarized in Tables 10 and 11 to show the DIF items with their respective content and the group favored.

Table 10. Math Items Exhibiting DIF with Matching Content Specifications

Item No.	Group Favored	Content Specification
1	Focal	Mathematical Decision Making: Probability
9	Focal	Patterns: Mathematical Communication
11	Focal	Patterns: Mathematical Communication
13	Reference	Shape and Size: Measurement
14	Reference	Mathematical Decision Making: Probability
16	Reference	Shape and Size: Measurement
20	Reference	Shape and Size: Geometry
21	Focal	Mathematical Decision Making: Data Analysis & Statistics
22	Focal	Mathematical Decision Making: Data Analysis & Statistics
23	Focal	Mathematical Decision Making: Probability
28	Focal	Numbers and Operations: Numbers & Number Sense

Note: Reference group = non-accommodated student; Focal group = accommodated students

On the math exam of the 7 items favoring the Focal group, 4 were labeled as Mathematical Decision Making problems. Three of the 4 items favoring the Reference group were in the Shape and Size content area.

Table 11. Science Items Exhibiting DIF with Matching Content Specifications

Item No.	Group Favored	Content Specification
7	Focal	Nature & Implications of Science: Communication
11	Reference	Earth & Space Sciences: The Universe
13	Reference	Earth & Space Sciences: Continuity & Change
14	Reference	Earth & Space Sciences: Continuity & Change
15	Reference	Nature & Implications of Science: Communication
16	Focal	Nature & Implications of Science: Inquiry & Problem Solving
17	Focal	Physical Sciences: Motion
18	Focal	Physical Sciences: Structure of Matter
19	Focal	Life Sciences: Classifying Life Forms
22	Reference	Physical Sciences: Energy
23	Reference	Life Sciences: Ecology
26	Focal	Nature & Implications of Science: Inquiry & Problem Solving
27	Reference	Nature & Implications of Science: Implications of Science & Technology
30	Reference	Earth & Space Sciences: The Universe
31	Reference	Earth & Space Sciences: Continuity & Change

Note: Reference group = non-accommodated students; Focal group = accommodated students

On the science exam, of the 9 items favoring the Reference group, 5 were in the Earth and Space Sciences content area. There were 6 items that favored the Focal group and half of those items were in the Nature and Implications of Science content area.

CHAPTER V

CONCLUSIONS

In the past testing has been the subject of a wide body of research. However, it is also important to research impact of the testing identifiable subpopulations. In particular, this study looked at a state's eighth-grade assessment program to see if and how those exams behaved differently across students with disabilities who used testing accommodations and those students who did not. Comparing populations across tests can shed light on a number of test issues that affect fairness and validity but in this comparison study there was no psychometric evidence to conclude that the use of accommodations provided an advantage. This chapter will discuss perceived study limitations and directions for future research, but will first summarize the findings of the previously posed research questions:

1. Is there a statistically significant difference in the covariance structure for students who were accommodated and for those who were not?
2. On a given assessment, is there differential dimensionality across the groups?
3. Does the same factor structure hold across the two groups?
4. Is there a difference in item functioning for those students with disabilities who received accommodations versus those without?

First, looking at classical item and test statistics revealed noticeable differences on each subject exam for accommodated and non-accommodated students. Average discrimination values were higher for accommodated student except for on the math exam. Total test scores were significantly different across the two groups on all three subject exams. P-values for non-accommodated students tended to be higher which explains the higher mean scores for this group, because the sum of the p-values is equal to the mean.

The second step in this study was to compare accommodated to non-accommodated students was to examine the covariance structure of the data using a Box's M statistic which tests the null hypothesis that there is no difference between the covariance matrices of the two groups being tested—i.e., homogeneity of covariance matrices. With the distinct differences between the groups being highlighted in the classical item and test statistics, it was to be expected that Box's M would detect a significant difference between the covariance matrices. The response data differed enough for each of the groups so that Box's M yielded a statistically significant difference. After determining that the groups did not have equal covariance matrices, it was natural to then test whether the dimensionality and the structure were the same across groups.

Examining the differential dimensionality across groups included several analyses and procedures in order to be thorough. First an exploratory factor analysis was performed and to extract the principal components for each of the subject tests. The EFA was done separately for accommodated and non-accommodated students for each test to

examine the presence of differential dimensionality. For both groups on the reading test, factor loadings and scree plot evidence favored unidimensionality, but scree plots for the math and science exams could have been interpreted in a way that suggested a second dimension. On the scree plots for science test (see Figure 8) the break from linearity for both the accommodated and non-accommodated groups suggested a possible second dimension. On the math test, the scree plot for the non-accommodated group suggested only one dominant dimension while the scree plot for the accommodated group looked slightly different and suggested a possible second dimension (see Figure 6). This difference on the math exam was a hint of possible differential dimensionality. The dendrograms output in the hierarchical cluster analysis from SPSS also looked different for the two groups on each of the three different exams, but strikingly different on the math exam. Because of the varied results, the use of other dimensionality and clustering techniques were used for cross-validation.

Using software from MIRT that was created to assess dimensionality allowed for further exploration of differential dimensionality and detection of multidimensionality. HCA/CCPROX was used in addition to the dendrograms as another item clustering technique. For each group on the three different tests, the two-cluster solution did not show that a distinct group of items clustered together and neither did the three-cluster or four-cluster solutions. Largely, individual items formed single item clusters. Given that the EFA, hierarchical cluster analysis, and HCA/CCPROX did not confidently lead to a concluding set of items that formed a separate distinct dimension, the use of DIMTEST became illogical. Consequently, the DETECT program was used in an exploratory

manner because it computes a statistic that measures the amount of multidimensionality and the tests were seemingly unidimensional. The DETECT program found the number of dimensions that maximized the DETECT statistic for accommodated and non-accommodated students on each test. For each run the DETECT value never exceeded 0.3, and values near and above 1.0 indicate multidimensionality. Therefore, it was safe to conclude that there was no evidence of multidimensionality or differential dimensionality. If there was in fact a second dimension or factor, it is likely that the factors were so highly correlated that the tests functioned in a unidimensional fashion. If two factors are highly correlated, it is difficult to argue that they are truly different factors measuring a different set of skills.

The application of SEM was used next on the math exam to test the factor structure across the groups to target the concept of measurement invariance thoroughly discussed by Horn and McArdle (1992). Initially a one-factor CFA was used to confirm unidimensionality for both groups. Next a series of SEM multi-group models that loosened constraints on the factor loading and error variances were performed and indicated no measurement invariance across the groups. This analysis also revealed that the most restrictive model which kept factor loadings and error variance invariant did fit the data. However, the less restrictive models also had acceptable fit indices.

Furthermore, to compare the models a chi-square difference test was used to evaluate the change in fit as model constraints were relaxed. The most constrained model was compared to the next most constrained model and assessed to determine if relaxing

the constraints improved model fit significantly. Because a more relaxed model is better to preserve parsimony, Model D seemed to be the most appropriate model choice overall.

The path coefficients from the SEM Model D were further examined to compare the coefficients across groups in order to see if items with the largest differences across groups would be the items to exhibit DIF. However, it turned out that the items flagged through SIBTEST did not have the largest differences between the two sets of path coefficients (see Table 19). To further examine the relationship between SEM and DIF, future studies could opt to examine various SEM multisample models to see how path coefficients change across groups and how that relates to differential item functioning.

After examining the covariance structure, dimensionality, and factor structure across groups, the last psychometric analysis that was performed to compare the two different testing populations was differential item functioning (DIF). This analysis was done to compare the item functioning across groups. SIBTEST was used as the DIF detection procedure which calls for the split of examinees into a Reference and a Focal group, and in this case the group split was based on receipt of a testing accommodation. The split resulted in the non-accommodated students being the Reference group and the accommodated students being the Focal group. After running three separate DIF analyses for each subject exam, it became evident that some items did function differently across groups. Like previous DIF studies that looked at students with disabilities, there was no clear consensus on if one group benefited more (Abedi, Leon, & Kao, 2008; Finch, Barton, & Meyer, 2009; Willingham et al., 1988). There were 11, 10, and 15 items flagged on the math, reading, and science exams, respectively. For the

reading and science exams, the majority of the DIF items favored the Reference group, while on the math exam most of the items favored the Focal group. There was a major overlap between the items flagged by the SIBTEST method and those items flagged using the Mantel-Haenszel (MH) method, however MH consistently flagged more items.

Although approximately one-third up to nearly one-half of the items were flagged as exhibiting DIF, without substantive analysis of the items, it is not possible to conclusively determine whether the DIF is due to bias or impact. As defined by Guhn, Gaderman, and Zumbo, 2007, impact is due to true ability differences, which is what a test intends to measure, whereas bias is due to irrelevant differences.

Having used dichotomous test data proved to be somewhat of a limitation although this problem was manageable. Not examining actual test items was a severe limitation to this study, but having content specifications for each item on the math and science exams allowed this DIF analysis to be further explored. When doing a DIF analysis access to the content of the items is very important in order to draw conclusions about flagged items because only then can the items be compared and grouped with respects to their common features such as content, cognitive processes, or skills used to correctly answer the items. On the math exam, the majority of the items exhibiting DIF favoring accommodated student were in the Mathematical Decision Making content area, while items in the Shape and Size content area tended to favor the non-accommodated students. On the science exam, the majority of items that favored the non-accommodated students were Earth and Space Sciences problems. Items favoring the accommodated students were in the Nature and Implications of Science content area. This information

can be used to extend the research into why these types of items would show tendency to favor one group more often than another. To further analyze the source of the DIF, other information such as sentence structure or distractor length could be examined when actually looking at the test items.

Accommodations have to be used and distributed fairly. Does simply prescribing an accommodation equal instant fairness? The answer appears to be no because recipients must be appropriately identified and the format and delivery must be standardized so use of accommodations do not falsely inflate test scores (Fuchs, Fuchs, & Capizzi, 2005; Ofiesh & Bisagno, 2008). However, Camara (2009) emphasized a very important point that was significant in driving this research study—that accommodations should not ever change the intended construct being measured, but they should only make information more accessible to those students at a disadvantage due to a learning disability. In other words, accommodations are an attempt at equality and should only be employed to level the playing field.

In this research study, although unidimensionality and factor structure seemed to hold across the groups of accommodated and non-accommodated students, there was a marked difference in test score distributions. Accommodated students scored significantly lower than non-accommodated students on each of the subject exams. This naturally raises the question of the utility of accommodations if their intent is indeed to level the playing field. However, it would be premature to attribute the large difference in test scores to ineffective accommodations or to even say that accommodations have

not served their purpose because there just may be a true difference in latent ability for the two groups.

Knowing that accommodations should not alter test constructs (Camara, 2009), this study examined dimensionality, DIF, and structural invariance across groups. Testing for differential dimensionality directly related to Camara's notion and sought out to determine whether or not accommodated students were being tested on different constructs than non-accommodated students. It is important to remember that constructs are a function of student-item interaction and perception of what is being measured. For this reason, tests makers must clearly articulate valid constructs. This study found that across the math, reading and science tests, there was strong evidence to prove that the tests measured one dimension for both accommodated and non-accommodated students. In other words, the one-factor structure held for both groups. This structural or factor invariance is important to scientific validity (Horn & McArdle, 1992). The only difference that should matter on a test is difference in ability, not differences in race, gender, SES, or exceptionality status. Equal dimensionality and factor invariance speaks to a well designed test.

It may initially seem counterintuitive that DIF still manifests itself if dimensionality and factor structure are equal. However, items were flagged as statistically exhibiting DIF in this study but it cannot immediately be concluded that those particular items were biased even though SIBTEST identified the direction of the DIF. Bias implies that an irrelevant construct was being measured, which cannot be convincingly stated without actual test items. Accommodated and non-accommodated

students appeared to interact with the items differently, showing that DIF is a result of more than just a dimensionality difference. Due to the variety of DIF detection procedures available, future studies could opt to compare the results of those across groups. Some other DIF methods that could be used are DFIT or the CDM (cognitive diagnostic model) Mantel-Haenszel method in which examinees are matched on an attribute mastery profile as opposed to being matched on total test score.

An alternative to DIF that could extend the foundation of group comparison involves the use of a distractor analysis that would shed more light on how accommodated and non-accommodated students interacted differently with specific test items. The analysis could be done for all items or specifically for items exhibiting DIF. Kopriva, Cameron, Carr, and Taylor (2008) discussed the limits of DIF and how examining distractors and their distributions could possibly provide more insight on problematic items. Looking at distractors could also reveal valuable diagnostic information.

The literature review of this research study reviewed several different accommodation types, but treated all of them as one to prevent possible low sample size issues. Future research may look at group differences across the different types of accommodations such as oral administration or extended time. More nuanced differences may be found in when separating students by the specific type of accommodation received. It may also be interesting to see how impact of accommodations varies across grade level. This study included only eighth grade student data because it was felt that this was a pivotal year in educational growth and that dimensionality differences, if

present, might be greatest at this level (Gutman, 2006; Isakson & Jarvis, 1999; Neild, Stoner-Eby, & Furstenburg, F, 2008; Reyes, Gillock, Kobus, & Sanchez, 2000).

Although accommodations are only one way of creating a fair testing environment for students with disabilities, it is important not to disregard alternate assessments in future dimensionality research. The dimensional structure should be the same across groups of students who use accommodations and those who do not, but should also be the same for students who qualify to take alternate assessments and those who take traditional assessments. This scenario is slightly different, but compares similar populations in which dimensionality should certainly be invariant if an alternate assessment claims to measure the same constructs as a traditional assessment. The dimensionality of alternate assessments as compared to their standard counterparts would allow for further documentation of the psychometric properties of such type of alternate assessments, since it is currently minimal. Looking at the dimensionality across various alternate and standard assessment pairs could provide valuable insight into what developmental methods work best, which allows for advancement in the field. Showing that the two different tests measure the same constructs is a step closer to being able to link or equate the test scores.

Future studies should explore using additional SEM models that purposely constrain specific items or allow previously identified suspect items to be freely estimated across groups. More specifically, if a set of items proved to be dimensionally distinct, the path coefficients of those particular items could be set to free in attempts to

find a better fitting SEM model. Comparing this methodology across groups would be particularly interesting to see if any noticeable group differences emerged.

Exploring the psychometric characteristics of state exams deepens the literature of applied psychometrics, so any future analysis would be valued. Researching students with disabilities has become increasingly important since the inclusion of this population in national accountability standards created by legislation such as *No Child Left Behind*. These students have often been the subject of various research topics ranging from how they learn to how they are tested. This study subset the population to compare students who received and used testing accommodations to those who did not. Despite the methodology or subgroup, it is important that educational research on this population continues in order to improve the quality of their education.

REFERENCES

Abedi, J., Seth, L., & Kao, J. C. (2008). Examining differential item functioning in reading assessments for students with disabilities. CRESST Report 744. University of California, Los Angeles.

Ackerman, T. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement, 29*(1), 67-91.

Ackerman, T. (1994). Using Multidimensional Item Response Theory to Understand What Items and Tests Are Measuring. *Applied Measurement in Education, 7*(4), 255-278.

Ackerman, T. (2010). ITEMAL (Version 6). [Computer software]. Greensboro, NC: University of North Carolina at Greensboro.

Ackerman, T. A., Gierl, M. J., Walker, C. M. (2003). Using multidimensional item response theory to evaluate educational and psychological tests. NCME Instructional Module.

Bachor, D. (1990). Toward improving assessments of students with special needs: Expanding the database to include classroom performance. *The Alberta Journal of Educational Research*, 36, 65-77.

Barton, K. & Finch, W. H. (2004, April). *Using DIF analyses to examine bias and assumptions of unidimensionality across students with and without disabilities*

and students with accommodations. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Beratan, G. (2008). The song remains the same: Transportation and the disproportionate representation of minority students in special education. *Race, Ethnicity, and Education, 11*(4), 337-354.

Berlak, H. (2001). *Academic Achievement, Race, and Reform: Six Essays on Understanding Assessment Policy, Standardized Achievement Tests, and Anti-Racist Alternatives.* Retrieved November 28, 2009, from ttp://www.eric.ed.gov/ERICWebPortal/contentdelivery/servlet/ERICServlet?accno=ED464973.

Bintz, W., & Harste, J. (1994). Where are we going with alternative assessment and is it really worth our time? *Contemporary Education*, 66, 7-12.

Bolt, S. (2004, April). *Using DIF analysis to examine several commonly-held beliefs about testing accommodations for students with disabilities.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Bolt, S., & Bielinski, J. (2002, April). *The effects of the read aloud accommodation on math test items.* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Bolt, S. E. & Thurlow, M. L. (2006). Item-level effects of the read-aloud accommodation for students with reading disabilities. Synthesis Report 65. Minneapolis, MN: National Center on Educational Outcomes.

Burgstahler, S. (2009). Universal Design of Instruction (UDI): Definition, Principles, Guidelines, and Examples. *DO-IT, 1*, 1-4.

Byrnes, M. (2008). Removing barriers to learning: A primer. *Principal Leadership, 9*(3), 34-37.

Cahalan-Laitusis, C., Cook, L. L., Cope, J. M., & Rizavi, S. (2004, April). *Detecting item bias for students with disabilities.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Camara, W. (2009). Validity evidence in accommodations for english language learners and students with disabilities. *Journal of Applied Testing Technology, 10*(2) 1-23.

Cohen, J. (1988). *Statistical power for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

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

Darling-Hammond, L. (1994). Setting standards for students: The case for authentic assessment. *The Educational Forum*, 59, 14-21.

Douglas, J., Kim, H., Roussos, L., Stout, W., Zhang, J. (1999). LSAT dimensionality analysis for the december 1991, and october 1992 administrations. Statistical Report. LSAC Research Report Series. Newtown, PA: Law School Admissions Council.

Elbaum, B. (2007). Effects of an oral testing accommodation on the mathematics performance of secondary students with and without learning disabilities. *Journal of Special Education, 40*(4), 218-229.

Elliot, S. N., & Fuchs, L. (1997). The utility of curriculum-based measurement and performance assessment as alternatives to traditional intelligence and achievement tests. *School Psychology Review*, 26, 224-233.

Elliot, S. N., Kratochwill, T. R., & McKevitt, B. C. (2001). Experimental Analysis of the Effects of Testing Accommodations on the Scores of Students with and without Disabilities. *Journal of Psychology, 39*(1), 3-24.

Elliot, S. N., Kratochwill, T. R., & McKevitt, B. C., Malecki, C. K. (2009). The effects and perceived consequences of testing accommodations on math and science performance assessments. *School Psychology Quarterly, 24*(4), 224-239.

Elliott, S. N. & Roach, A.T. (2007). Alternate Assessments of Students with Significant Disabilities: Alternative Approaches, Common Technical Challenges. *Applied Measurement in Education, 20*(3), 301-333.

Finch, H., Barton, K, & Meyer, P. (2009). Differential item functioning analysis for accommodated versus nonaccommodated students. *Educational Assessment, 14*(1), 38-56.

Fuchs, L. S. & Fuchs, D. (2001). Helping teacher formulate sound test accommodation decisions for students with learning disabilities. *Learning Disabilities Research and Practice, 16,* 174-181.

Fuchs, L. S., Fuchs, D., & Capizzi, A. M. (2005). Identifying Appropriate Test Accommodation for Students With Learning disabilities. *Focus on Exceptional Children, 37*(2), 1-8.

Fuchs, L. S., Fuchs, D., Eaton, S. B., Hamlet, C. L., Binkley, E., & Crouch, R. (2000). Using objective data sources to enhance teacher judgments about test accommodations. *Exceptional Children, 67,* 67-81.

Fuchs, L. S., Fuchs, D., Eaton, S. B., Hamlet, C. L., & Karns, K. (2000). Supplementing teacher judgments of mathematics test accommodations with objective data sources. *School Psychology Review, 29,* 65-85.

Froelich, A. G. & Habing, B. (2008). Conditional covariance-based subtest selection for DIMTEST. *Applied Psychological Measurement, 32,* 138-155

Geisinger, K. F. (1994). Psychometric issues in testing students with disabilities. *Applied Measurement in Education, 7*(2), 121-140.

Gibson, D., Haeberli, F. B., Glover, T. A., & Witter, E. A. (2005). Use of recommended and provided testing accommodations. *Assessment for Effective Intervention, 31,* 19-36.

Goh, D. (2004). *Assessment accommodations for diverse learners*. Boston, MA: Allyn and Bacon.

Guhn, M., Gadermann, A., Zumbo, B. D. (2007). Does the EDI Measure School Readiness in the Same Way Across Different Groups of Children? *Early Education and Development.* 18(3), 453-472.

Gutman, L. M. (2006). How student and parent goal orientations and classroom goal structures influence the math achievement of African Americans during the high school transition. Contemporary Educational psychology, 31(1), 44-63.

Harman, H. H. (1976). *Modern factor analysis* (3rd ed., revised). Chicago: University of Chicago Press.

Haugers, J. B., & Sireci, S. G. (2008). Detecting differential item functioning across examinees tested in their dominant language. *International Journal of Testing, 8*(3), 237-250.

Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance and aging research. *Experimental Aging Research, 18*(3), 118-144.

Hosp, J. L. & Reschly, D. L. (2004). Disproportionate representation of minority students in special education: Academic, demographic, and economic predictors. *Exceptional Children, 75*(2), 185-199.

Huynh, H. & Barton, K. E. (2006). Performance of students with disabilities under regular and oral administrations of a high-stakes reading examination. *Applied Measurement in Education, 19,* 21-39.

Isakson, K., & Jarvis, P. (1999), The adjustment of adolescents during the transition into high school: A short-term longitudinal study. Journal of Youth and Adolescence, 28(1), 1-26.

Johnson, E., & Arnold, N. (2004). Validating an alternate assessment. *Remedial and Special Education, 25*(5), 266-275.

Joreskog, K. & Sorbom, D. (1993a). *LISREL 8: Structural equation modeling with the SIMPLIS command language.* Hillsdale, NJ: Lawrence Erlbaum Associates Publishers.

Joreskog, K. & Sorbom, D. (1993b) LISREL 8.0. [Computer software]. Chicago: Scientific Software International.

Kettler, R. J., Elliott, S. N., & Beddow, P. A. (2009). Modifying achievement test items: A theory-guided and data-based approach for better measurement of what students know. *Peabody Journal of Education, 84*(4), 529-551.

Kim, H. R. (1996). A new index of dimensionality—DETECT. *Pure Applied Mathematics, 3*(2), 141-153.

Kim, J. & Mueller, C. W. (1978). Factor *analysis: Statistical methods and practical issues.* Beverly Hills, CA: Sage Publications.

King-Sears, M. (2009). Universal design for learning: Technology and pedagogy. *Learning Disability Quarterly,* 32(4), 199-201.

Kopriva, R., Cameron, C., Carr, T., & Taylor, M (2008, April). *The limits of DIF: Why this item evaluation tool is flawed for english learners, hearing impaired, and students with learning disabilities.* Presented at the annual meeting of the National Council of Measurement in Education, New York, NY.

Lam, T. C., (1993). Testability: A Critical Issue in Testing Language Minority Students with Standardized Achievement Tests. *Measurement and Evaluation in Counseling and Development, 26*(3), 179-191.

Lang, S., Elliott, S. N., Bolt, D. M., Kratochwill, T. R. (2008). The effects of testing accommodations on students' performances and reactions to testing. *School Psychology Quarterly, 23,* 107-124.

Lewandowski, L. J., Lovett, B. J., Rogers, C. L. (2008). Extended time as a testing accommodation for students with reading disabilities: Does a rising tide lift all ships? *Journal of Psychoeducational Assessment, 26*(4), 315-324.

Mahoney, K. (2008). Linguistic influences on differential item functioning for second language learners on the national assessment of educational progress. *International Journal of Testing, 8,* 14-33.

Martiniello, M. (2008). Language and the performance of english-language learners in math word problems. *Harvard Educational Review, 73*(2), 333-368.

Nandakumar, R. (1994). Assessing dimensionality of a set of item responses—comparison of different approaches. *Journal of Educational Measurement, 31*(1), 17-35.

Neild, R. C., Stoner-Eby, S., & Furstenberg, F. (2008). Connecting entrance and departure: The transition to ninth grade and high school dropout. Education and Urban Society, 40(5), 543-569.

Ofiesh, N. S. & Bisagno, J. M. (2008). Enduring and unresolved issues in the test accommodation decision process for individuals with learning disabilities: Test validity and documentation of learning disabilities. *Learning Disabilities: A Multidisciplinary Journal, 15*(3), 147-151.

Passman, T. & Green, R. A. (2009). Start with the syllabus: Universal design from the top. *Journal of Access Services, 6*(1), 48-58.

Phillips, S. E. (1994). High-stakes testing accommodations: Validity versus disabled rights. *Applied Measurement in Education, 7,* 93-120.

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

Reyes, O., Gillock, K. L., Kobus, K., & Sanchez, B. (2000). A longitudinal examination of the transition into senior high school for adolescents from urban, low-income status, and predominantly minority backgrounds. American Journal of Community Psychology, 28(4), 519-544.

Roussos, L. (1995). Hierarchical agglomerative clustering: HCA/CCPROX. [Computer software]. Urbana-Champaign: Statistical Laboratory for Educational and Psychological Measurement, Department of Statistics, University of Illinois.

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

Salend, S. J., Duhaney, L. M., & Montgomery, W. (2002). A comprehensive approach to identifying and addressing issues of disproportionate representation. *Remedial and Special Education 23*(5), 289-299.

Salvia, J., Ysseldyke, J, & Bolt, S. (2006). *Assessment: In special and inclusive education*. Florence, KY: Wadsworth Publishing.

Schafer, W. D., & Lissitz, R. W. (2009). *Alternate assessments based on alternate achievement standards: Policy, practice and potential.* Baltimore, MD: Brookes Publishing Company.

Shapiro S., Lasarev M., and McCauley L. (2002). Factor analysis and gulf war illness: What does it add to our understanding of possible health effects of deployment. *American Journal of Epidemiology, 156 (6),* 578-585.

Sinharay, S, Dorans, N. J., Liang, L. (2009). First language of examinees and its relationship to differential item functioning. Research Report. ETS. RR-09-11. Princeton, NJ: Educational Testing Service.

Sireci, S. G., (2009). No more excuses: New research on assessing students with disabilities. *Journal of Applied Testing Technology, 10*(2), 1-23.

Sireci, S. G., Scarpati, S. E. & Li, S. (2005). Test accommodations for students with disabilities: An analysis of the interaction hypothesis. *Review of Educational Research, 75*, 457-490.

Stout, W., Habing, B., Douglas, J., Kim, H. R., Roussos, L., Zhang, J. (1996). Conditional covariance-based nonparametric multidimensionality assessment. *Applied Psychological Measurement, 20*(4), 331-354.

Stout, W.F., & Roussos, L.A. (1995) SIBTEST. [Computer software]. Urbana-Champaign: University of Illinois, Department of Statistics.

Tindal, G., & Fuchs, L. S. (2000) *A summary of research on test changes: An empirical basis for defining accommodations.* Lexington, KY: Mid-South Regional Resource Center Interdisciplinary Human Development Institute.

Tindal, G., Heath, B., Hollenbeck, K., Almond, P., & Harniss, M. (1998). Accommodating students with disabilities on large-scale tests: An experimental study. *Exceptional Children, 64,* 439-450.

U.S. Department of Education (2001). *Table II-10. Twenty-third annual report to congress on the implantation of the individuals with disabilities act.* Retrieved March, 28, 2009, from https://www.theadvocacyinstitute.org.

U.S. Department of Education (2007). *Table 1-5. Students Ages 12 through 17 Served Under IDEA, Part B, by Disability Category and State: Fall 2006*. Retrieved November 28, 2009, from https://www.ideadata.org/arc_toc8.asp#partbCC.

Willingham, W. W., Ragosta, M., Bennett, R. E., Braun, H., Rock, D. A., & Powers, D. E. (1988). *Testing handicapped people.* Needham Heights, MA: Allyn & Bacon.

Willingham, W. W. (1986). *Testing handicapped people—the validity issue.* ETS Research Report. Princeton, NJ.

Wilson, M. (2004). Assessment, accountability and the classroom: A community of judgment. In M. Wilson (Ed.), *Towards coherence between classroom assessment and accountability* (pp.1-19) 103rd Yearbook of the National Society for the Study of Education, Chicago: University of Chicago Press.

Yovanoff, P. & Tindal, G. (2007). Scaling early reading alternate assessments with statewide measures. *Exceptional Children*, 73, 184-201.

Zhang, J. & Stout, W. (1999a). Conditional covariance structure of generalized compensatory multidimensional items. *Psychometrika,64*(2), 129-152.

Zhang, J. & Stout, W. (1999b). DETECT. [Computer software]. St. Paul, MN: Assessment Systems Incorporated.

APPENDIX A. TABLES AND FIGURES

Table 12: 2007-08 Approved Accommodations: Usage Frequencies

	Grade 8: Subject		
Timing Accommodations	**Reading**	**Math**	**Science**
Extended time (same day)	1,416	1,414	1,394
Extended time (several days)	249	253	251
Multiple or frequent breaks	477	472	460
Beneficial time of day/week	146	145	143
Flexibility in content area order	171	174	167
Setting Accommodations			
Site other than regular classroom	1,005	1,012	996
Out-of-school setting	27	26	25
Presentation Accommodations			
Individual administration	212	216	213
Small group administration	1,808	1,812	1,786
Using a human reader	526	837	827
Using sign language	6	8	8
Opportunity to move, stand, or pace	169	163	159
Alternative assistive technology	8	7	7
By school personnel other than teacher	797	803	798
Large print version	21	19	20
Braille version	2	2	2
Word-to-word bilingual dictionary	19	20	20
"Sheltered English"	102	105	103
Response Accommodations			
Using a scribe or recording device	300	263	288
Alternative assistive communication devices	13	14	12
Other assistive devices	189	251	200
Student use of word processor	244	204	223
Student use of brailler	2	3	2
Student use of visual aids	14	13	13
Word-to-word bilingual dictionary	11	12	11
Verification of comprehension of directions	1,234	1,243	1,227
Side-by-side placement of 2 test booklets	31	30	29
Rewrite of illegible student responses	26	22	27
Other Accommodations			
Other accommodations not documented	1	1	1

Table 13. ITEMAL Results: Non-Accommodated Math

Item	P-Value	Dis.	Biserial	Pt. Biserial	Reliability
1	.9045	.1855	.4625	.2662	.0782
2	.3919	.5181	.5615	.4412	.2154
3	.7710	.2706	.3725	.2670	.1122
4	.6827	.4280	.4924	.3761	.1751
5	.5774	.5455	.5624	.4452	.2199
6	.3257	.4893	.5686	.4355	.2041
7	.3828	.5364	.5800	.4551	.2212
8	.5253	.5343	.5558	.4430	.2212
9	.7990	.3080	.4508	.3154	.1264
10	.2697	.3197	.4160	.3086	.1369
11	.6588	.4605	.5044	.3902	.1850
12	.6561	.5404	.5952	.4596	.2183
13	.4658	.4434	.4713	.3754	.1873
14	.2463	.3053	.4278	.3122	.1345
15	.5587	.6893	.6948	.5520	.2741
16	.5935	.6058	.6202	.4894	.2404
17	.4387	.4767	.4996	.3965	.1968
18	.5836	.6480	.6635	.5241	.2584
19	.5036	.6810	.6819	.5441	.2720
20	.3106	.4109	.4894	.3723	.1723
21	.5165	.5882	.5962	.4753	.2375
22	.5358	.5075	.5197	.4140	.2065
23	.7914	.3602	.5296	.3715	.1510
24	.5923	.5335	.5610	.4425	.2175
25	.3795	.3392	.3796	.2974	.1443
26	.5253	.4748	.4889	.3896	.1946
27	.6966	.5609	.6516	.4940	.2271
28	.8306	.3226	.5306	.3559	.1335
29	.5833	.5161	.5512	.4353	.2146
30	.6081	.4455	.4881	.3836	.1872
31	.3848	.4411	.4953	.3882	.1889
32	.5332	.5335	.5516	.4393	.2191

Table 14. ITEMAL Results: Accommodated Math

Item	P-Value	Dis.	Biserial	Pt. Biserial	Reliability
1	.6659	.4951	.5353	.4128	.1947
2	.1944	.2453	.4891	.3377	.1336
3	.6731	.2206	.2632	.2023	.0949
4	.4621	.4039	.4586	.3651	.1820
5	.3085	.4142	.5504	.4174	.1928
6	.2030	.1078	.2800	.1951	.0785
7	.2232	.2551	.4535	.3230	.1345
8	.3327	.2782	.3953	.3038	.1432
9	.6022	.4051	.4235	.3337	.1634
10	.1810	.1773	.3370	.2287	.0881
11	.3772	.4709	.5357	.4189	.2031
12	.4086	.4306	.5086	.4010	.1971
13	.3857	.1767	.2626	.2057	.1001
14	.1958	.1822	.3729	.2590	.1028
15	.3076	.3365	.5305	.4027	.1858
16	.3934	.3896	.4970	.3903	.1906
17	.3018	.2688	.3910	.2969	.1363
18	.3350	.4323	.5808	.4476	.2112
19	.2793	.3072	.5049	.3773	.1693
20	.2398	.2378	.3950	.2869	.1225
21	.2753	.2144	.3995	.2980	.1331
22	.3121	.3266	.4255	.3249	.1505
23	.5712	.4223	.4649	.3687	.1825
24	.3687	.4283	.5246	.4094	.1975
25	.2636	.1452	.2229	.1644	.0725
26	.3691	.2893	.3363	.2624	.1266
27	.4041	.5023	.5903	.4651	.2282
28	.6098	.3954	.4212	.3312	.1616
29	.3633	.4108	.4884	.3811	.1833
30	.4055	.3716	.4129	.3260	.1601
31	.2771	.2056	.3168	.2359	.1056
32	.3314	.3692	.4783	.3680	.1732

Table 15. ITEMAL Results: Non-Accommodated Reading

Item	P-Value	Dis.	Biserial	Pt. Biserial	Reliability
1	.6697	.3386	.3876	.2985	.1404
2	.7726	.1406	.2097	.1507	.0632
3	.7574	.3417	.4840	.3526	.1511
4	.7924	.3844	.6204	.4361	.1768
5	.9133	.1548	.5006	.2777	.0781
6	.6676	.2738	.3135	.2410	.1135
7	.7960	.2536	.4058	.2847	.1147
8	.7110	.3324	.4236	.3187	.1445
9	.7811	.4115	.6156	.4382	.1812
10	.7406	.5166	.6819	.5024	.2202
11	.6187	.5185	.5496	.4303	.2090
12	.8810	.2631	.6422	.3944	.1277
13	.7770	.4438	.6698	.4773	.1987
14	.7144	.3377	.4495	.3375	.1524
15	.6917	.4683	.5751	.4362	.2014
16	.7537	.4933	.6811	.4971	.2142
17	.8514	.2626	.5402	.3492	.1242
18	.7974	.3440	.5730	.3997	.1606
19	.7456	.4464	.6265	.4585	.1997
20	.8534	.2874	.6137	.3947	.1396
21	.7896	.3141	.5057	.3565	.1453
22	.9121	.2206	.7054	.3943	.1116
23	.7199	.4363	.5428	.4052	.1820
24	.7749	.3846	.5776	.4133	.1726
25	.9181	.2144	.7461	.4075	.1117
26	.8504	.3242	.6544	.4262	.152
27	.6979	.4464	.5634	.4276	.1964
28	.8981	.2557	.7200	.4186	.1266
29	.6397	.4996	.5455	.4248	.2040
30	.8036	.3831	.6294	.4366	.1735
31	.6561	.5350	.6192	.4782	.2271
32	.6299	.4335	.5006	.3904	.1885

Table 16. ITEMAL Results: Accommodated Reading

Item	P-Value	Dis.	Biserial	Pt. Biserial	Reliability
1	.4863	.2750	.2874	.2292	.1146
2	.6371	.2396	.2839	.2208	.1062
3	.5367	.4854	.5048	.4019	.2004
4	.4520	.5853	.6064	.4820	.2399
5	.7226	.4685	.5572	.4147	.1857
6	.5308	.3318	.3392	.2703	.1349
7	.5633	.4040	.4371	.3471	.1722
8	.4611	.4025	.4271	.3401	.1696
9	.5056	.4854	.5078	.4051	.2025
10	.4475	.5207	.5618	.4463	.2219
11	.4183	.3886	.4302	.3404	.1679
12	.6578	.5653	.6059	.4684	.2222
13	.4611	.4716	.5068	.4036	.2012
14	.5232	.5300	.5419	.4321	.2158
15	.4462	.5407	.5654	.4494	.2234
16	.4993	.5806	.5885	.4695	.2348
17	.5957	.5806	.6080	.4791	.2351
18	.5002	.5392	.5669	.4523	.2261
19	.4548	.5791	.6094	.4847	.2414
20	.5534	.5776	.5908	.4695	.2334
21	.5340	.4762	.5102	.4064	.2027
22	.6898	.5284	.6272	.4773	.2208
23	.5178	.3948	.4006	.3194	.1596
24	.5092	.5269	.5415	.4320	.2159
25	.6339	.6559	.6978	.5436	.2619
26	.5894	.5238	.5639	.4453	.2191
27	.4822	.4762	.5067	.4041	.2019
28	.6497	.5515	.5979	.4634	.2211
29	.4606	.3318	.3733	.2973	.1482
30	.5389	.5361	.5594	.4455	.2221
31	.3674	.4455	.5056	.3949	.1904
32	.3791	.4562	.5035	.3946	.1915

Table 17. ITEMAL Results: Non-Accommodated Science

Item	P-Value	Dis.	Biserial	Pt. Biserial	Reliability
1	.9373	.1295	.4984	.2505	.0607
2	.7623	.2966	.4231	.306	.1303
3	.5997	.4484	.4754	.3742	.1833
4	.6843	.3655	.4406	.3370	.1567
5	.7464	.4107	.5511	.4038	.1757
6	.7051	.4585	.5581	.422	.1924
7	.7973	.4016	.6046	.4216	.1695
8	.4990	.5805	.5803	.4630	.2315
9	.8311	.2748	.4822	.3239	.1213
10	.4511	.4375	.4544	.3612	.1798
11	.6374	.2829	.3176	.2470	.1188
12	.7564	.2308	.3311	.2409	.1034
13	.6600	.4859	.5427	.4185	.1982
14	.6189	.4411	.4754	.3722	.1807
15	.5650	.5377	.5653	.4484	.2223
16	.6086	.4391	.4792	.3766	.1838
17	.3957	.4604	.4856	.3820	.1868
18	.6687	.3618	.4116	.3167	.1491
19	.9176	.1818	.6057	.3344	.0920
20	.6129	.4551	.4888	.3838	.1870
21	.5646	.3968	.4268	.3385	.1678
22	.7746	.4480	.6429	.4598	.1921
23	.6019	.4831	.5070	.3994	.1955
24	.7886	.4183	.6241	.4393	.1794
25	.8613	.2362	.5044	.3214	.1111
26	.7876	.4431	.6691	.4739	.1938
27	.8619	.3251	.6784	.4331	.1494
28	.7984	.3479	.5518	.3856	.1547
29	.8109	.3404	.5725	.3925	.1537
30	.3701	.4158	.4512	.3519	.1699
31	.7409	.5266	.6692	.4933	.2161

Table 18. ITEMAL Results: Accommodated Science

Item	P-Value	Dis.	Biserial	Pt. Biserial	Reliability
1	.7915	.3931	.5780	.4056	.1648
2	.5644	.4998	.5312	.4212	.2089
3	.4046	.3686	.3890	.3064	.1504
4	.5075	.4337	.4513	.3601	.1800
5	.5234	.5381	.5639	.4496	.2246
6	.4520	.5018	.5327	.4234	.2107
7	.5740	.4799	.5014	.3972	.1964
8	.3209	.3889	.4653	.3561	.1662
9	.6204	.5676	.6018	.4715	.2288
10	.2626	.3199	.4265	.3150	.1386
11	.5853	.2867	.3165	.2501	.1232
12	.6340	.3552	.4076	.3175	.1529
13	.4939	.4343	.4587	.3660	.1830
14	.4957	.4892	.4984	.3976	.1988
15	.3828	.4885	.5364	.4209	.2046
16	.4133	.2728	.3062	.2421	.1192
17	.2057	.1956	.3173	.2219	.0897
18	.4829	.3725	.3965	.3162	.1580
19	.7278	.4605	.5811	.4325	.1925
20	.4597	.2903	.3292	.2620	.1306
21	.4274	.3930	.4245	.3362	.1663
22	.5931	.5806	.6070	.4790	.2353
23	.5020	.4331	.4515	.3602	.1801
24	.5640	.5737	.5816	.4612	.2287
25	.6964	.4939	.5700	.4320	.1986
26	.5244	.5406	.5448	.4341	.2168
27	.6545	.6047	.6500	.5034	.2394
28	.5922	.5443	.5597	.4415	.2169
29	.5781	.5394	.5621	.4451	.2198
30	.2913	.2797	.3700	.2793	.1269
31	.5421	.6026	.5964	.4747	.2365

Table 19: Model D Path Coefficients Comparison Across Groups

Path Coefficients			
Item No.	**Group 1 (non-accommodated)**	**Group 2 (accommodated)**	**Path Coefficient Difference**
1	.41	.52	-.11
2	.52	.43	.09
3	.30	.15	.15
4	.43	.40	.03
5	.52	.52	.00
6	.53	.18	.35
7	.54	.39	.15
8	.51	.32	.19
9	.39	.36	.03
10	.36	.25	.11
11	.45	.50	-.05
12	.56	.47	.09
13	.41	.16	.25
14	.37	.29	.08
15	.68	.48	.20
16	.58	.45	.13
17	.45	.32	.13
18	.63	.55	.08
19	.67	.44	.23
20	.43	.31	.12
21	.56	.33	.23
22	.47	.36	.11
23	.48	.42	.06
24	.51	.48	.03
25	.31	.12	.19
26	.43	.25	.18
27	.61	.55	.06
28	.48	.37	.11
29	.49	.43	.06
30	.42	.33	.09
31	.43	.21	.22
32	.51	.42	.09

Table 20. Complete DIF Summary: Math

		SIBTEST			Mantel-Haenszel	
Run No.	Beta-uni	SIB-uni z-statistic	SIB-uni p-value	Chi-square	p-value	(D-DIF)
1	.061	6.982	.000 E	97.97	.000 E	-1.65
2	.018	1.162	.245 E	1.79	.181 E	-.22
3	.020	1.457	.145 E	1.52	.218 E	.18
4	.030	2.166	.030 E	4.85	.028 E	-.30
5	.040	2.632	.008 E	11.85	.001 E	-.50
6	-.002	-.144	.886 E	3.98	.046 E	.33
7	-.020	-1.272	.203 E	6.03	.014 E	.39
8	.010	.658	.510 E	.51	.475 E	.10
9	.041	3.159	.002 E	8.65	.003 E	-.43
10	-.030	-2.114	.035 E	4.23	.040 E	.35
11	.063	4.243	.000 E	29.60	.000 E	-.76
12	.017	1.249	.212 E	.38	.538 E	-.09
13	-.044	-2.868	.004 E	29.36	.000 E	.74
14	-.077	-4.919	.000 E	36.79	.000 E	1.01
15	-.002	-.183	.855 E	.18	.669 E	.07
16	-.032	-2.332	.020 E	10.46	.001 E	.46
17	-.023	-1.459	.145 E	10.46	.001 E	.47
18	-.014	-1.064	.287 E	.61	.435 E	.12
19	-.008	-.525	.600 E	1.79	.181 E	.21
20	-.084	-5.551	.000 E	40.97	.000 E	.99
21	.064	4.217	.000 E	6.48	.011 E	-.38
22	.046	2.911	.004 E	4.33	.038 E	-.30
23	.036	2.973	.003 E	9.55	.002 E	-.45
24	-.011	-.843	.399 E	.08	.781 E	.04
25	.034	2.118	.034 E	.09	.764 E	.05
26	.019	1.205	.228 E	1.93	.164 E	.19
27	.028	2.209	.027 E	9.27	.002 E	-.44
28	.059	4.857	.000 E	21.02	.000 E	-.67
29	.011	.696	.486 E	.22	.641 E	.07
30	.034	2.105	.035 E	.25	.616 E	-.07
31	-.026	-1.676	.094 E	12.52	.000 E	.52
32	-.009	-.613	.540 E	.60	.437 E	.11

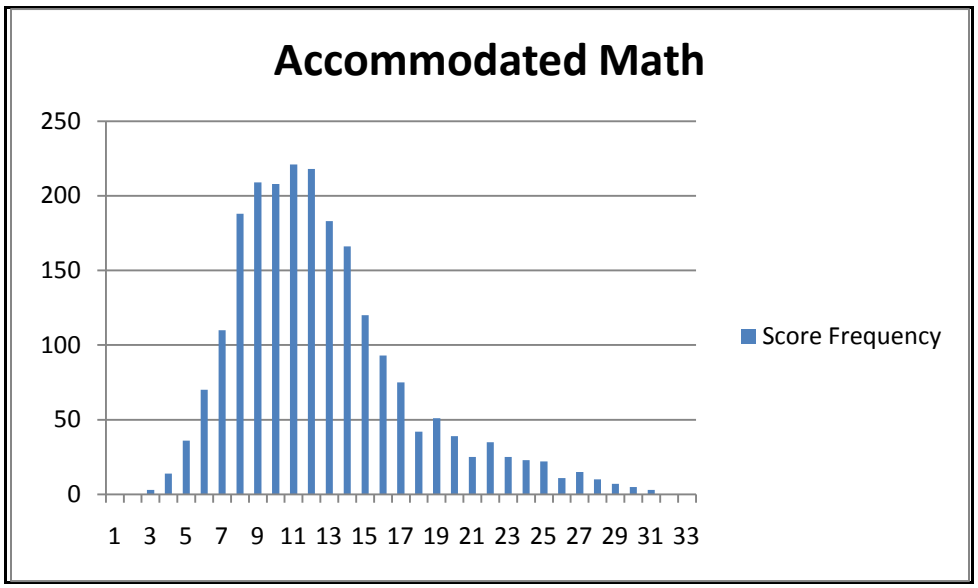
Table 21. Complete DIF Summary: Reading

	SIBTEST			Mantel-Haenszel		
Run No.	Beta-uni	SIB-uni z-statistic	SIB-uni p-value	Chi-square	p-value	(D-DIF)
1	.039	2.309	.021 E	.88	.349 E	-.14
2	.067	4.162	.000 E	4.83	.028 E	-.33
3	-.010	-.702	.483 E	1.37	.242 E	.18
4	.058	3.918	.000 E	21.19	.000 E	-.70
5	.003	.326	.744 E	4.56	.033 E	-.41
6	.003	.151	.880 E	2.65	.104 E	.23
7	.057	3.676	.000 E	11.93	.001 E	-.52
8	.053	3.225	.001 E	8.73	.003 E	-.43
9	.005	.393	.694 E	.82	.366 E	-.14
10	-.030	-2.351	.019 E	.88	.347 E	.15
11	-.049	-3.099	.002 E	15.81	.000 E	.59
12	-.026	-2.438	.015 E	1.20	.273 E	.20
13	.028	1.987	.047 E	7.61	.006 E	-.42
14	-.063	-4.574	.000 E	15.99	.000 E	.61
15	-.061	-4.389	.000 E	10.32	.001 E	.49
16	-.062	-4.968	.000 E	19.50	.000 E	.70
17	.008	.669	.503 E	2.55	.110 E	-.27
18	.025	1.743	.081 E	7.69	.006 E	-.43
19	-.017	-1.220	.222 E	.56	.455 E	.12
20	.031	2.465	.014 E	17.84	.000 E	-.68
21	.002	.175	.861 E	3.85	.050 E	-.30
22	-.016	-1.849	.064 E	.01	.912 E	-.03
23	-.019	-1.240	.215 E	5.70	.017 E	.36
24	-.008	-.586	.558 E	.01	.917 E	-.02
25	.016	1.727	.084 E	15.27	.000 E	-.74
26	-.012	-1.169	.242 E	.54	.463 E	-.13
27	-.056	-3.854	.000 E	19.52	.000 E	.67
28	.000	-.020	.984 E	1.27	.260 E	-.20
29	-.053	-3.559	.000 E	22.35	.000 E	.70
30	-.016	-1.266	.206 E	.26	.610 E	.08
31	.004	.269	.788 E	.00	.946 E	.02
32	-.007	-.402	.688 E	.21	.650 E	.07

Table 22. Complete DIF Summary: Science

		SIBTEST		Mantel-Haenszel		
Run No.	Beta-uni	SIB-uni z-statistic	SIB-uni p-value	Chi-square	p-value	(D-DIF)
1	.018	2.437	.015 E	18.06	.000 E	-.87
2	.022	1.731	.084 E	4.69	.030 E	-.31
3	.031	2.184	.029 E	3.46	.063 E	-.26
4	.005	.402	.688 E	.05	.831 E	-.03
5	.001	.108	.914 E	2.37	.124 E	-.22
6	.028	2.165	.030 E	13.85	.000 E	-.53
7	.033	2.735	.006 E	6.42	.011 E	-.37
8	-.028	-2.013	.044 E	4.02	.045 E	.30
9	.008	.756	.449 E	11.31	.001 E	-.53
10	.029	1.968	.049 E	5.86	.016 E	-.36
11	-.043	-2.977	.003 E	25.74	.000 E	.68
12	.005	.348	.728 E	.34	.561 E	.09
13	-.034	-2.607	.009 E	4.90	.027 E	.31
14	-.088	-7.043	.000 E	26.09	.000 E	.72
15	-.039	-2.802	.005 E	9.20	.002 E	.43
16	.065	4.421	.000 E	8.28	.004 E	-.39
17	.074	5.467	.000 E	17.08	.000 E	-.65
18	.049	3.361	.001 E	5.67	.017 E	-.32
19	.027	2.974	.003 E	19.16	.000 E	-.80
20	.011	.759	.448 E	.86	.354 E	.13
21	-.029	-2.009	.045 E	7.02	.008 E	.36
22	-.055	-5.053	.000 E	13.87	.000 E	.57
23	-.089	-6.690	.000 E	51.93	.000 E	1.01
24	-.003	-.265	.791 E	.96	.328 E	-.15
25	-.012	-1.251	.211 E	.19	.665 E	-.08
26	.035	2.976	.003 E	12.09	.001 E	-.51
27	-.024	-2.847	.004 E	.02	.885 E	-.03
28	.000	.008	.994 E	.84	.360 E	-.14
29	.027	2.345	.019 E	10.18	.001 E	-.48
30	-.078	-5.436	.000 E	25.15	.000 E	.74
31	-.054	-4.927	.000 E	12.98	.000 E	.55

Figure 3. Graphical Score Distributions for Math



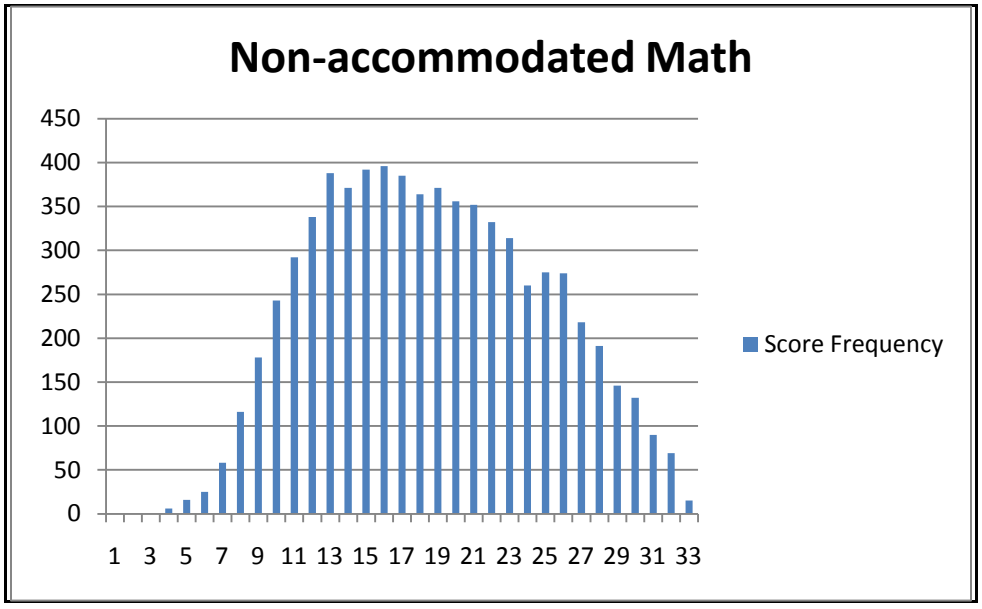
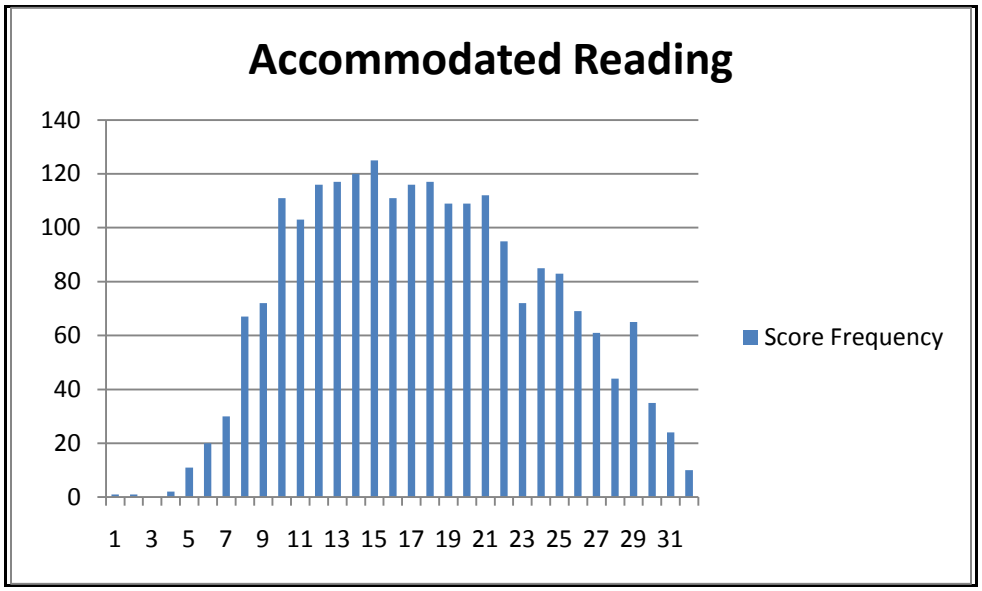


Figure 4. Graphical Score Distributions for Reading



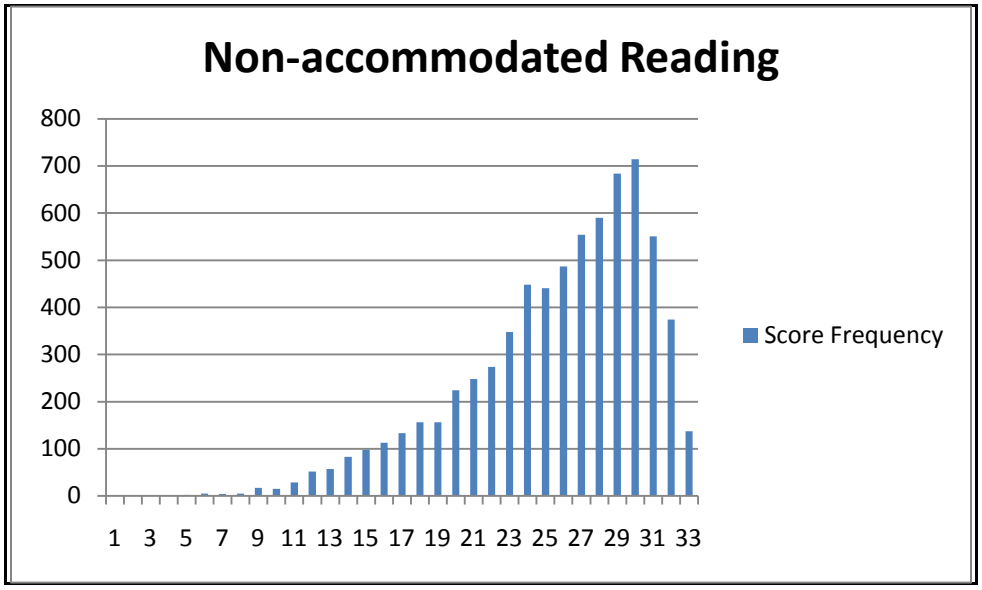
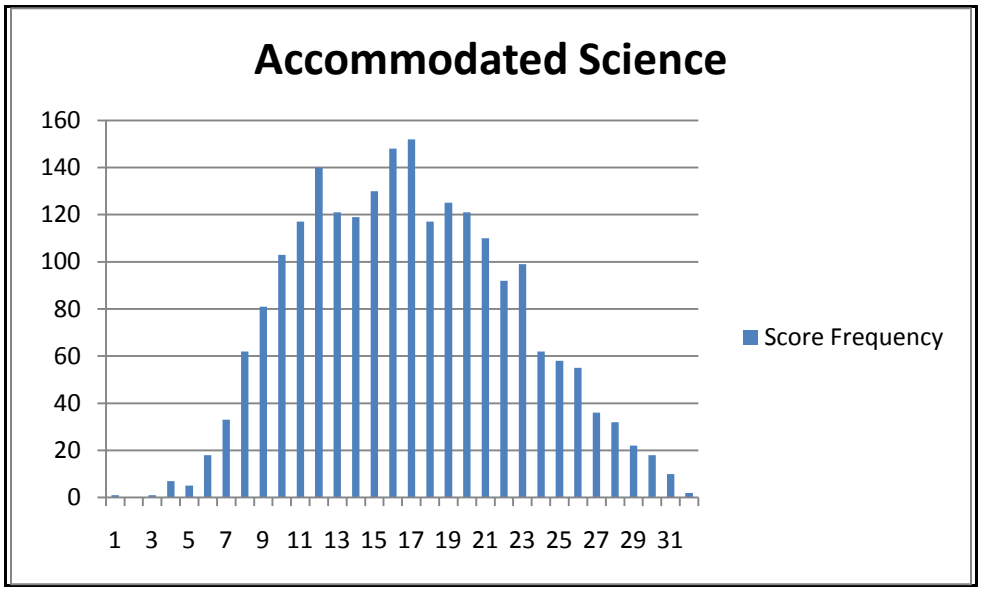


Figure 5. Graphical Score Distributions for Science



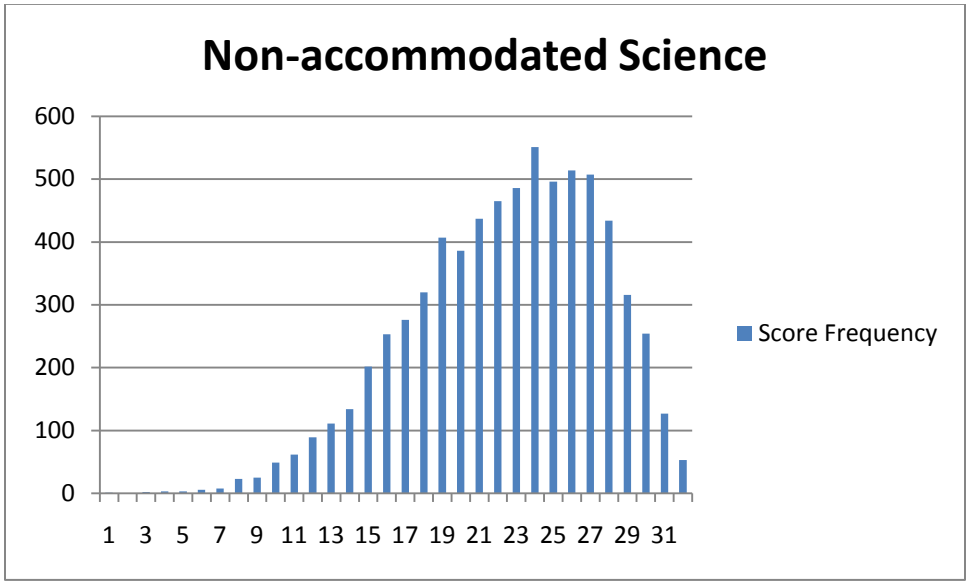


Figure 6. Non-accommodated Versus Accommodated Math Scree Plots

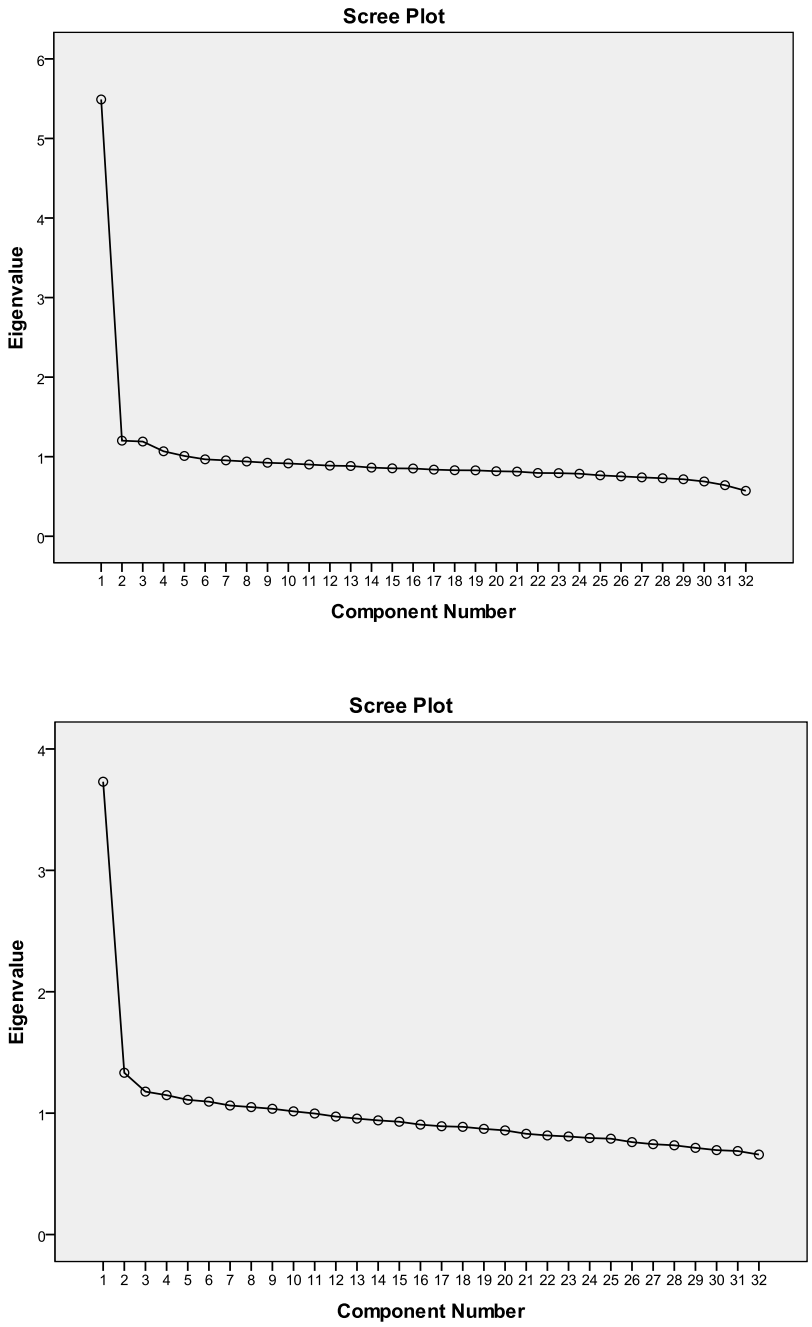
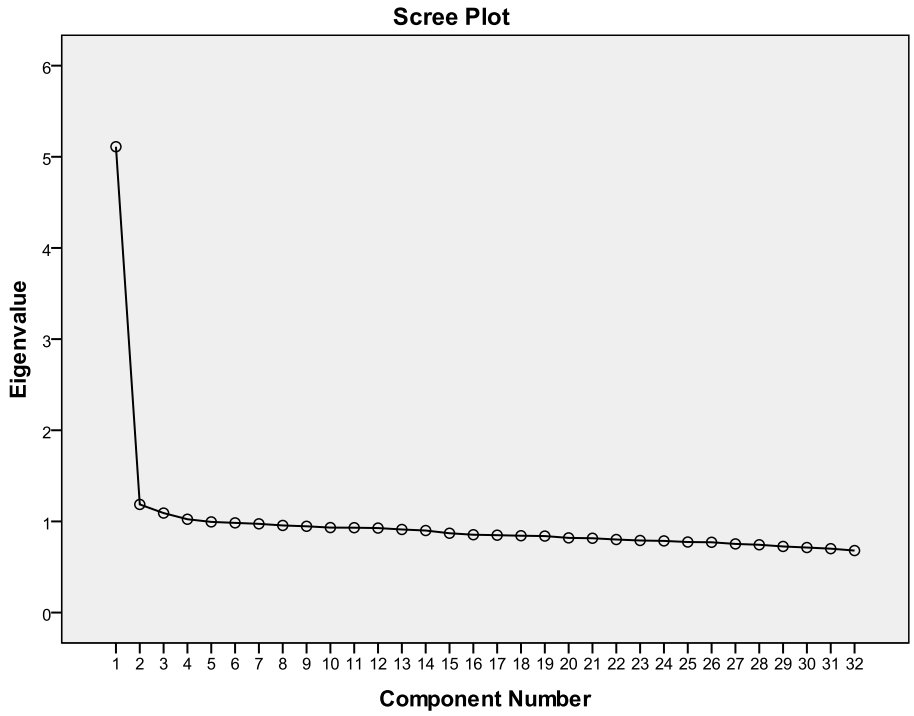


Figure 7. Non-accommodated Versus Accommodated Reading Scree Plots



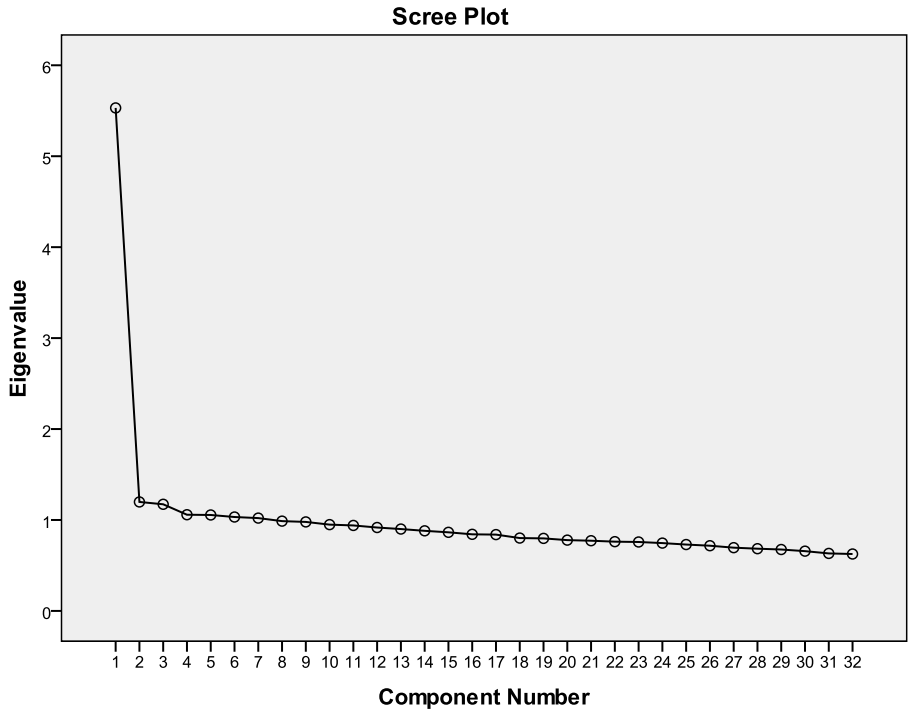


Figure 8. Non-accommodated Versus Accommodated Science Scree Plots

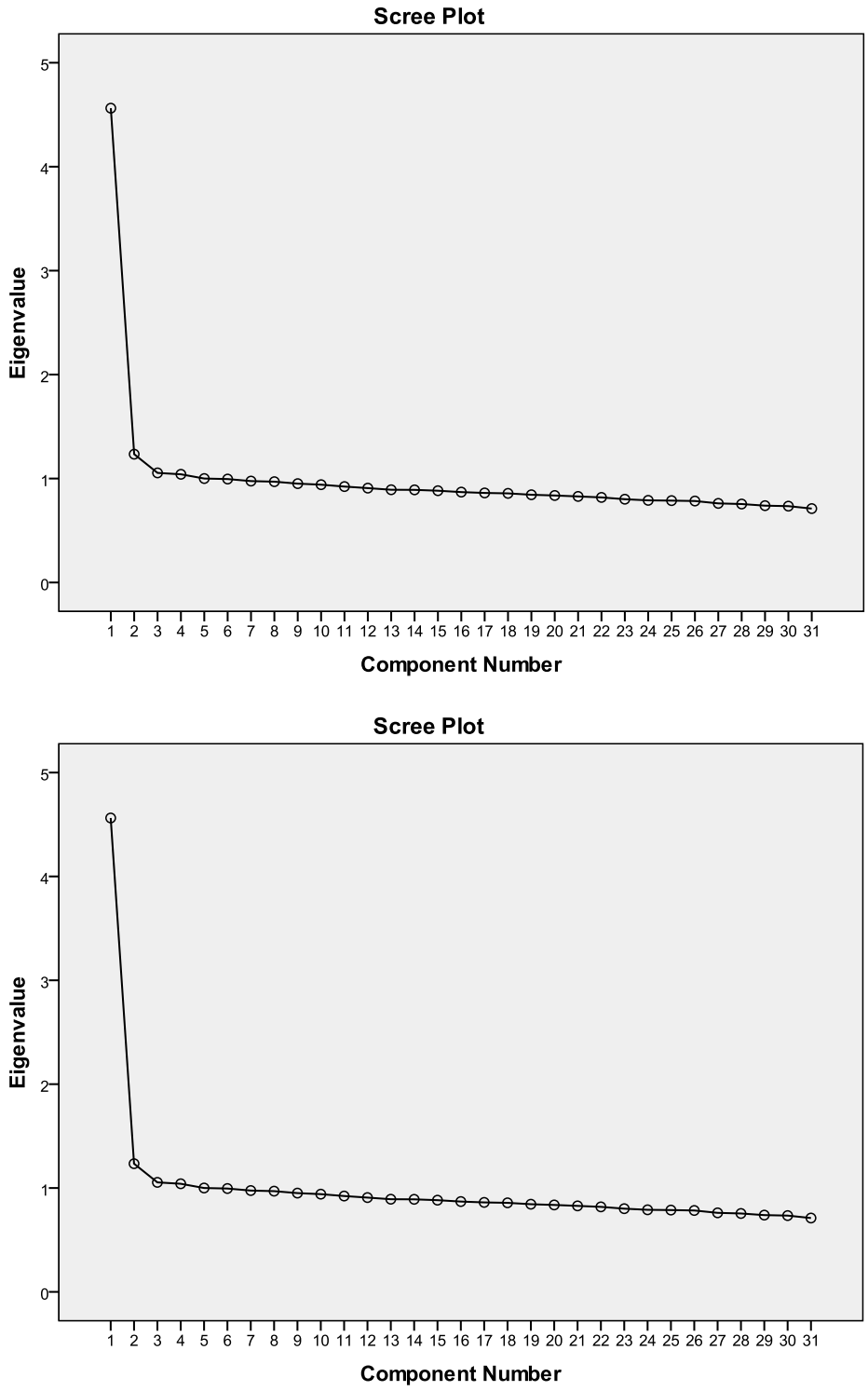


Figure 9. Math Dendrograms: Accommodated and Non-accommodated

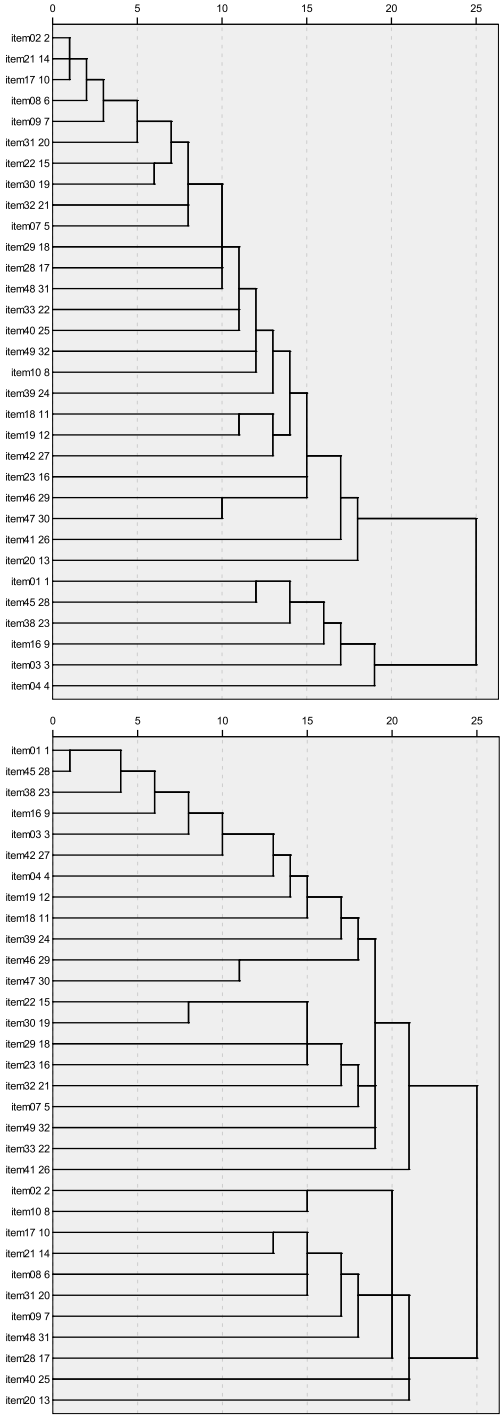


Figure 10. Reading Dendrograms: Accommodated and Non-accommodated

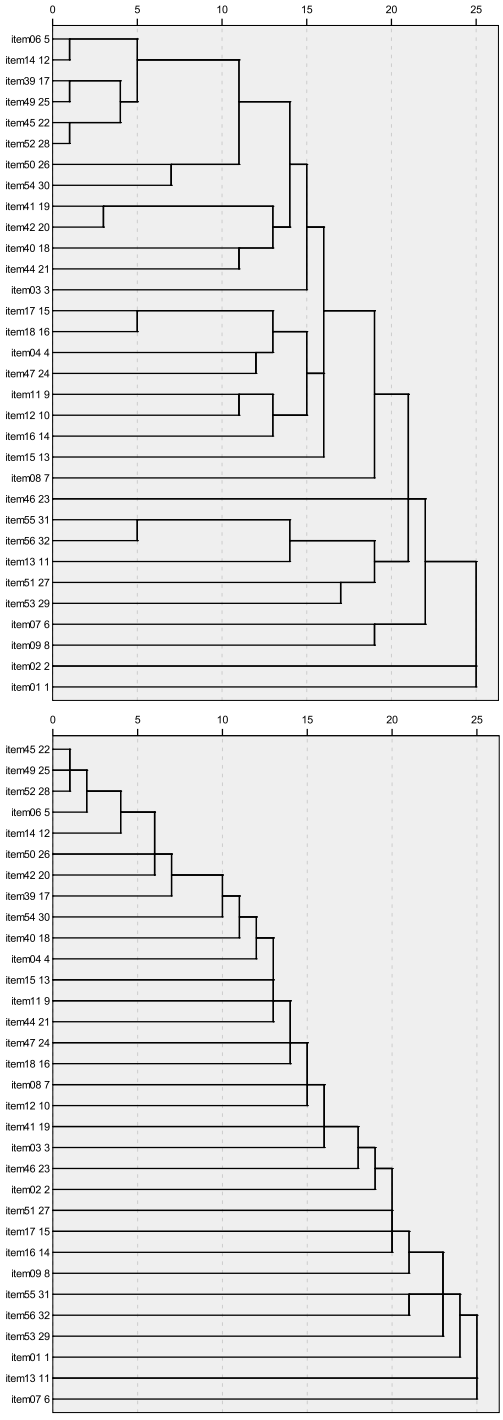


Figure 11. Science Dendrograms: Accommodated and Non-accommodated