Examining the Impact of Accommodations and Universal Design on Test Accessibility and Validity

Author: Maureen Kavanaugh

Boston College
Lynch School of Education

Department of
Educational Research, Measurement and Evaluation

EXAMINING THE IMPACT OF ACCOMMODATIONS AND UNIVERSAL DESIGN ON TEST ACCESSIBILITY AND VALIDITY

Dissertation
by

MAUREEN KAVANAUGH

submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy

May, 2017

ABSTRACT

EXAMINING THE IMPACT OF ACCOMMODATIONS AND UNIVERSAL DESIGN ON TEST ACCESSIBILITY AND VALIDITY

Maureen Kavanaugh

Michael Russell, Chair

Large-scale assessments are often used for statewide accountability and for instructional and institutional planning. It is essential that the instruments used are valid and reliable for all test takers included in the testing population. However, these tests have often fallen short in the area of accessibility, which can impact validity for students with special needs. This dissertation examines two strategies to addressing accessibility: the use of technology to implement principles of universal design to assessment and the provision of accommodations.

This study analyzed test data for students attending high schools in New Hampshire, Vermont and Rhode Island who participated in the 2009 11th grade New England Common Assessment Program (NECAP) science assessment. Three test conditions were of interest: (1) no accommodations with a paper-based form (2) accommodated test administration with a paper-based form and (3) accommodated test administration using a universally designed computer-based test delivery system with embedded accommodations and accessibility features.

Results from two analyses are presented: differential item functioning (DIF) and confirmatory factor analysis (CFA). DIF was used to explore item functioning, comparing item difficulty and discrimination under accommodated and non-

accommodated conditions. Similarly, CFA was used to examine the consistency of underlying factor structure as evidence that constructs measured were stable across test conditions. Results from this study offered evidence that overall item functioning and underlying factor structure was consistent across accommodated and unaccommodated conditions, regardless of whether accommodations were provided with a paper form or a universally designed computer-based test delivery system. These results support the viability of using technology-based assessments as a valid means of assessing students and offering embedded, standardized supports to address access needs.

ACKNOWLEDGEMENTS

I dedicate this work to my husband, Scott.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

Beginning in the 1990s and continuing into the 21st century, educational policy, referred to as standards-based reform, emphasized access to a quality education for all students. At the center of these reform efforts have been the use of large-scale assessments. These assessments have been used for accountability purposes and also to inform instructional and institutional planning. It is therefore essential that the instruments used are valid and reliable for all test takers. However, questions have been raised about whether students with special needs can meaningfully participate on state assessments designed for more typical learners. Alternative assessments and test accommodations are two ways states have tried to facilitate more meaningful participation. Another way assessments can be made more inclusive is through the application of Universal Design for Learning (UDL) during test development. Some states have attempted to apply these principles to paper and pencil assessments, but may be more successful doing so using technology-based assessments. The latter format can offer greater flexibility while improving consistency and efficiency of test administration. Moving in that direction, both the Partnership for Assessment of Readiness for College and Careers (PARCC) and the Smarter Balanced Assessment Consortium (SBAC) have adopted technology-based assessment systems.

This dissertation explores how measurement may be altered, if at all, based on test format and how accommodations are provided. Two strategies for addressing test accessibility are considered – the use of technology to implement principles of Universal Design to assessment and the provision of accommodations. This study explores whether a technology-based approach results in scores that have similar psychometric and

underlying structural qualities as scores collected under non-accommodated, paper-based administration.

BACKGROUND

For most of the 20th century, students with disabilities and other special needs have largely been excluded from the general curriculum and large-scale assessments. This allowed states, districts and schools to avoid accountability for the academic progress of such students (Bechard, 2000). Beginning in the 1990s and continuing into the 21st century, new regulations were put in place to ensure that students with disabilities received greater access to the general curriculum and were included in accountability efforts along with other traditionally neglected groups (Koretz & Hamilton, 2001).

Two of the most influential pieces of legislation in that effort had been the Individuals with Disabilities Education Act (IDEA, 2004) and the No Child Left Behind Act of 2001 (NCLB, 2002). NCLB required states receiving Title I funding to demonstrate proficient student achievement in reading, mathematics and science. To do so, states were required to develop annual academic assessments in these areas, testing at least 95% of all students, including those with disabilities. The most recent iteration of the law, the Every Student Succeeds Act (ESSA) continues to include similar requirements for testing and inclusion of students. The latest version of IDEA, passed in 2004, also reinforces this, stipulating the inclusion of all children with disabilities in all state and district-wide assessments (Crawford, 2007).

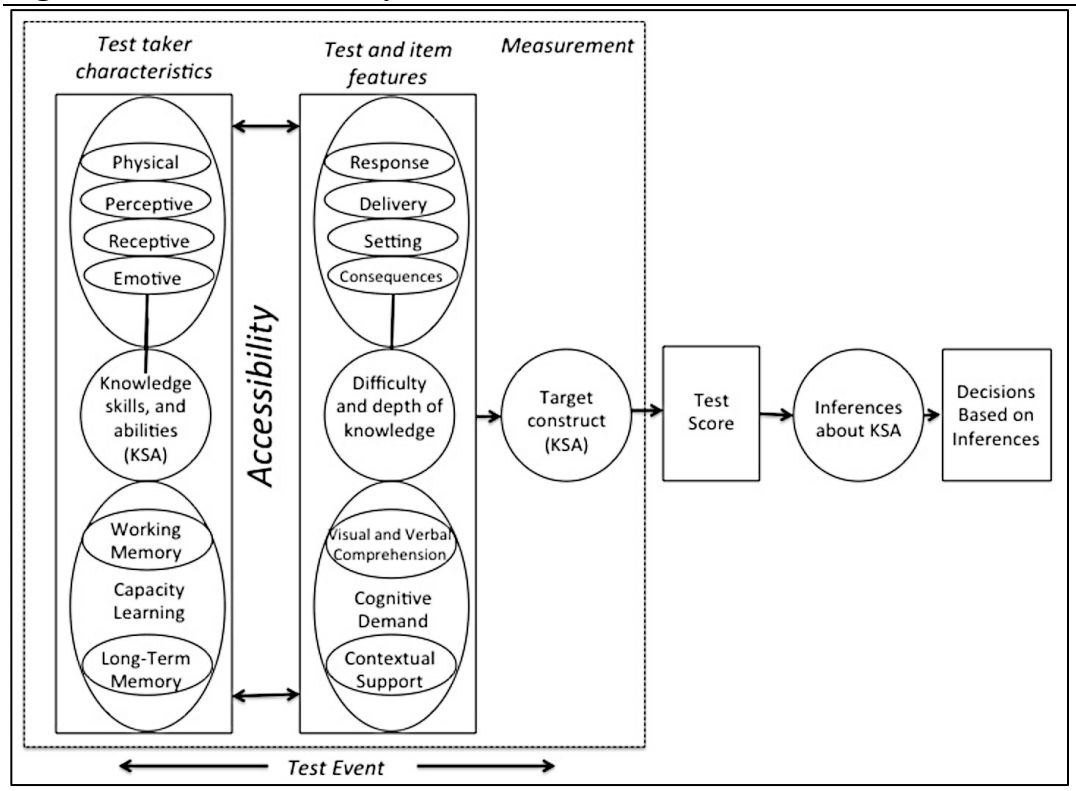
Although these legislative efforts have been successful in increasing participation of students with disabilities and English language learners on statewide assessments, successful inclusion requires states to provide students with appropriate opportunities to

learn and demonstrate their abilities. Questions have been raised about whether many students with special needs could meaningfully participate on statewide assessments used to satisfy requirements of NCLB and now ESSA. Although improvements have been observed in the overall performance of students with disabilities, many of these students continue to lag far behind their non-disabled peers (Chudowsky, Chudowsky, & Kober, 2009). Among other factors, research suggests that this group's low performance is likely the consequence of a lack of appropriate opportunities to learn the general curriculum and a mismatch between assessments and test taker characteristics (Abedi, Leon & Kao, 2008).

STATEMENT OF THE PROBLEM

Concerning the latter issue, state tests tend to fall short in the area of test accessibility. Test accessibility has been defined as "the extent to which a test and its constituent item set permit the test taker to demonstrate knowledge of the target construct" (Beddow, Elliott, & Kettler, 2009, p. 1). As depicted in Figure 1.1, accessibility will be influenced by the interaction between test taker attributes (e.g. cognitive, sensory, linguistic and physical characteristics) and test design and delivery features (e.g. format and presentation method) (Kettler-Geller & Crawford, 2011). Accessibility is compromised when the interaction between individual attributes unrelated to the target construct and test features negatively impacts a test taker's ability to demonstrate what he/she knows and can do (Kettler-Geller & Crawford, 2011).

Figure 1.1. Test Accessibility: Interaction Between Test Taker and Test Features



(Beddow, Elliott & Kettler, 2013)

For example, the intended construct for a state assessment may be to solve problems involving measurement and estimation. Such items may contain substantial printed text that students must read. In order for a student to understand the problem presented and subsequently respond, he or she must be able to detect and decode this printed text. A student with a reading disability or visual impairment may be unable to do so. The test taker is prevented from demonstrating his or her knowledge due to an inability to access item content. According to Beddow (2011), "the test-taker characteristics that interact with test or test item features and either promote or inhibit one's access to the test event are referred to as *access skills*" (p. 381-382). As illustrated, often implicit in the design of many state tests is an assumption that test takers will possess certain access skills (e.g. ability to decode printed text, see a graph, hold a pencil

and legibly handwrite their responses, maintain attention and motivation throughout the test) that are necessary for meaningful participation. However, the extent to which individual students actually possess these skills can vary substantially.

Validity and Test Accessibility

Although access skills are generally considered tangential to the intended construct, a student's lack of those skills may prevent him or her from accessing the target construct. Consequently, test scores will not support the intended inferences about the test taker in relation to the target construct (Beddow, 2011). Thus, issues with test accessibility will have an impact on test validity. As described by Beddow (2011), "conceptually, the validity of inferences made about a test-taker from his or her achievement test score is proportional to the accessibility of the test for that individual" (p. 381).

According to Messick (1995), validity is "an overall evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores or other modes of assessments" (p. 1). The *Standards* offers a similar definition as "the degree to which evidence and theory support the interpretation of test scores for proposed uses of the test" (AERA, NCME, & APA, 2014, p. 11). An important component of validity is the extent to which a test measures a targeted construct without contamination from unintended constructs (Messick, 1989). This refers to a threat known as construct irrelevant variance (CIV), or the introduction of systematic error variance to scores as a result of measurement of constructs irrelevant to the intended construct (Haladyna & Downing,

2004). The presence of such systematic error will lead to faulty interpretations of scores and negatively impacts validity.

In attempts to limit potential sources of CIV, many large-scale assessments incorporate standardized elements. Standardization often includes uniform test content, procedures and conditions, such as scripted test directions, uniform response requirements (e.g. separate answer booklet, filling in a bubble sheet with a number two pencil) and scoring rubrics used for all test takers (Sireci, 2008). The rationale underlying standardization is to keep the measurement instrument and procedures constant so that any observed differences are reflective of true differences among test takers, rather than measurement artifacts (Sireci, 2008).

For some students, however, strict adherence to standardized procedures may prevent demonstration of what they know and can do. Messick (1994) points out that validity is a function of test items, but also of the persons responding to items and the context and conditions in which an assessment takes place. In other words, for a given purpose, an instrument (including standardized procedures) may allow valid inferences for certain students, but perhaps not for others. The concern for statewide assessments, administered to nearly all students is that due consideration of the unique characteristics of students with special needs is not adequately applied during test development. Test design that fails to account for the varied access needs of this population may introduce construct irrelevant variance for these test takers. Consequently, without alteration, these measures may be poor indicators of what these students know and can do.

Addressing test and/or item accessibility can be very challenging. Attempts to improve accessibility for one test taker or type of test taker may hinder access for others.

For example, to promote access for students with visual impairments, items may be verbally delivered in the place of printed text. However this would diminish accessibility for students with hearing impairments. Furthermore, students often have multiple accessibility related difficulties (e.g. inability to read items and handwrite responses) or vary in the degree of support needed. Test developers must therefore consider a variety of solutions to address the full range of accessibility issues within the student population.

Definition and Rationale for Using Test Accommodations

States have attempted to address accessibility by providing individualized test accommodations to students with special needs. Test accommodations are defined as any change in the way a test is administered or how a student responds (Driscoll, 2007) and are often grouped in the following categories:

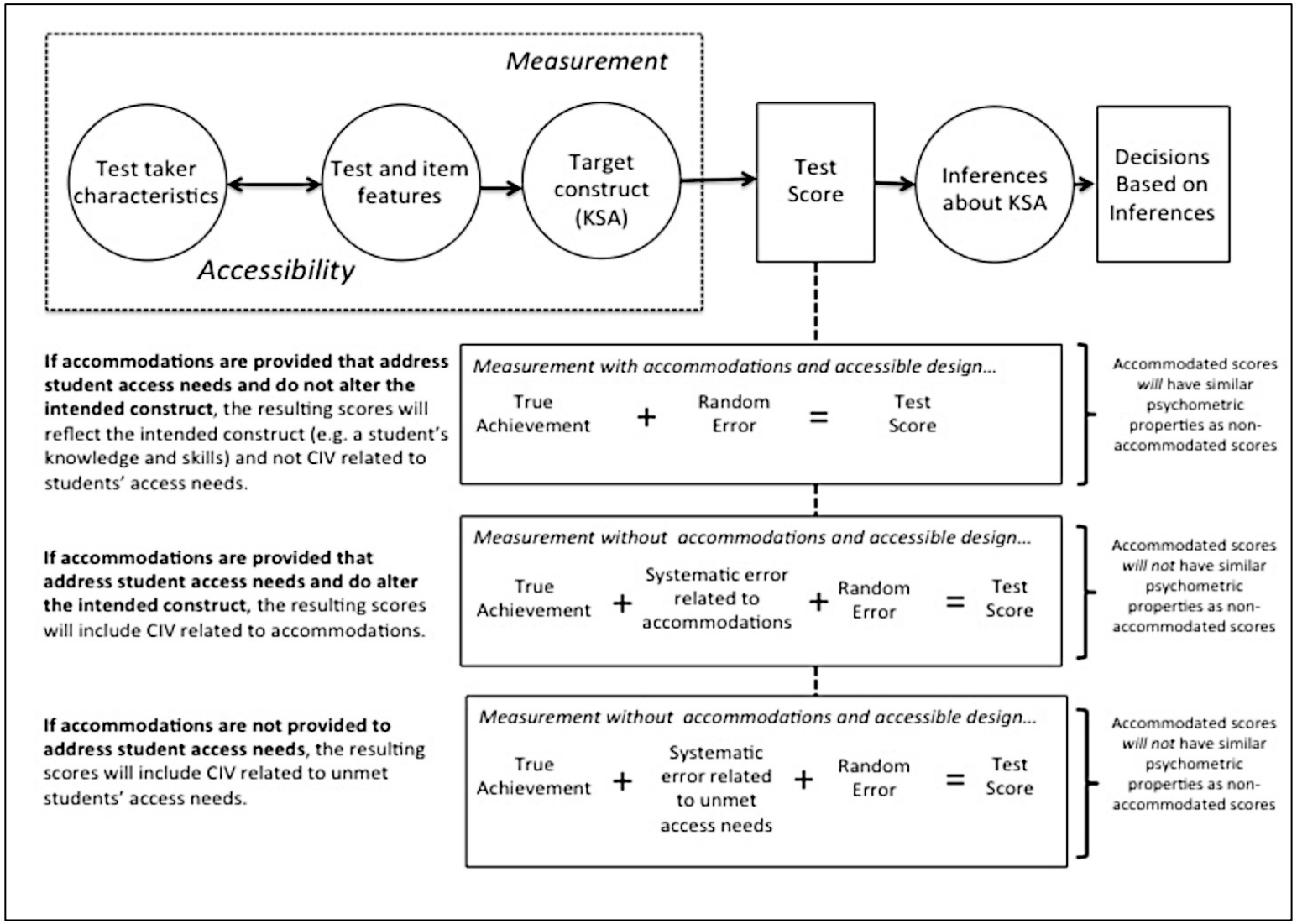
- *Presentation* (e.g. read aloud, native language translation, large print)
- *Equipment and material* (e.g. calculators, manipulatives, amplification equipment)
- *Response* (e.g. scribe records answer, record answers in test booklet)
- *Timing*/*scheduling* (e.g. extended time, breaks)
- *Setting* (e.g. small group or individual administration, study carrel, separate room) (National Center for Educational Outcomes, 2009).

These supports are said to provide students with the opportunity to participate on a more equal playing field with their non-disabled peers (Driscoll, 2007) and improve validity by removing access barriers that result from disability or other disadvantages (Elliot, Kratochwill & Schulte, 1999).

The goal in providing accommodations is to ensure that test scores reflect the intended construct (e.g. a student's knowledge and skills) and not a student's disability or access needs. Figure 1.2 depicts the effect test accommodations are intended to have on the measure provided by a test. The top section of Figure 1.2 depicts the interaction between students' characteristics, including their access needs, and test item(s) as well as the subsequent inferences and decisions. The resulting effect of this interaction on measurement will manifest itself in a student's test score, which will include true achievement and any random and non-random error resulting from an unmet accessibility need or unintended alteration of the intended construct produced by an accommodation. The subsequent rows depict the true score model and the effects that accommodations may have on the error associated with a test score. If the accommodations provided address a student's access needs and do not alter the intended construct, the resulting scores will reflect only that construct (e.g. a student's knowledge and skills) and not CIV related to students' access needs. Under these conditions, CIV related to accessibility is reduced or eliminated and accommodated scores should exhibit similar psychometric properties as non-accommodated scores (e.g. similar item functioning and underlying factor structure). If the accommodations provided address a student's access needs and do alter the intended construct, the resulting scores will include CIV related to accommodations. Under these conditions, non-random error is introduced and item functioning and underlying factor structure is expected to vary between accommodated and non-accommodated scores. Finally, if students are not provided accommodations or not provided with accommodations that address all their access needs, the resulting scores will include CIV related to unmet access needs. The result, again, is the

introduction of non-random error, which would alter the psychometric properties of accommodated scores.

Figure 1.2. Measurement Without and With Accommodations and Accessible Test Design



For example, a read aloud accommodation is intended to provide access to students who have difficulty or cannot decode (e.g. students with reading related learning disabilities) or detect print text (e.g. students with visual impairments). If the read aloud successfully permits access to test content without altering the intended construct (e.g. application of scientific method), the resulting test score should reflect a student's true achievement and correct inferences can be made. Without this accommodation, a student may be unable to access test content (e.g. a scientific problem presented through printed text) and would face a severe disadvantage. In this example, this could result in the test underestimating a student's ability to apply the scientific method. This, in turn, could lead to a false impression of individual and institutional performance when interpreting test results.

In addition to meeting the access needs of students, it is equally critical that an accommodation itself does not alter the construct being applied by the student (Phillips, 1994). A valid accommodation should reduce construct irrelevant variance (often related to accessibility barriers) (Elliot, Kratochwill, & Schulte, 1999), but should not become a source of construct irrelevant variance itself. When an accommodation reduces the difficulty of a task, for instance, providing an unfair advantage for accommodated test takers, the accommodation becomes a source of what Messick (1989) referred to as construct irrelevant easiness. Extending the example of read aloud to a literacy assessment, if the accommodation does impact the intended construct (e.g. reading comprehension), a student's ability could be overestimated and result in faulty inferences about a student's literacy skill development. Adaptations that alter the target construct

are typically referred to as “modifications” or “non-standard accommodations” and the resulting scores are generally not included in state accountability calculations.

Although accommodations have the potential to promote greater equity, Sireci (2008) points out two major questions that have fueled the ongoing debate concerning the value of accommodations: “Do the test scores that come from non-standard test administrations have the same meaning as test scores resulting from standard administrations?” and “Do current test accommodations lead to more valid test score interpretations for certain groups of students?” (p. 82). The validity of an accommodated test will depend both on the extent to which the special need(s) of a test taker are met and the impact the accommodation has on the measured construct (Thurlow et al., 2000).

Challenges in Providing Accommodations

Given the requirements of ESSA and IDEA, it seems likely that students with diverse needs will continue to take part in statewide assessments and many will do so using accommodations. A summary of public accountability reports found that more than 50 percent of students with disabilities received accommodations on high school state reading tests in 38 states and in 34 states for mathematics tests in the 2013–14 school year (National Center for Education Outcomes, 2016). In spite of widespread use of accommodations, students with disabilities continue to underperform in academic achievement compared to their non-disabled peers (Abedi, Leon & Mirocha, 2003; Ysseldyke et al., 1998). Abedi, Leon and Kao (2008) speculate that among other factors (e.g. fewer learning opportunities), test format and inappropriate or incomplete test accommodations play a role in this group’s low achievement. The range of accessibility

issues resulting from poor test design may not be fully addressed by accommodations (e.g. a read aloud would not remedy issues arising from distracting graphics or pictures).

Accessibility issues can be further compounded by inconsistent or low quality test accommodations (e.g. lack of or improper training of human readers). Some commonly used accommodations, such as sign language interpretation, read aloud or use of a scribe, are provided by an access assistant. This is a person who provides test takers with specialized support during an assessment (Clapper, Morse, Lazarus, Thompson & Thurlow, 2005). Training and multiple checks by experts are necessary to ensure consistency among access assistants, but this process can be costly and time-consuming (Clapper, Morse, Lazarus, Thompson & Thurlow, 2005). Without proper training, the quality and consistency of accommodations may vary and result in the inequitable provision of appropriate support across a testing program. The presence of such variability will have implications for the validity of test scores for students who rely on access assistants (Clapper, Morse, Thompson & Thurlow, 2005).

Further, given limited resources, space and personnel available, schools may be unable to provide students with all needed accommodations. It is often not feasible to provide students with individual access assistants (e.g. scribes, readers, sign language interpreters) or develop multiple translated tests to address the needs of multiple language groups, for example. A 2003 Rhode Island Department of Education study revealed that some schools "bundle" accommodations for groups of students rather than follow individual IEP recommendations and noted considerable inconsistency between daily accommodations provided during instruction and those made available during

assessments. Some recommended accommodations in particular, such as computers and other assistive devices, were rarely used during assessments.

Universal Design

Despite challenges and setbacks in improving test accessibility through accommodations, these efforts do represent progress toward more inclusive assessments. However, sole reliance on accommodations fails to get at the heart of the issue: that many traditionally used large-scale assessments were not designed from the start with diverse learners in mind (Johnstone, Anderson & Thompson, 2006). Universal Design for Learning offers a framework that may guide development of more inclusive assessment systems and calls for attention to accessibility needs starting from test conception.

The ultimate goal in developing universally designed assessments is to "reduce the need for accommodations and various alternative assessments by eliminating or reducing access barriers to content associated with the tests themselves" (Thompson, Johnstone & Thurlow, 2002, p. 6). To accomplish this, Universal Design for Learning calls for the creation of more flexible instruments that provide multiple means for students to engage, recognize, and express understanding of test material.

Working with the fixed medium of paper, some states have attempted to integrate universal design during assessment development (e.g. New England Common Assessment Program in Vermont, Rhode Island and New Hampshire). However, this still required multiple versions of test materials and test proctors with specialized skills (e.g. ability to translate test into American Sign Language or another language) (Russell, Hoffmann, & Higgins, 2009), which may be difficult to provide in systems where

resources are scarce. Technology-based testing would permit multiple accessibility features within a single interface, eliminating the need for multiple versions and in some cases, human access assistants. Also, technology-based accommodations can be offered with more consistency. Due to its inherent flexibility, the use of technology for assessment is likely to play a key role in incorporating principles of UDL, making an individualized approach more feasible (Dolan, Hall, Banerjee, Chun, & Strangman, 2005).

Changes in both federal law and policy suggest support for broader application of UDL, including in the design of statewide assessments. The latest authorization of the Elementary and Secondary Schools Act signed in 2015, the Every Student Succeeds Act (ESSA) includes several references to universal design. Universal Design for Learning is also emphasized in the national educational technology plan, Future Ready Learning (Office of Educational Technology, 2016).

Applicants for the federal Race to the Top (RttT) grant program were required to submit proposals for assessment systems that were valid, fair and reliable and "designed to assess the broadest possible range of students, including English learners and students with disabilities" (U.S. Department of Education, 2010, p. 18172). Applicants were also required to "use technology to the maximum extent appropriate to develop, administer and score assessments and report assessment results" (U.S. Department of Education, 2010, p.18175). The result was approximately three hundred fifty million dollars awarded to two consortia of states for the development of a new generation of common assessments – the Smarter Balanced Assessment Consortium (SBAC) and the Partnership for Assessment of Readiness for College and Careers (PARCC). Both test consortia have

developed and piloted computer-based assessments with embedded accessibility tools as well as locally provided accommodations. Both consortia reference principles of Universal Design for Learning in their test design frameworks (Partnership for Assessment of Readiness for College and Careers, 2016; Measured Progress & National Center on Educational Outcomes, 2014).

RESEARCH QUESTIONS

The central problem of this dissertation concerns the validity or what Sireci (2008) refers to as the "construct equivalence" of non-accommodated and accommodated test scores. Do scores collected under accommodated and non-accommodated conditions support the same inferences about what students know and can do? A secondary aim is to explore whether a universally designed, technology-based approach to testing and accommodations was effective at addressing accessibility concerns and enhanced the validity of scores for students with special needs. This dissertation examines the underlying structural and psychometric characteristics of student scores across accommodated and non-accommodated conditions and paper-based and technology-based assessment forms. In doing so, the following research questions are considered:

- Is the underlying factor structure consistent for scores gathered under accommodated and non-accommodated conditions?
- Do items function similarly under accommodated and non-accommodated conditions? Specifically, holding ability constant, are item difficulty and discrimination equivalent for accommodated students and non-accommodated students?

- If differential item functioning is exhibited, do patterns of DIF and item characteristics suggest that accommodations or use of accessibility supports may be related to DIF?

The underlying premise of these analyses is that an assessment should measure the same construct(s) and function similarly regardless of test taker characteristics unrelated to the target construct. Variation in scores should reflect only differences in achievement levels. Principles of UDL applied during test development along with the use of accommodations and other supports are intended to minimize sources of construct irrelevant variance related to accessibility barriers. If these efforts are successful and measurement is consistent for all students, various item and score characteristics (e.g. item parameters and factor loadings) should be similar across accommodated and non-accommodated testing conditions (Bolt & Ysseldyke, 2007).

To answer these questions, a secondary data analysis of assessment data collected during the 2009 administration of the 11th grade NECAP science assessment was conducted. During this administration, schools were given the option of using a paper assessment with traditionally offered accommodations or a computer-based test that offered embedded accommodations and accessibility supports. The remaining majority of students took a paper assessment with no accommodations. Using differential item functioning analysis and factor analytic techniques, the psychometric and structural qualities of student scores collected under different testing conditions were examined.

SIGNIFICANCE OF STUDY

A key component of standards-based reform has been the use of mandatory large-scale tests based on state content and performance standards. These tests are used to monitor the progress of states and schools toward more widespread student proficiency and act as a mechanism to ensure that all students are provided with appropriate opportunities to learn the general curriculum. The goal of building inclusive assessment systems is to ensure that these educational systems are accountable for all students. Historically, this was not the case.

Without uninhibited access to test materials and procedures, some students may be unable to show their knowledge and understanding. This will impact the validity of inferences drawn from student scores. This, in turn, can result in misleading conclusions and affect decisions about individuals and institutions. Given the inclusion of students with special needs on statewide assessments and the need for accurate performance information, it is essential that the instruments used to measure student achievement are valid and reliable for all students, and that construct-irrelevant variance related to accessibility is minimized (Abedi, Leon & Kao, 2008).

CHAPTER SUMMARY

This chapter introduced the problem of test accessibility and its relationship with validity with respect to test takers who participate on large-scale assessments and have unique access needs. Strategies for addressing test accessibility and enhancing validity for the full range of test takers were also discussed. This included the provision of test accommodations and application of Universal Design for Learning (UDL) during test

development. Finally, research questions for this dissertation were introduced. The next chapter will explore previous literature and research relevant to these issues. This includes attempts to study the impact of test accommodations, conceptual definitions of universal design applied to assessment, the role of technology and the related challenges.

CHAPTER 2: REVIEW OF RELATED LITERATURE

Although legislative efforts, including NCLB and IDEA, have been successful in increasing participation of students with disabilities on statewide assessments, successful inclusion requires states to provide students with appropriate opportunities to learn and demonstrate their abilities. Test accommodations are one way states have tried to facilitate meaningful participation of students with special needs. Another way assessments can be made more inclusive is through the application of UDL during test development. This dissertation focuses on both these strategies and their impact on test validity for students who receive accommodations.

This chapter provides an overview of relevant research and literature on test accommodations and UDL in the context of large-scale assessment. The first section focuses on test accommodations and includes a discussion on individualized accommodations, research implications, a discussion of the criteria used to judge accommodations, and a summary of previous research examining the impact of individualized accommodation packages on test validity. The second half of this chapter focuses on Universal Design. This section includes a conceptual definition of Universal Design related to assessment and a description of design elements and principles. This is followed by a discussion of the role of technology and accommodations in universally designed assessments. This chapter concludes with a summary of available research examining attempts to integrate UDL during large-scale assessment design and its impact on the test validity.

THE IMPORTANCE OF INDIVIDUALIZED ACCOMMODATIONS

A considerable number of accommodations studies have focused on the impact of a single accommodation (e.g. Harker & Feldt, 1993; Helwig, Rozek-Tedesco & Tindal, 2000; Helwig, Rozek-Tedesco, Tindal, Heath & Almond, 1999; Kosciolek & Ysseldyke, 2000; Maihoff, 2000) or standard packages of accommodations (e.g. Abedi, Hofsetter, Baker, & Lord, 2001; Fletcher, et al. 2006; Hafner, 2001) for a general group of students with disabilities, regardless of whether included participants demonstrate a need for accommodations (Bolt & Ysseldyke , 2007). Some researchers considered this approach inauthentic (Elliot, Kratochwill, McKevitt, & Malecki, 2009; Kim, Schneider & Siskind, 2009a) and believe it fails to account for the diversity of needs among students with disabilities (Bolt & Ysseldyke, 2007). These researchers argue that in operational settings, access needs among students vary and therefore recommended accommodations also tend to vary from student to student, even among those who share the same disability label. Furthermore, students often receive more than one accommodation to address the full range of their access needs. Thus, construct irrelevant barriers to test material may still exist for test takers whose access needs have not been met fully in single accommodation studies. Consequently, results may be misleading.

Rather than attempt to isolate the impact of individual accommodations, some researchers have focused on "individualized accommodations" instead (Lang, Elliot, Bolt & Kratchowill, 2008). These studies examined the overall effect of a group of accommodations or various combinations of accommodations assigned based on the individual needs of participants. A limitation of this research is that it can be difficult to disentangle the impact of individual accommodations and the extent to which specific

accommodations address specific access needs (Schulte, Elliot & Kratochwill, 2001). Nevertheless, accommodations packages must be matched to a student's needs to be beneficial (Ketterlin-Geller, 2005). Given the research design and questions posed in this dissertation, the next sections summarize only studies that examine the impact of individualized accommodations on test validity. That is, only studies in which accommodations were matched to participants' individual needs were selected. Student participants may have been purposely selected for inclusion based on their demonstrated need for the target accommodation, or participants may have been assigned a package of accommodations based on their individual characteristics.

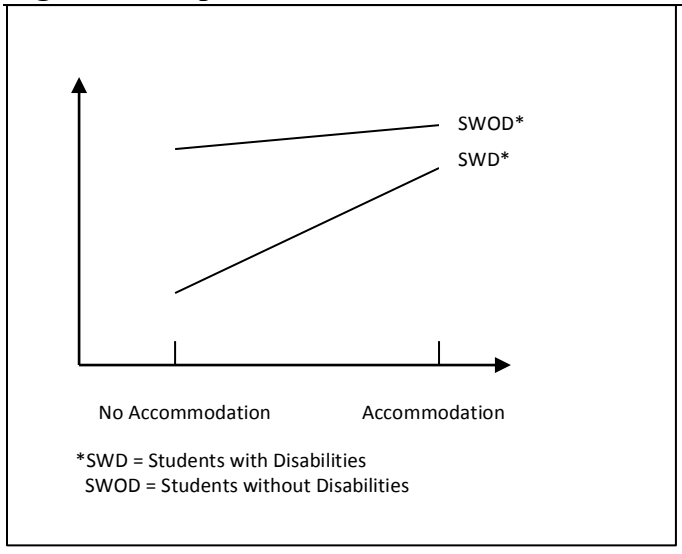
RESEARCH ON THE VALIDITY OF ACCOMMODATIONS

Controversy around test accommodations concerns the concept of test validity and the comparability of score interpretation collected under accommodated and non-accommodated conditions. Thurlow and colleagues (2000) offered three commonly used criteria to judge the impact of accommodations on validity. The first two criteria specify that a valid accommodation or set of accommodations positively affects the test performance of students with the targeted need, but shows no impact on the performance of students who do not demonstrate that need. An accommodation should have a differential effect for students with special needs (Phillips, 1994). This pattern is generally referred to as "differential boost" (Bolt & Ysseldyke, 2007). The third is that the accommodation does not change the psychometric properties of the measure, which concerns measurement consistency. The following sections summarize research that applied to each of these criteria.

Differential Boost

In determining if accommodations have a differential effect on students (as evidenced by a "differential boost"), students with and without disabilities are generally testing under accommodated and non-accommodated condition. If an accommodation is appropriate, students identified with a particular access need due to disability or other disadvantage addressed by the accommodation are expected to derive a greater benefit than students who do not have the same identified need (Bolt & Ysseldyke, 2007). That is, students with special needs are expected to exhibit a greater boost in total test score moving from non-accommodated to accommodated conditions, than students without special needs. Accommodations are likely to have a negative impact on validity if both students with and without disabilities demonstrate the same boost in scores (Phillips,1994)

Figure 2.1. Expected Outcomes Under Differential Boost Hypothesis



Below is a summary of findings from four studies that examined differential boost. Three out of the four studies provided some evidence that accommodations have a differential effect for students with disabilities, though results varied by grade and subject.

In a dissertation completed by McKevitt (2001), 79 eighth grade students with and without disabilities were asked to complete two halves of a standardized reading test. Students completed one half with no accommodations and the other half with teacher-recommended accommodations (21 students with disabilities and 20 students without disabilities) or teacher accommodations plus read aloud accommodation (19 students with disabilities and 19 students without disabilities). The order in which each half of the test was administered was randomly assigned, but accommodations were always received during the second test administration. Mckevitt found that individualized accommodation packages with and without read aloud had little effect on student test scores for either group. As a group, no differential boost was exhibited for students with disabilities. For individual students within groups, accommodations did result in a small, but positive boost (based on effect size) in scores for 50% of students with disabilities and 38% of students without disabilities.

Kettler, Niebling and Mroch (2003) found that test accommodations did have a significant impact on reading test performance for fourth grade students with disabilities, but not for eighth grade students. Their study included 196 fourth and eighth grade students (118 fourth graders, including 49 with disabilities, and 78 eighth graders, including 39 with disabilities). Students were asked to complete two forms of the math subtest and two forms of the reading subtest for the TerraNova Multiple Assessment

Battery. For each pair of subtests, students completed one form with accommodations and one form without accommodations. On accommodated administrations, students with disabilities were provided accommodations recommended on their IEPs. Students without disabilities were randomly paired with a student with a disability and given the same set of accommodations. A statistically significant interaction between disability status and accommodation condition was present for fourth grade reading scores ($F(1, 66) = 7.58$, $p < .05$), but not fourth grade mathematics scores ($F(1, 15) = 2.83$, $p < .05$). There was no significant interaction between disability status and accommodation condition present for either test at the eighth grade level. Gains were noted for individual students within each group.

Lang, Elliott, Bolt and Kratochwill (2008) conducted a similar study with 102 fourth grade (including 32 identified as having a learning disability) and 68 eighth grade students (including 32 identified as having a learning disability). They also administered two forms of reading and math subtests of the TerraNova Multiple Assessment Battery. Educators assigned students with disabilities individualized accommodation using the Assessment Accommodations Checklist (AAC, Elliot, Kractochwill & Schulte, 1999), and based on students' IEPs. As in Kettler, Niebling and Mroch (2003), students with disabilities were matched to one or two students without disabilities. Students within pairs received the same accommodations and took two forms of the both the math and reading test in a randomly assigned order. Although both groups showed gains under accommodated conditions, test accommodations were found to have a significant differential positive effect on the performance of students with disabilities in reading. In

math, students in both groups benefited from accommodations and the effect was not significantly different among students with and without disabilities.

Schulte, Elliot and Kratochwill (2001) found that fourth grade students with disabilities benefited more from accommodations than did students without disabilities on multiple-choice items, but not for constructed response items. Eighty-six fourth grade students (including 43 students with disabilities) were asked to complete two forms of the TerraNova Multiple Assessment mathematics subtest. Each student with at least one identified disability was assigned accommodations based on teacher recommendations and paired with another student with no identified disability who received the same package of accommodations. There was a significant interaction between test condition and disability status for multiple-choice items ($F = 7.92$, $p = .004$), but not for constructed response items ($F = .05$, $p = .42$). For multiple-choice items, there was a small to medium effect for students with disabilities (.41) and no effect for students without disabilities. For constructed response items, effect sizes were similar for both groups (.31 and .35, respectively).

Summary and Limitations of Differential Boost Research

Findings for research that applied differential boost criteria have been mixed. For groups of students, differential boost was not consistently observed. Results varied by grade, subject area and item type. However, substantial performance differences were often noted for individual students. This may indicate ineffective assignment of accommodations for some participants. Among differential boost research, no studies were located that examined the impact of individualized accommodations on science

assessments. All located research focused on fourth and eighth grade— grades at which students commonly participate in statewide testing. No differential boost studies involving high school students were located.

Bolt and Ysseldyke (2007) pointed out a number of general drawbacks to differential boost studies. The first is that treatment groups have often been formed on the basis of disability status, as was the case in the studies summarized. This approach rests on the assumption that disability status alone is indicative of need for an accommodation. Bolt and Ysseldyke caution that research suggests students who have no identified disability may in some cases have similar access needs as students with identified disabilities, and will consequently also demonstrate a boost in scores. Furthermore, students with the same disability label may vary in their need for accommodation.

Also, repeated measures designs, where fatigue or practice effects may be a concern, were employed in all of the above mentioned studies. To address these concerns, researchers often counterbalanced or randomly assigned participants to the order in which they receive test conditions, as was the case in Kettler, Niebling and Mroch (2003), Lang, Elliott, Bolt and Kratochwill (2008), and Schulte, Elliot and Kratochwill (2001), but not in McKevitt (2001). To address practice effects, researchers administered two different, but parallel forms of the test. Lang, Elliot, Bolt and Kratochwill, (2008) also noted that testing did not include the same high stakes present in operational tests, which may have influenced student motivation during research. This limitation also applied to other differential boost studies summarized here.

Another limitation cited by researchers was the inability to provide explanation for why students do not perform better, or actually do worse, when provided accommodations. One hypothesis is that accommodations have been incorrectly provided or mis-assigned to students. Although it was assumed that accommodations appropriately matched student needs, Schulte, Elliot and Kratochwill (2001) found that roughly a third of the students with disabilities performed worse when provided with individualized accommodations. Researchers theorized that this negative effect was due to mis-assignment of accommodations.

Measurement Comparability

Score comparability across student groups and/or testing conditions is considered important evidence that a test will support fair and unbiased decision making (AERA, APA & NCME, 1999). In accommodation studies, examination of measurement comparability is based on the assumption that if a test measures the same construct for all students, various item and score characteristics (such as item parameters and factor loadings) should be similar across accommodated and non-accommodated testing conditions (Bolt & Ysseldyke, 2007). Two types of analyses have been commonly used in accommodations research to examine score comparability: differential item functioning (DIF) and factor analysis. In the next sections, findings are presented from research that examined differential item functioning and factor structure for scores collected under accommodated and non-accommodated conditions. In all ten studies that considered differential item functioning, some evidence of DIF was identified, suggesting that accommodations may have an impact on item difficulty and discrimination. Like

research on differential boost, results varied by grade and subject area, but also by item type. Research that considered factor structure suggested similar constructs were measured under accommodated and non-accommodated conditions. In the six studies presented, researchers noted factor invariance among accommodated and non-accommodated scores.

Differential Item Functioning: Previous Research

Koretz (1997) analyzed student assessment data from the 1995 Kentucky Instructional Results Information System (KIRIS) in fourth and eighth grade mathematics and reading. Koretz found no evidence that test items discriminate less well for students with disabilities who had received accommodations based on item-to-total score correlations, a finding that contradicts previous research. However, he cautioned that a large number of students left items blank, thereby limiting the confidence one can have in the reported correlations.

Using logistic discriminant function analysis (DFA) described by Miller and Spray (1993), Koretz found statistically significant differential item functioning for 13 of 22 common items in both subject areas for students in both grades who received accommodations. Of the items exhibiting DIF, seven were more difficult for students with accommodations, and six were easier for students with accommodations. Further analysis showed that for most items, DIF was modest in reading and more substantial in mathematics. Most items exhibiting DIF showed non-uniform DIF (nine items). For these items, differential difficulty was more pronounced for students with accommodations who had total test scores within a specific range. This range varied

among items. Some items that were found to be differentially easy for students with accommodations required more reading than other items and did not involve nonverbal representations of information. Koretz speculated that accommodations, particularly those that support reading, might have reduced the difficulty that reading demands contribute to this item. However, Koretz was unable to determine if DIF was associated with particular accommodations due to small n's for these subsamples.

In a follow up study, Koretz and Hamilton (1999) used state test data from fourth, fifth, seventh, eighth and eleventh grade to replicate findings, apply analyses in additional subject areas, and compare performance across different item types. Using the same technique as the previous study, researchers again found DIF for both open-response (38 out of 48 items) and multiple-choice items (112 out of 190), which was generally limited to students who received accommodations. For open response items, most items exhibiting DIF were found on mathematics assessments. However, an item on the social studies assessment and an item on the science assessment also exhibited DIF. Items that favored non-disabled students tended to have substantial reading requirements or included fairly advanced mathematical vocabulary. For multiple-choice items, fewer items displayed DIF at higher grade levels than at lower grade levels. The authors speculated that this pattern might relate to more frequent assignment of inappropriate accommodations that have an impact on item difficulty at lower grade levels than at higher grades. Koretz and Hamilton recommended more research to uncover the underlying reasons or cause of DIF.

Bielinski, Thurlow, Ysseldyke, Freidebach and Freidebach (2001) analyzed data from the 1998 Missouri assessment program to examine the impact of read aloud

accommodation on the characteristics of multiple-choice items from a third grade reading and fourth grade mathematics assessment. Using operational test data, the authors examined item functioning for 4 groups: 1000 randomly selected general education students who took tests without accommodations; randomly selected general education students who took tests without accommodations who were matched in ability to students with IEPs in reading (n = 995; reference group); all students with IEPs in reading who took tests without an accommodation (n = 600) and; all students with IEPs in reading who took tests with a read aloud accommodation (n = 661). Students who received a read aloud accommodation also typically received other accommodations.

Item difficulty estimates were generated using an IRT three parameter logistic model and then compared for each group. Many more items exhibited DIF for students who received accommodations than other student groups. Nineteen reading items out of 41 and six mathematics items out of 32 exhibited DIF for students with accommodations. Ten reading items and one mathematics item exhibited DIF for students with disabilities who did not receive a read aloud accommodation. Only one reading and no mathematics items exhibited DIF for general education students matched in ability to IEP students. Based on the number of items exhibiting DIF for students who received a read aloud accommodation, results of this study suggested that this accommodation might not be an appropriate solution to improve test accessibility (Bielinski et al., 2001).

Bolt (2004) similarly focused on read aloud accommodation by examining the degree of DIF present on reading and math tests. In this study, the authors analyzed data from multiple states, across multiple years (2001-2003 for one state program, 1998-2001 and 2002 for two others) in both math and reading in elementary and high school. Bolt

looked at DIF for several targeted accommodations, including read aloud, setting accommodations, dictated response and extended time, along with packages of accommodations (e.g. accommodations for sensory/physical vs. cognitive disabilities). In some cases, students who received other accommodations in addition to the target accommodation were included in groups.

For each analysis, groups of 1,000 randomly selected students without disabilities served as a reference group. Like Bielinski et al., (2001), a three parameter logistic model was applied to estimate item parameters. The magnitude of DIF was determined by calculating the weighted average of the vertical distance between focal and reference item characteristic curves as described by Wainer (1993). Magnitudes were evaluated using criteria established by Dorans and Holland (1993). Bolt found that overall, the results followed expected patterns. For example, in all three datasets analyzed, a larger proportion of items exhibited DIF for students who received read aloud accommodations for reading sections than for math sections. Students with sensory/physical disabilities and students with cognitive disabilities who received accommodation did not substantially differ in the number of items exhibiting DIF. Varying levels of measurement comparability were observed across different accommodations, supporting the belief that in general, accommodations should not be considered as either "appropriate" or "inappropriate." Finally, Bolt noted that while some patterns were consistent across datasets, others were not. This finding is important because it suggests that findings based on data from one test may not necessarily generalize to another, even within the same subject area. Bolt suggested that this inconsistency might be due to

varying skills and knowledge being tested by the different assessments, resulting in a differential impact by accommodations.

Finch, Barton and Meyer (2009) analyzed data from students in grades three through eight on language arts and mathematics subtests used during the standardization of a nationally normed educational achievement test. In this study, students with disabilities who received accommodations (between 92 and 197 students, depending on grade level and subject) were compared to those with disabilities who did not receive accommodations (between 212 and 346 students, depending on grade level and subject). Accommodation assignment was determined by local administrators using the same procedures normally used for other large scale tests. The majority of students in the accommodated group received multiple accommodations.

Using SIBTEST (Shealy & Stout, 1993), Finch and colleagues identified a number of items exhibiting DIF in both language and mathematics and at all grade levels. For many of the language items, DIF was non-uniform, suggesting that accommodations may not have the same impact for students at different ability levels. All language items demonstrating uniform DIF favored students with disabilities not receiving accommodations. The authors also noted that all language items demonstrating DIF required the test taker to refer back to the passage to locate a numbered sentence. In mathematics, the majority of flagged items displayed uniform DIF. The authors could not identify any shared characteristics across these items that might explain differential functioning. The magnitude of DIF was considered "large" using guidelines from Roussos and Stout (1996) for almost all flagged items in both mathematics and language.

Further analysis using logistic regression found that uniform DIF was not associated with any specific type of accommodation (i.e. question read aloud, directions read aloud, alternative test setting, or extended time). According to Finch, Barton and Meyer (2009), this may suggest that DIF identified using SIBTEST was the result of the combined impact of multiple accommodations for these items. Specific accommodations were found to be associated with non-uniform DIF for at least some items, suggesting that students at lower proficiency levels may not benefit from these accommodations for these items.

Research by Cohen, Gregg and Deng (2005) suggests that student content knowledge may more accurately explain differential functioning on mathematics items than accommodation status. Data for this study came from a statewide mathematics test administered in 2003 for 1,250 students with learning disabilities receiving extended time, and 1,250 non-accommodated randomly selected ninth-grade students with no identified disabilities. Of the 29 multiple-choice items analyzed using IRT, 22 items exhibited DIF with 13 being easier for accommodated students and nine being easier for non-accommodated students. Items demonstrating DIF were found among all four sub domains (algebra, plane geometry, intuitive geometry and arithmetic). Additional analysis suggested accommodation status might not be a sufficient explanatory variable for differential item functioning. Results indicated that level of skill development or ability in specific areas of mathematics appeared to better explain differential item difficulty. Cohen, Gregg and Deng (2005) concluded that given the findings of this study, "accommodations are more appropriately viewed as simply leveling the playing

field; they do not supply the knowledge necessary to pass tests" (p. 231), and that accommodations should be viewed as "no different than reading glasses" (p. 232).

Bolt and Ysseldyke (2007) also identified a large number of DIF items on statewide mathematics tests for two groups of accommodated students with disabilities in fourth and eighth grade. Bolt and Ysseldyke performed a DIF analysis of data from three consecutive annual administrations for accommodated students with physical disabilities (n = 17,978 fourth graders; n = 16758 eighth graders), accommodated students with mental disabilities (n = 361 fourth graders; n = 253 eighth graders), and non-accommodated students without disabilities (two groups of 1000 randomly selected fourth and eighth graders). Among all accommodated students, extended time and small group or individual setting accommodation were the most common. Again, the three parameter logistic model was used to estimate item difficulty, discrimination and pseudo-guessing parameters for each focal and reference group. Similar patterns of DIF were found across both grade levels. The largest proportion of moderate to large DIF items was identified for accommodated students with physical disabilities and a smaller proportion of such items was found for students with mental disabilities. The authors concluded that measurement in this case was not highly comparable between students with physical and/or mental disabilities receiving accommodations, and non-accommodated students with disabilities.

However, Bolt and Ysseldyke (2007) also pointed out that it might not necessarily be the case that accommodations result in poorer measurement for students with disabilities. Instead DIF may be evidence that target skills were perhaps more accurately measured for accommodated students and less well among students without

accommodations. To explore this, one should examine flagged items within the context of the accommodations received. Researchers did not do this for this study.

While examining differential distractor functioning, Middleton and Laitusis (2007) also looked at differential item functioning among accommodated and non-accommodated students on reading/language arts assessments. Middleton and Laitusis (2007) examined item functioning for students with learning disabilities who received no accommodation, students with learning disabilities who received a read aloud accommodation, and students with learning disabilities who received some form of accommodation other than read aloud. Data came from responses to 81 multiple-choice items included on a fourth grade English language arts test. Thirty thousand students without disabilities were randomly selected as a reference group, bringing the total number of students included in the analysis to approximately 45,000.

Using students without disabilities as a reference group, nine items exhibited DIF for students with a learning disability who received a read aloud accommodation. Four of these nine items were differentially easier and five were differentially more difficult for students who received a read aloud accommodation. Only one item displayed DIF for students with learning disabilities receiving no accommodations and two items exhibited DIF for students with learning disabilities that received accommodations other than read aloud, all of which favored students without disabilities. Using students with disabilities receiving no accommodation as a reference group, no items exhibited DIF for students with a learning disability receiving accommodations other than read-aloud. Two items exhibited DIF for students with a learning disability who received a read aloud. Of these two items, one item favored students without disabilities and one item favored students

with a disability receiving read aloud. The authors did not include further analysis or discussion of the possible causes of DIF.

Stone, Cook, Laitusis and Cline (2010) looked specifically at item functioning for students with visual impairments who completed either large print or braille testing forms. Using the Mantel Haenszel method, the authors looked at fourth and eighth grade multiple-choice items included on a statewide English language arts test for students with no disability and students who are blind or have other visual impairments. At the fourth grade level, of the 75 items analyzed, ten items exhibited DIF for students who were visually impaired and completed the large print form. Half of these items were reading items and half were writing. Only one item (in writing) exhibited large DIF in favor of students without disabilities. Other items were split in terms of which group was favored, but these differences in item function were not considered substantial. When the two accommodated groups were combined (large print and braille test forms), four items exhibited DIF (two in reading and two in writing). None of these items exhibited substantial DIF. Similarly, at grade eight, only one item exhibited substantial DIF favoring students without disabilities for both analyses (large print and braille). Of the 75 items analyzed, five items exhibited "intermediate" DIF (four in reading, and one in writing) when just students who completed a large print form were included as a focal group. When both students who completed a braille form and students who completed a large print form are analyzed together as a focal group, nine items exhibited DIF (seven in reading and two in writing) and only one of these items exhibited large DIF (i.e. an effect size that exceeds 1.5 delta units as defined by Dorans and Holland, 1993). Across items, the authors noted several patterns across items exhibiting DIF. Items that involve

visualizing and metaphors seemed to be differentially easy for students with visual impairments. In some cases, item and passage content seemed related to differential functioning. For example, items and passages that included content that may be of little or no interest or potentially offensive to students with visual impairments (e.g. photography) may have contributed to differential functioning. With only two items exhibiting substantial DIF, Stone and colleagues concluded that overall, their results suggested sufficient accessibility and validity of items for students who are visually impaired and received these two accommodations.

Summary of Differential Item Functioning Research

Although studies used different criteria for identifying DIF, it was common for researchers to identify at least some items that exhibited DIF for students with disabilities who received at least one accommodation. Instances of DIF among accommodated and non-accommodated conditions were noted in multiple subject areas (i.e. math and reading/language arts). This research suggested that accommodated scores may not be interpreted the same as non-accommodated scores in some instances. No studies could be identified that examine science items for differential functioning. Authors of these studies frequently pointed out the need for complementary experimental studies to further investigate the impact of accommodations (e.g. Bolt & Ysseldyke, 2007).

Factor Analysis: Previous Research

Pomplun & Omar (2000) used confirmatory factor analysis to examine factor invariance of a statewide fourth grade mathematics assessment. The authors fit a two-

factor model to data from three groups of students: students without disabilities (n = random sample of 1,500), students with disabilities who received a read aloud accommodation (n = 240) and students with disabilities who did not receive accommodations (n = 1,369). The results indicated factorial invariance across all three student groups, though test reliability was found to be highest for students without disabilities. Their findings support the practice of aggregating student scores across each of these three student groups and provided evidence of score comparability among testing conditions (Pomplun & Omar, 2000).

Kim, Schneider and Siskind (2009a) conducted a similar study examining factor invariance across read aloud administrations and non-accommodated conditions for students with and without disabilities. They also concluded that science test scores could be interpreted across testing conditions and student groups. To reach their conclusions, researchers analyzed data from a 2005 administration of statewide science tests in grades six, seven and eight. Tests included both multiple-choice and constructed response items for three student groups: students with disabilities who received a read aloud accommodation, students with disabilities who received standard administration and students without disabilities who had standard administration. The authors conducted a confirmatory factor analysis based on established test specifications using the following criteria for model fit: Standardized Root Mean (SRM) < .08, Root Mean Square Error of Approximation (RMSEA) < .06 and Comparative Fit Index (CFI) > .95. The authors found similar factor structure among all three student groups, providing evidence of measurement consistency.

Huynh and Barton (2006) also looked at the effect of read aloud accommodations on underlying factor structure, but did so for a state reading test administered to tenth grade students. Like Kim and colleagues (2009), they analyzed data for three groups of students: students with disabilities who completed a special oral form (n = 822), students with disabilities who completed the standard form without accommodations (n = 3,022), and students without disabilities who also completed the standard form without accommodations (n = 85,457). A preliminary principle components analysis indicated a single factor model was appropriate. The authors then performed confirmatory analysis using the following criteria to assess model fit: Root Mean Square Error of Approximation (RMSEA) < .06, Standardized Root Mean Residual (SRMR) < .09 and Goodness-of-fit indexes (GFI) > .90. Huynh and Barton found that all indexes were within the acceptable range, indicating good model fit for all three groups and suggesting that oral accommodations had little impact on the internal structure of the test. They concluded that in this case oral accommodations had leveled the playing field for students who received them, and did not have a negative impact on the validity of test scores for this group.

Cook, Eignor, Sawaki, Steinberg and Cline (2010) also concluded that accommodations had little impact on the underlying factor structure of a fourth grade statewide English language arts test. Cook and colleagues evaluated the impact of read aloud accommodation along with other test accommodations though confirmatory factor analyses that included four groups: students without disabilities who did not receive accommodations, students with disabilities who did not receive accommodations, students with learning disabilities who received accommodations specified in their 504 or

IEP plans, and students with learning disabilities who took the test with a read aloud accommodation. Random samples of 500 students were selected from each group. Researchers first performed a series of single group confirmatory factor analyses for each group to determine the fit of a one- or two-factor model. The authors found a one-factor model more appropriate for their data and then carried out multi-group confirmatory analysis to compare factor structure across student groups. Though results were not clear cut, the authors concluded that factor structure was very similar across all four groups. The authors presented their findings as support for the provision of read aloud accommodations on reading tests, and of state policies aggregating student scores collected under accommodated and standard conditions for accountability purposes.

Cook, Eignor, Steinberg, Sawaki and Cline (2006) examined factor invariance under accommodated and non-accommodated conditions on the Gates-MacGinitie Reading Test (GMRT). Their analysis included data from a sample of fourth grade public school students with and without reading-based learning disabilities, divided into four groups:

- Group One: Students without disabilities who took the test without accommodations
- Group Two: Students without disabilities who took the test with a read aloud accommodation
- Group Three: Students with a reading based learning disability that took the test without accommodations
- Group Four: Students with a reading based learning disability that took the test with a read aloud accommodation

Researchers first conducted single group exploratory analyses to determine the underlying factor structure for each group. This was followed by a single group confirmatory analysis for each group, and two multi-group confirmatory analyses, used to look at factor invariance under non-accommodated and read aloud conditions for students without disabilities (Groups one and two) and students with disabilities (Groups three and four). Researchers found that a one factor model was appropriate for their data collected under non-accommodated and accommodated conditions. The results also indicated factorial invariance among groups for both comparisons. They cautiously concluded that it is likely that the same construct is measured under both accommodated and standard conditions, but also cited contradictory evidence from previous research.

In a dissertation, Harris (2008) used confirmatory factor analysis with structural equation modeling to test for factorial stability among scores collected under two modes of administration: human read aloud and a computer-based read aloud. Analysis was conducted using data from a statewide English language arts assessment in grades six, seven and eight administered in 2006 and 2007. Harris analyzed data from students who whose IEP or 504 teams assigned a read aloud by a human reader following a script or a recorded read aloud of the same script delivered via CD-ROM. Results suggested factorial invariance among testing conditions. Harris also conducted a multivariate analysis of covariance (MANCOVA) to compare student performance and found no significance performance differences between read aloud conditions, after controlling for prior English language arts performance.

Summary of Research Examining Factor Structure

Accommodations research that examined underlying factor structure across testing conditions has been limited to evaluations of read aloud accommodations. These studies found evidence of factor invariance among accommodated and non-accommodated conditions and across multiple subject areas and grade levels. At least one study (Kim, Schneider & Siskind, 2009a) examined the impact of read aloud on a science assessment and observed similar factor structure for groups of middle school students. Given similar underlying factor structure between accommodated and non-accommodated conditions, Cook et al. (2010) suggested that read aloud in combination with other accommodations may be offered as a valid support to students with disabilities. Similarly, results from Harris (2008) suggested that read aloud delivered by a live reader or as a technology-based accommodation might both be offered as a valid support based on factorial invariance across testing conditions. More research applying this approach is needed to evaluate the effects of other combinations of accommodations during operational testing.

Limitations of Studies Examining Measurement Comparability (DIF and Factor Analysis)

Both DIF and factor analysis require large data sets to obtain stable estimates. Thus, studies presented in the last two sections that take this approach often relied on data from operational large-scale assessments. Bolt and Ysseldyke (2007) pointed out two advantages of using this type of data. The first is that test takers are more likely to be provided with accommodation packages tailored to their individual needs.

Accommodation assignment for these studies was generally based on IEP team decisions. The second advantage of using operational test data concerns the conditions under which the test data is collected. Student test performance will have real consequences for test takers and schools, and represents normal testing conditions to which researchers aim to generalize.

However, researchers also noted a number of limitations of using operational data. For instance, the use of non-experimental designs limited the ability of researchers to make causal statements about the impact of accommodations (Bolt, 2004; Bolt & Ysseldyke, 2007; Finch, Barton & Meyer, 2009; Kim, Schneider & Siskind, 2009a). Also, students often receive multiple accommodations during operational testing. This can make it difficult for researchers examining operational data to disentangle the impact of a particular accommodation on item functioning (Finch, Barton & Meyer, 2009) and underlying factor structure. Operational conditions can also introduce the threat of confounding factors, such as variation in accommodation assignment procedures between schools and provision of accommodations (Kim, Schneider & Siskind, 2009a). Finch, Barton and Meyer (2009) pointed out that accommodation policies varied across school districts in their study and differences in item functioning may have related to differences in assignment and provision of accommodations. Another consideration is the challenge of ensuring appropriate administration and provision of accommodations (Bielinkski et al., 2001; Bolt, 2004; Bolt & Ysseldyke, 2007). For example, Bielinski et al. (2001) suggested that a flawed read aloud accommodation may be responsible for DIF, which in their study was provided by a human proctor to small groups or individual students.

In studies that attempted to analyze data for specific accommodations, small samples were mentioned as a limitation (Bolt, 2004; Finch, Barton and Meyer, 2009) that may have resulted in diminished statistical power. Stone et al. (2010) stated that small sample sizes prevented analysis of particular subgroups altogether. In some cases, English language learners (ELLs) were excluded from studies (e.g. Huynh & Barton 2006; Middleton & Laitusis, 2007). Consequently, findings may not generalize to ELLs, who are frequently provided with accommodations. Some authors also suggested that findings be interpreted with caution due to ability differences between groups (e.g. Middleton & Laitusus), though some researchers did attempt to use groups matched roughly in ability or adjusted for ability differences statistically. Finally, authors noted that studies did not address the question of whether accommodations benefit students without disabilities (Huynh & Barton, 2006). Researchers highlighted the need for other types of complementary research (e.g. cognitive lab studies, experimental research) to provide further evidence of the validity of accommodated administrations (Kim, Schneider & Siskind, 2009a; Cook et al., 2010).

For DIF research, not all studies examined the content of items to identify potential causes of DIF (Bolt & Ysseldyke, 2007). This was likely a consequence of test security concerns and restricted access to state test items. However, examination of flagged items is needed to make more conclusive statements about the possible reasons for differential functioning and the role accommodations may play in item functioning.

Challenges of Accommodations Research

Researchers face several challenges when conducting accommodations research in general. First, authors often noted that results generalized only to the subject areas tested during their research. Previous research has most frequently examined reading and mathematics assessments, leaving other content areas such as science and social studies understudied. Also, disability status has often been used as a grouping variable. Since disability status is a pre-existing characteristic, it is not possible to randomly assign students to treatment and control groups. Researchers must also contend with ethical issues associated with withholding accommodations when a student has demonstrated a need, particularly in operational settings in which high stakes may be attached to student performance. Johnson, Kimball, Brown and Anderson (2001) also cited educator reluctance to participate in additional student testing for the purposes of research as potential barrier.

Finally, researchers must carefully think about how accommodations are operationally defined and assigned to students, which can vary from state-to-state and among studies. For example, a read aloud accommodation may refer to the reading aloud of directions only, but in other instances refer to read aloud of directions and item stems, or directions, item stems and answer choices. How an accommodation is provided (e.g. human reader versus screen reader, separate braille form versus refreshable braille tablet) can also vary. The assignment procedures and the operational definition of accommodations used during research will impact the extent to which results can be generalized to different testing situations.

UNIVERSALLY DESIGNED ASSESSMENTS

Another approach to achieving inclusive assessment systems is the integration of universal design principles during test development. Universal design in education is intended to increase access for a wide range of learners by addressing three potential barriers to learning and assessment (Rose, 2001; Dolan & Hall, 2001). These barriers concern the means of recognition, expression and engagement required by an assessment. This first barrier, the means by which students recognize materials, relates to how a student accesses content. When a test is restricted to a fixed medium, such as paper, a student's ability or inability to work within that medium can confound measurement of their knowledge or skill (Dolan & Hall, 2001). The second barrier concerns a student's level of engagement with material. This barrier relates to a test taker's level of motivation during a test performance and the extent to which students feel engaged with test material. The last barrier concerns the means by which students must communicate their understanding and express responses to test items and tasks (Russell, Hoffman & Higgins, 2009).

When applied, universal design has the potential to assure that all test takers are provided with appropriate opportunities to demonstrate what they know and can do. From the perspective of universal design, barriers to test material are often the result of restrictions placed on the means by which students are expected to recognize and engage with material and express understanding. These restrictions are often unrelated to the target construct(s). Any impact unrelated restrictions have on test performance can have implications for validity and should be addressed during the design process.

Conceptualization and Definition of Universally Designed Assessments (UDA)

There is no universally agreed upon definition of a universally designed assessment (Ketterlin-Geller, 2005). However, assessment is often referenced under the larger umbrella of Universal Design for Learning (UDL). CAST, a leader in the area of Universal Design for Learning, has suggested three principles that can be helpful during assessment development. These principles are to:

- Provide multiple means of engagement (the why of learning or assessing) to support interest, motivation, and persistence,
- Provide multiple means of representation (the what of learning or assessing) such that information and content is presented in different ways and connections are made between them, and
- Provide multiple means of action and expression (the how of learning or assessing) to ensure different ways for students to work with information and content and to demonstrate what they know and can do (CAST, 2016).

CAST has also emphasized the importance of specifically defined constructs. Evaluation of test accessibility and attempts to enhance it should be considered only within the context of very clearly defined constructs. For example, in their critique of an early version of SBAC and PARCC accessibility and accommodation frameworks, staff from CAST (Hall et al. 2013, Hall et al. 2013a) advised that test developers first narrow down and clearly define constructs and then make recommendations for accommodations and accessibility features.

Universal Design for Learning is also broadly defined or referenced in several federal policies. In some cases, the phrase "Universal Design" is used seemingly

interchangeably with UDL. IDEA defines Universal Design as "a concept or philosophy for designing and delivering products and services that are usable by people with the widest possible range of functional capabilities, which include products and services that are directly accessible (without requiring assistive technologies) and products and services that are interoperable with assistive technologies" (IDEA, 2004, Section 611, 16(E)). The No Child Left Behind Act of 2001 (NCLB, 2002) stated that universally designed assessments should "be designed to be valid and accessible with respect to the widest possible range of students, including students with disabilities and students with limited English proficiency" (NCLB Regulation (July 5, 2002), Section 200.2(b)(2)). Referencing the Higher Education Act of 1965 (1998), the Every Student Succeeds Act (ESSA), defines Universal Design as:

> A scientifically valid framework for guiding educational practice that— (A) provides flexibility in the ways information is presented, in the ways students respond or demonstrate knowledge and skills, and in the ways students are engaged; and (B) reduces barriers in instruction, provides appropriate accommodations, supports, and challenges, and maintains high achievement expectations for all students, including students with disabilities and students who are limited English proficient.

Universal Design is also mentioned in the Assistive Technology Act of 2004 (P.L. 108-394 -ATA, 2004) and is similarly defined as:

> A concept or philosophy with the widest possible range of functional capabilities, which include products and services that are directly usable (without requiring assistive technologies) and products and services that are made usable with assistive technologies.

In an effort to clarify what application of Universal Design for Learning to assessment development looks like some have identified specific elements or

characteristics of universally designed assessments. For example, building from a set of general universal design principles developed by the Center for Universal Design (1997), Thompson, Johnstone and Thurlow (2002) described seven elements of universally designed assessments, which included:

(1) *Inclusive assessment population*: Test developers should consider the entire population of students who take a test at the very start of test development. In the case of statewide tests intended for accountability purposes, this includes all students attending public schools, including students with disabilities and other special needs.

(2) *Precisely defined constructs*: Every test item should measure what is intended. To begin, consensus must be reached on how a construct is defined to allow for appropriate evaluation of whether elements of the assessment accurately reflect the intended construct.

(3) *Accessible, non biased items*: Items should be developed in such a way that no subgroup of test takers has an advantage or disadvantage compared to others.

(4) *Amendable to accommodations*: Although the goal of universal design is to improve accessibility for all students, a small population of students may still need accommodations. Test development efforts should aim to facilitate the use of appropriate accommodations and address potential threats to validity and comparability of accommodated scores.

(5) *Simple, clear and intuitive instructions and procedures*: Instructions and procedures should be easy to understand for all students expected to participate on

an assessment, and should be comprehensible regardless of student background, experience or language skills.

(6) *Maximum readability and comprehensibility*: Items should be constructed so that students understand what is being asked of them. When the construct does not relate to reading ability, failure to achieve maximum readability could unintentionally introduce construct irrelevant variance.

(7) *Maximum legibility*: Text, tables, figures, illustrations, and response formats should be clearly designed such that they are easy to decipher by test takers, and irrelevant and potentially distracting physical features are eliminated.

In the area of reading, the National Accessible Reading Assessment Projects (NARAP), which included the Designing Accessible Reading Assessments (DARA), Technology Assisted Reading Assessment (TARA) and the Partnership for Accessible Reading Assessment (PARA), identified five principles for developing more accessible reading tests. These principles are described in a report published in 2009 and included the following:

(1) Reading assessments are accessible to all students in the testing population, including students with disabilities.
(2) Reading assessments are grounded in the definition of reading that is composed of clearly specified constructs, informed by scholarship, supported by empirical evidence and attuned to accessibility concerns.
(3) Reading assessments are developed with accessibility as a goal through rigorous and well documented test design, development and implementation procedures.
(4) Reading assessments reduce the need for accommodations, yet are amenable to accommodations that are needed to make valid inferences about a student's proficiencies.
(5) Reporting of reading assessment results is designed to be transparent to relevant audiences and to encourage valid interpretation and use of these results (Thurlow et al. 2009, p 4).

Ketterlin-Geller (2005) suggested that a universally designed assessment is one that includes "an integrated system with a broad spectrum of possible supports so as to provide the best environment in which to capture student knowledge and skills" (p. 5). From this point of view, designers should consider the compatibility between the test environment and the test taker's characteristics and access needs when developing assessments. Incompatibility between the testing environment and the test taker will limit the degree of test accessibility (Ketterlin-Geller, 2005) and, consequently, the quality of information a test yields about a given test taker. Ketterlin-Geller pointed out that since test takers will have both fixed characteristics (e.g. permanent visual impairment) and fluid characteristics (e.g. current mastery of a certain access skill such as reading or writing), the testing environment must be flexible in structure and format to meet the diverse and changing needs of a testing population. Flexibility can be considered during the design of test format, presentation, delivery, and/or administration (Ketterlin-Geller, 2008).

Rather than outline specific elements of Universal Design for Learning applied to assessment, Ketterlin-Geller (2005; 2008) described the process of developing accessible assessments. In general, the approach described is similar to steps taken in the creation of any test. The distinguishing feature is "the conscious and deliberate consideration of individual needs within the design of the testing environment" (Ketterlin-Geller, 2005, p 11). The characteristics of the entire testing population should drive design decisions regarding procedures, structure and format to ensure that all test takers' access needs are met. For example, during item creation, Ketterlin-Geller suggested that when

determining an item's format, developers allow for flexibility so that users can choose a combination of response modes based on their individuals needs.

Johnstone (2003) compared student performance on a traditionally designed test with performance on an assessment that included features of UDL described by Thompson, Johnstone and Thurlow (2002). This study used a paper-based mathematics assessment that included multiple-choice items from an actual statewide assessment as a control. A second test was created with revised items in accordance with UDL principles. The aim in revising items was to keep constructs constant, but remove construct irrelevant features. Examples of revisions included changes to font size, response formats and timing. Both tests were administered to 231 sixth graders (with form order randomly assigned). This sample included 31 students with specific learning disabilities, 109 English language learners, and 132 who were reading below grade level. Students with disabilities were provided with the accommodations indicated in their IEP on both tests. Out of the 231 students tested, 155 scored significantly higher on the universally designed test. Only 17 scored significantly lower. An analysis on specific subgroups found that all subgroups (e.g. students with disabilities, English language learners and students from various ethnic groups) performed significantly higher on the universally designed test. On a post-test interview of a subsample of 23 students, students generally preferred UDL features.

ACCOMMODATIONS, UNIVERSAL DESIGN FOR LEARNING, AND TECHNOLOGY

The aim of UDL is not to create a "one-size fits all" test, but rather one that offers alternatives to cover the broadest range of needs among the test taker population (Rose & Meyer, 2000). To an extent, many of the needs within the testing population can be anticipated, and appropriate design decisions can be made. However, in some cases additional accommodations may still be necessary. Universal Design will not eliminate the need for accommodations (Thompson, Johnstone, Anderson, & Miller, 2005). The goal is to anticipate common accommodations needed by students, and then design tests to allow for more effective integration into the test format and procedures (Thompson, Johnstone, Anderson, & Miller, 2005), while minimizing or avoiding potential validity impacts.

When accommodations are considered this way, rather than dealing with them as retrofitted test elements, accommodations can be offered as embedded accessibility features that provide alternative means of engaging, interacting and responding to assessment tasks. However, Kettler-Geller (2005) pointed out that in many state assessment systems, accommodations are applied after test conceptualization and development that has been mainly driven by the needs of the general education population. She argued "externally imposed accommodations" can lead to several problems, including insensitivity to individual differences, accommodations that fail to meet the needs of users, the inability to provide multiple accommodations together, and inconsistent assignment and provision of accommodations.

TECHNOLOGY-BASED ASSESSMENT AND ACCOMMODATIONS: OPPORTUNITIES AND CHALLENGES

More extensive integration of technology into assessment design may allow for accommodations to be more feasibly and more cost effectively included in the architecture of assessment systems. The use of technology-based assessments can provide the opportunity to seamlessly build in accommodations on a single interface used by all test takers (Kavanaugh & Russell, 2011). Also, technology-based accommodations and assistive technology can be better standardized and administered more consistently. This limits the amount of construct irrelevant variability introduced by variable conditions sometimes created by human administered accommodations. Other benefits include more efficient test administration, availability of immediate results, better organization of data, and increased authenticity of items (Thompson, Thurlow, & Moore, 2003). Furthermore, technology is being used with growing frequency to provide a range of supports for diverse learners during classroom instruction, allowing students the benefit of independent self-paced access to instructional and testing material (Dolan et al. 2005). Integration of the same technology-based supports on large-scale assessments would lead to better continuity between instruction and statewide assessments (Dolan et al., 2005) and better alignment with student preferences for technology-based testing over traditional paper-and-pencil assessment (Thompson, Thurlow & Moore, 2003).

There are several challenges associated with large-scale use of technology-based assessments. Thompson, Thurlow, Quenemoen and Lehr (2002) provided a summary of these challenges. First, questions still remain about whether students' familiarity with technology and frequency of use puts some students at a disadvantage when technology-

based tests are used (Trotter, 2001). Research suggests that technology-based testing imposes different demands than paper-based tests (e.g. typing, scrolling through multiple screens, recalling information not currently displayed on screen, reading from a screen) (Hollenbeck, Tindal, Harniss, Almond, 1999; Ommerborn & Schuemer, 2001) and may consequently decrease accessibility for test-takers who lack these skills.

Thompson and colleagues (2002) also described the technical challenges that come along with technology-based testing. For example, schools and states need staff with appropriate technical expertise to set up and keep computer-based testing systems running. Also, appropriate infrastructure (e.g. high speed internet, computers or tablets, headphones, keyboards) to support testing systems may require a large initial investment by schools and states. The quality and consistency of equipment and consistency among equipment available to schools may have an impact on the level of standardization that is ultimately achieved. Finally, Thompson and colleagues mentioned the security of online data as an ongoing concern for schools and states interested in technology-based testing.

It is also important to keep in mind that not all technology-based assessments are universally designed (Thompson, Johnstone & Thurlow, 2002). In 1998, Bennett observed, many "computerized tests automate an existing process without re-conceptualizing it to realize the dramatic improvements that the innovation could allow" (p. 3). This observation still applies and includes the failure to use technology to improve accessibility of assessments. A greater number of states have embraced technology based testing in recent years, but have failed to take full advantage of technology in enhancing their assessments. For example, these systems often use separate interfaces for different student groups, rather than integrating multiple accessibility features into one common

interface (e.g. offering a technology-based version for students who receive a read accommodation while all other students use paper and pencil form). The use of separate interfaces can confuse test administrators and test takers, can increase costs, can result in a meaningfully different test experience among test takers (i.e. one that correlates with test performance), and does not reflect principles of UDL.

Moving Towards Large Scale Technology-Based Assessments: PARCC and SBAC Assessment Consortiums

As part of the American Recovery and Reinvestment Act of 2009 (ARRA), a 4.35 billion dollar fund was established for a competitive grant program, known as Race to the Top (RttT). This program was designed to encourage states to promote education reform and innovation. Grant applicants were required to submit proposals for assessment systems that were valid, fair and reliable and "designed to assess the broadest possible range of students, including English learners and students with disabilities" (Department of Education, 2010, p. 18172). Applicants were also required to "use technology to the maximum extent appropriate to develop, administer and score assessments and report assessment results" (DOE, 2010, p.18175). Just over three hundred fifty million dollars were awarded to two multi-state consortia for the development of a new generation of common assessments – the Smarter Balanced Assessment Consortium (SBAC) and the Partnership for Assessment of Readiness for College and Careers (PARCC). Both assessment consortia have developed and piloted computer-based assessments with embedded accessibility tools as well as locally provided accommodations and reference

Universal Design principles in their test design frameworks (PARCC, 2014, Measured Progress & National Center on Educational Outcomes, 2014).

In a summary of findings from the PARCC 2014-15 field tests (PARCC, 2014), evaluators did not report on the use or impact of accessibility features and accommodations, but did report on the operational and student reported experiences with technology-based assessment. In their report, they stated no system-wide technology issues were identified during the pilot test, but noted that many local issues did occur. These local occurrences were described as "an expected result when school districts introduce computer testing for the first time as was the case in most PARCC states" and that issues were "quickly and easily resolved" (PARCC, 2014, p 3). Examples of such issues included needed adjustments to firewall settings or computer settings, students needing help logging in, devices that stopped working, devices that worked slowly, or lost internet connection during testing. They also reported that student survey results and observations by test administrators suggested students found assessments more engaging, easy to use and generally reported a positive experience with testing.

The Smarter Balanced Assessment Consortium (SBAC) also offered a similar technology-based assessment and released results of their field test in October 2014. Again, they did not specifically report on the impact of accessibility features or accommodations, but did report on student reported experiences with the technology-based test interface and technical issues. They reported similar findings as PARCC and noted that the use of technology-based features seemed to positively impact student engagement. They also reported that across the five states for which there was pilot test data available, on average 67% of responding students found the test interface "easy" or

"very easy" to use. The report authors did note that in order to obtain valid results for all students, all students would need to be familiar with and be able to easily navigate the testing interface and tools.

Additional Research on Technology Assisted Accommodations

Currently, research on technology- based accommodations and the use of assistive technology during testing to improve accessibility is limited and represents only a small portion of accommodations research. There is more research available on the comparability of paper and pencil and computer administered tests, sometimes offered as an accommodation itself. Generally the aim of this research did not include the evaluation of computer administration as an accommodation or as a means of improving accessibility. Researchers often did not examine the impact on scores for students with disabilities and other special needs separately (e.g. Choi & Tinker 2002; Russell & Plati 2000). Other studies examined computers as a vehicle to provide a single accommodation, such as read aloud. Some studies have explored the use of a computerized read aloud as one tool for creating more individualized and flexible assessments (Brown & Augustine, 2000; Burk, 1999; Calhoon, Fuchs & Hamlett, 2000; Dolan, Hall, Banerjee, Chun & Strangman 2005; Hollenbeck et al., 2000; HumRRO, 2003; Miranda, Russell & Hoffman, 2004). However, among the research that has examined the impact of computer administration and other types of technology-based accommodations on student scores specifically for students with special needs, very few studies examined individualized packages of accommodations. The following is a

summary of those studies that examine the impact of technology-assisted accommodations assigned according to participants' individual needs.

In a dissertation Ketterlin-Geller (2003) analyzed test data from 187 third grade students attending six schools in the Pacific Northwest to determine whether Universal Design leads to more accurate measurement of mathematics ability. Students were given two mathematics assessments: a computer-based test, and a universally designed, adaptive computer-based test. The universally designed test was developed using guidelines from the National Center on Educational Outcomes (Johnstone, 2003). The tests differed primarily in two ways: (1) computer-based accommodations were available to any student based on scores on basic skills assessments for the universally designed assessment, and; (2) the universally designed assessment was adaptive. For example, a student with poor reading comprehension, determined based on performance on a Maze task, was permitted a computer-based read aloud accommodation. In total, four options were available on the universally designed assessment: standard (no accommodation) (n = 128), simplified language (n = 37), read aloud (n = 14), or read aloud with simplified language (n = 8). The standard testing interface was identical for both computer-based test forms. Items were presented one at a time on the left of the screen, with answer choices listed vertically on the right. All relevant information for a single item was displayed on a single screen in black 18-point font and directions were written in simple language with a read aloud option.

Over a two week period, students completed both tests on two separate occasions in a random, counterbalanced order. To evaluate the effect of accommodations and educational placement on estimates of mathematics ability, a three-way within subjects

ANOVA was conducted. This analysis included test format (computer adaptive and computer-based) as a within subjects factor and accommodation condition (standard, read aloud, simplified, and read aloud with simplified) and educational classification (special education vs. general education) as between subject factors. According to results, students who received an accommodation on the universally designed CAT performed significantly lower than students who had not received an accommodation. Students placed in special education scored significantly lower than general education students. Students performed consistently across testing formats (computer-based and computer adaptive), regardless of accommodation condition or educational classification. Based on these findings and the correlation between student performance on the CAT and other measures of mathematics ability (i.e. Stanford Diagnostic Mathematics test, a paper and pencil test), Ketterlin-Geller concluded that the universally designed testing system may accurately measure students' mathematics ability, but accommodation assignment based on basic skills tests may not be effective. She also noted that there was limited empirical evidence to support a claim that accommodated conditions were equivalent to non-accommodated conditions for the universally designed CAT.

The use of dictation and speech recognition technology was the focus of a study by MacArthur and Cavelier (1999). Using a repeated measures group design, students with (n = 27) and without (n = 10) a learning disability that affected their writing were asked to complete a writing test under the three conditions: (1) handwriting, (2) dictation to a human scribe, and (3) dictation to a computer using speech recognition software. The order in which conditions were completed was randomly assigned. For each condition, students were asked to prepare an essay in response to prompts similar to those

used for statewide assessments. Prior to scoring, researchers typed all handwritten compositions, including errors. Significant main effects were detected for both disability status and testing condition (handwritten, human scribe, and speech recognition software), but interaction between disability status and testing condition was not significant. Follow up t-tests found that students with learning disabilities scored significantly higher when provided a human scribe or speech recognition software than when handwriting their own responses. Students with learning disabilities scored the highest on writing tasks when provided a human scribe. There were no significant differences observed in test scores across conditions for students without learning disabilities.

Researchers concluded that dictation assisted students with writing-related learning disabilities to produce better essays, exhibiting their best performance with a human scribe. Researchers hypothesized that the provision of a human scribe freed students from thinking about mechanics and to concentrate on content and organization. Speech recognition software eliminated student concerns about spelling and handwriting, but students still needed to be mindful of punctuation and had the added concern of speaking clearly and monitoring their writing for errors. Overall, their findings suggested that dictation was a valid accommodation for students with disabilities, and did not appear to provide an extra advantage over peers without disabilities. However, improvement to speech recognition software may be necessary to enhance its utility as an accommodation.

Dolan, Hall, Banerjee, Chun and Strangman (2005) examined the impact of a technology-based read aloud accommodation on student performance on a multiple-

choice U.S. history and civics test. Researchers compared student performance on two test forms for each subject under two conditions: (1) using a computer-based testing (CBT) system with a text-to-speech read aloud tool, and; (2) a paper and pencil version. Participants included nine 11th and 12th grade students recommended by teachers. All students received special education services, but were also partially or fully included in general education classes. Under both conditions students were presented one item at a time and asked to respond directly on either a test booklet or computer screen, eliminating the need for a separate answer sheet. During the CBT, students were also provided a computer-assisted read aloud. Test designers of the CBT aimed to allow test takers the same flexibility permitted on the paper and pencil version. That is, students were permitted to complete test items in any order they chose, were able to skip items, review and change answers, and reread any text (i.e. items, answers, passages and directions).

Dolan and colleagues found that students performed slightly better on the CBT. The difference was associated with an effect size of .49, but was not statistically significant. Students did perform significantly higher on the CBT than on paper and pencil tests for items associated with longer passages (more than 100 words; effect size = .60), but performed better on paper and pencil tests for items associated with shorter passages (100 words or less). This latter difference in performance represented an effect size of .29 and was not statistically significant. Researchers speculated that shorter passages were less challenging and participants did not require read aloud support as much for these items. Overall, these findings were consistent with other research that had found computer-based read aloud effective for students with disabilities.

The authors did note several limitations. Students lacked extensive experience with the computer interface and read aloud tool, a condition that is not ideal for high stakes testing. They also cautioned that since students were told their scores would not affect their grade in any way, test taker motivation might have been lower than operational test conditions. With more training and performance incentives, students may have yielded greater benefit from the computer-based testing interface and computer-assisted read aloud for both short and long passages. Finally, they noted a small sample size may have resulted in insufficient power to detect smaller score differences. They recommended future studies including larger sample sizes to further investigate the impact of computer-assisted read aloud.

Kopriva, Emick, Hipolito-Delgado and Cameron (2007) explored whether individualized computer-based accommodations were more effective than incomplete packages or no accommodations for third and fourth grade English language learners (ELL). A sample of 272 ELLs was randomly assigned to receive no test accommodations, a picture dictionary, a single accommodation (i.e. a bilingual glossary, oral reading of test items in English) or some combination of computer-based accommodations while completing a mathematics assessment. After testing, students were grouped according to whether they received accommodations that matched all teacher recommendations (recommended), matched some recommendations (incomplete), or received none of the recommended accommodations (none). Researchers found that students who received all the recommended accommodations performed significantly better than students who received incomplete packages or no accommodations. No difference was found between students who received incomplete

packages and those receiving no accommodations. Their findings highlight the impact incomplete assignment can have on student scores and the importance of accommodation packages that address all test taker access needs.

Research on NimbleTools

NimbleTools, a universally designed computer-based test delivery system, was one example of an attempt to capitalize on technology to improve test accessibility. Given this dissertation's focus on the impact of accommodations provided via NimbleTools on the validity of student scores, the following section summarizes available research on this test delivery system.

According to its developers, "NimbleTools embraces principles of universal design, incorporating a variety of accessibility tools, dynamically tailored for each student and which allow students to use those tools as desired while taking a test. Flexibly tailoring the availability of accessibility tools enables a test to be delivered to students across a testing program using a single computer-based test delivery system" (Russell, Hoffman & Higgins, 2009, p. 1). The aim of this system was to improve test validity by addressing accessibility issues up-front during the design process, and include flexible accommodation options to provide better support for students with more extreme access needs. With each added accessibility feature, developers redesigned and rebuilt the system, rather than simply adding supports to an existing version of the system (Russell, Hoffman & Higgins, 2009). This ensured that each tool could interact with each other and permitted the provision of multiple accommodations to a single user (Russell, Hoffman & Higgins, 2009).

NimbleTools was not a test itself, but rather a delivery system on which testing programs could insert their own test items. Embedded tools were selected and activated for individual students. Test takers receiving supports could use activated tools as needed. This approach might allow the specific needs of individual students to be more easily and effectively met. A full description of accessibility tools offered as accommodations during the 2009 NECAP administration is included in chapter three.

Early development research provided promising results for addressing accessibility barriers for students with special needs on state assessments in a cost effective and standardized manner. According to research described by Nimble Assessment Systems (2009), during initial pilot testing of NimbleTools in 2007, teachers found the control interface used to select accommodations for students easy to use. In a later pilot test, a sample of 31 students with various special needs was administered two short 10-item tests, containing both multiple-choice and open-ended items from the New Hampshire tenth grade mathematics test. Students were asked to complete a paper version of one form, and a second form using NimbleTools (referred to as "flexible test delivery system"). Overall, students scored significantly higher on multiple-choice items when taking the test on a computer (26% correct versus 35% correct; $p = .03$). An additional caveat was that when performing on paper, students performed at the chance level while computer performance was well above chance level. Student feedback collected from a focus group and surveys indicates that students had a positive experience using NimbleTools and preferred using it to the paper-based test.

A 2008 pilot study conducted by Nimble Assessment Systems and the National Center for Educational Outcomes found similar results. This study examined the use of

NimbleTools to deliver the Florida Comprehensive Assessment Test (FCAT) with a subset of accommodations to sixth and ninth grade students with and without special access needs. Accessibility tools that were offered via NimbleTools included read aloud of text (via digital recordings of human voice), and alternative contrast (ability to change background and font colors), magnification and auditory calming. Using a repeated measure design, students with and without special needs were administered two sets of items matched by content (math or science) and difficulty followed by a brief survey. After a brief training on how to use NimbleTools, all students completed the first form without any accessibility tools. Students with special needs completed the second form with accessibility tools assigned by their teachers, while general education students were permitted access to all accessibility features.

Based on the performance of 181 students who completed both forms (146 = Teacher Assigned Accommodations, 35 = No Teacher Assigned Accommodations), it appeared that NimbleTools had a differential effect on the performance of students who had an identified need for an accommodation compared with students who had no identified need. Students also reported that in general NimbleTools was easy to use, and expressed a preference for using NimbleTools on future assessments.

The authors of this study cite several design flaws to consider along with these results. First, accommodations were always provided to students on the second test form, making it impossible to separate the impact of accessibility tools on performance from other factors, such as fatigue or test difficulty. In some cases, teachers were unable to provide a break between test administration due to time constraints and availability of computer labs. Fatigue may have impacted test performance for these students in

particular. Finally, since testing took place very early in the school year, some teachers felt they were not given enough time to get to know students to allow more informed decisions about accommodation assignment.

There has also been research focused on the refinement of accessibility tools. Russell, Kavanaugh, Masters, Higgins and Hoffman (2009) described a randomized trial that compared the effect of two signing accommodations offered by NimbleTools (a recorded human and a signing avatar) on student performance and students' attitudes about using each accommodation. Ninety-six student participants ranging from grade 8 to 12 who communicated in American Sign Language were asked to complete two test forms matched by item difficulty made up of released items from the Grade 8 National Assessment of Educational Progress in mathematics. Using NimbleTools, students either viewed the first form using the recorded human or avatar and the second form using the alternative signing presentation. Students were then asked to complete a brief survey. In general students reported that both signing tools were easy to use and understand. Students also expressed a strong preference for completing future tests on a computer, and using either signing tool rather than a DVD. Finally, although most students preferred the recorded human to the signing avatar, this preference did not seem to affect actual performance. There was no difference in time required to complete items or performance on items based on the signing accommodation used.

During focus groups conducted later, students noted a few elements of the signing tools that could be improved. Some participants found the highlighting of text during signing videos to be distracting. Participants felt that this feature drew their focus away from the video and to the text. Also, for this study, signing videos played automatically

when a new item was displayed on the screen. Several students preferred to read items first and play videos as needed instead. Though it was already possible to implement the signing tool this way, these student comments highlighted the value of the option and the importance of offering it during future studies and operational testing. Some students found the avatar's movements not as natural looking as the human. A small number of students stated that the avatar's "jerky movements" were distracting, but most felt this was not problematic. Finally, several students found the tutorial presented before testing too difficult. This tutorial, which provided an introduction to test taking with NimbleTools, used text based explanations. Many participants with below grade level reading skills expressed a preference for a tutorial presented using ASL.

Summary of Research on Universal Design for Assessment

It is important to keep in mind that universal design for assessment is still developing. Test designers and consumers are still determining how best to improve access for the wide range of learners who participate on large scale assessments (Johnstone, Anderson & Thompson, 2006; Thompson & Thurlow, 2002). Given the high stakes often attached to statewide tests and the increased use of technology-based assessments, rapid progress in this area should be encouraged (Thompson & Thurlow, 2002).

Although technology has the potential to facilitate greater implementation of UDL principles to address accessibility concerns during large scale testing, research on technology-based accommodations and the use of assistive technology is relatively limited. Initial cost requirements associated with technology- based assessment systems

may pose an obstacle for some schools and states. These same systems may offer savings further down the line, particularly in the area of scoring. Preliminary research on NimbleTools, one example of a universally designed computer- based test delivery system, suggests this is a promising approach to addressing accessibility issues for unique learners.

CHAPTER SUMMARY

Although research examining the impact of test accommodations on validity has grown, results have been mixed and research can be difficult to carry out. These findings may be due in part to varying operational definitions of accommodations among studies and the extent to which assigned accommodations address participants' access needs. Among the available research, science assessments have been less frequently studied. This may be reflective of the relatively low priority given to the subject of science in standards-based reform of the late 1990s and into the 21st century. There has also been less research on accommodations used among high school students, with middle school and elementary populations studied more frequently. This may be due to the fact that accommodations are assigned to middle school and elementary school students in higher frequency on statewide tests (Thurlow, Altman, Cormier & Moen, 2008; Altman, Thurlow & Vang, 2010). Finally, less research has been conducted on Universal Design for Learning applied to assessment and the impact of technology-based accommodations. This dissertation attempts to address these gaps and focuses on the impact of accommodations provided through a universally designed, computer-based test delivery system on the validity of scores on a high school statewide science assessment. The next

chapter will revisit the research questions posed in this dissertation and describe the data and methods used in greater detail. This includes population and sample characteristics, instrumentation, procedures and analyses.

CHAPTER 3: METHODOLOGY

This dissertation examines two approaches to addressing accessibility for students with special needs: a paper-based assessment with individual accommodations and a universally designed computer-based assessment with embedded accessibility supports. If these approaches are successful in addressing accessibility needs, construct irrelevant variance related to accessibility will be minimized or eliminated. The resulting scores should exhibit similar psychometric properties across accommodated and non-accommodated conditions. To guide this investigation, the following research questions are considered:

- Is the underlying factor structure consistent for scores gathered under accommodated and non-accommodated conditions?
- Do items function similarly under accommodated and non-accommodated conditions? Specifically, holding ability constant, are item difficulty and discrimination equivalent for accommodated students and non-accommodated students?
- If differential item functioning is exhibited, do patterns of DIF and item characteristics suggest that accommodations or use of accessibility supports may be related to DIF?

To answer these questions related to validity, a secondary analysis of assessment data collected during the 2009 administration of the 11th grade NECAP science assessment was conducted. Schools were given the option of administering a paper and pencil assessment with traditionally offered accommodations or using a universally

designed computer-based test delivery system that included integrated accessibility supports.

Results from two types of analyses are presented in the next chapter: differential item functioning (DIF) and confirmatory factor analysis (CFA). DIF was used to explore item functioning, comparing item difficulty and discrimination under accommodated and non-accommodated test conditions. A large number of items exhibiting DIF, favoring one test condition over another could signal differences in overall test functioning. Items exhibiting DIF were examined within the context of accommodations and the NimbleTools test interface to determine whether DIF appeared related to accommodated conditions. Similarly, CFA was used to examine the consistency of underlying factor structure as evidence of potential constructs measured across test conditions.

POPULATION AND SAMPLE

Test data was collected from students attending high schools in New Hampshire, Vermont and Rhode Island who participated in the 11th grade New England Common Assessment Program (NECAP) science assessment in spring 2009. Within this consortium of states, "all" students were eligible for accommodations. Unlike many states, an IEP or 504 plan was not required to receive test accommodations on statewide assessments. Therefore, not all students assigned test accommodations had a formal disability classification, but should have had some otherwise identified need. There was no record of specific student need(s) beyond the assignment of a given accommodation available in this particular data set. Specific state policies for allowable accommodations

and decision-making as well as test condition group definitions are described in greater detail later in this chapter.

Student data used for this analysis were collected for a larger study examining the feasibility, effect and capacity to deliver state achievement tests using NimbleTools (Nimble Assessment Systems, Inc., 2009), a computer-based test delivery system with embedded test accommodations and accessibility features, designed using principles of Universal Design for Learning. A brief description of NimbleTools can be found below. Additional information about NimbleTools is also included in the previous chapter. Public schools, who are required to participate in annual statewide testing, had the option of administering the 11th grade NECAP science assessment using NimbleTools to some students. This option was not available for students completing the NECAP science test in other grades.

All schools in this study volunteered to use NimbleTools, introducing the possibility of self-selection bias at the school level, a noted limitation of this study. School level data was not available for this analysis so this could not be fully explored. In general, it is known that schools opted to use or not use NimbleTools for a variety of reasons. Among those who did not elect to use NimbleTools, school staff may not have felt prepared to administer a technology-based assessment. Some schools may have been unaware of the option. In other cases, school staff did not see this option as a benefit to their students. Among schools that did opt to use NimbleTools, some may have used it for all students requiring accommodations supported by NimbleTools while others only used it for some students.

Almost all NimbleTools users received one or more of the embedded test accommodations/accessibility features. The remaining majority of 11th grade students completed the NECAP science test using a paper and pencil form with or without traditionally offered test accommodations. In total 32,651 students participated in testing. Based on state assigned test status codes, a number of students were excluded from analysis, including those who tested incomplete (n = 742), tested with non-standard accommodations (n = 14) or had a test that was somehow damaged and therefore was not scored (n = 3). Students were considered as "tested incomplete" if they did not attempt all test sessions. Students who tested with a non-standard accommodation are those who received some modification to test procedures beyond the standard accommodations believed to have interfered with the target construct. An example of a non-standard accommodation for the NECAP science assessment is the use of a scientific or graphing calculator during the third testing session. To avoid possible confounding of a language disadvantage, students identified as English language learners (n = 449) were also excluded. The remaining cases (n = 31,463) were analyzed and sampled for analysis. Approximately one-third of schools across all three states volunteered to use NimbleTools to test at least some of their students (n = 656). The remaining schools administered only the paper-based form to students (n = 30,807).

This dissertation examined item functioning and underlying structure for three groups of students within the testing population: (1) students who completed the science assessment using NimbleTools and were assigned at least one embedded support (Accommodations - Nimble; n = 656); (2) students who completed the paper-based assessment and were assigned at least one accommodation traditionally offered by states

(Accommodations - Paper; n = 2,343); and (3) students who completed the paper-based assessment without any accommodations (No Accommodations; n = 28,464). To achieve similar sample sizes among groups, a random sample of 2,000 students from this latter group was selected for analyses. Table 3.1 summarizes the demographic characteristics among sampled students. There seemed to be differences between students assigned to use NimbleTools and those accommodated with a paper form. Specifically, the percentage of students with an IEP assigned to NimbleTools was noticeably larger than those assigned accommodations with a paper form. This was also the case in the original population data.

Table 3.1. Demographic Characteristics of Sampled Students by Test Condition

	No Accommodations (n = 2,000)	Accommodations - Paper (n = 2,343)	Accommodations - Nimble (n = 656)	Total (n = 4,999)
Gender (%)				
Male	48.6	58.0	64.3	55.0
Female	51.4	42.0	35.6	44.9
IEP Status (%)				
Yes	6.3	64.4	76.7	42.8
No	93.7	35.6	23.3	57.2

NECAP ACCOMMODATION ASSIGNMENT POLICY

All accommodations (provided via NimbleTools and traditional means) were assigned based on each student's individual need(s), teacher recommendations and state policies regarding acceptable test accommodations. As described in the *New England Common Assessment Program Accommodations Guide* (New Hampshire Department of Education, Rhode Island Department of Education, & Vermont Department of Education, 2009), accommodation decisions were made by a team of individuals who are responsible

for planning a student's academic program and by a student's parent(s) or guardian. Typically, this was a student's existing 504 or IEP team. For general education students, schools were encouraged to convene Student Support or Child Study Teams. A typical team might include a student's teachers, parent(s) or guardian and the student when appropriate. Schools were responsible for determining a specific process for making accommodation decisions, but states did advise that team members select "the least intrusive accommodations possible to meet the needs of the student while allowing the maximum level of independence possible for that student" (New Hampshire Department of Education, Rhode Island Department of Education, & Vermont Department of Education, 2009, p.6).

Although experimental designs are generally preferable in accommodations research, random assignment to test conditions was not possible in this case of operational testing. Thompson, Blount and Thurlow (2002) point out possible benefits of non-experimental accommodations research, including large sample sizes and real world test conditions. The benefit of large sample sizes in this case permits greater generalizability and increased power (i.e. the ability to detect an effect when one in fact exists). This is especially important when CFA and DIF are used since they require relatively large samples to achieve stable estimates (Thurlow et al. 2000). The second benefit suggests that operational testing may offer conditions in which students are likely to put forth their best efforts. The resulting scores are less likely to be artificially depressed due to lack of motivation. However, there is also the danger that non-experimental data can be more easily confounded by other variables (Tindal, 1998). For example, in this case, schools volunteered to use NimbleTools. These schools may differ

in meaningful ways (e.g. staff attitudes toward technology or available school infrastructure to support computer-based administration) that may directly and/or indirectly impact student performance. Unfortunately school level data was not available for this analysis and it was not possible to explore the impact of between school differences.

Another debate in accommodations research has been disentangling the effects of individual accommodations and whether or not researchers should attempt to do so. In operational testing, test takers often need multiple accommodations to permit full access to test items (Thompson, Blount, & Thurlow, 2002; Kim, Schneider & Siskind, 2009). Students may fail to do well if they do not receive all the necessary supports. Therefore, attempting to isolate the effect of single accommodations by assigning participants only one type of accommodation for research may not be authentic (Kim, Schneider & Siskind, 2009) and fail to demonstrate an effect. Also the needs of individual students will vary and the same solution (i.e. test accommodation) is not necessarily appropriate for all students. Thompson, Blount and Thurlow (2002) recommend considering the combined impact of multiple accommodations instead. Following these recommendations, this study explores the impact across all combinations of accommodations.

Likewise, attempts to improve access and, in turn, test validity using NimbleTools involve both the provision of technology-based accommodations and its universally designed interface. Providing test takers with only one of these conditions (either technology-based accommodations or a universally designed testing interface) may fail to address the full range of accessibility needs among test takers. The approach of

examining the impact of each of these design features in isolation would also fail to mimic the operational test condition and possibly fail to yield an effect. This dissertation considered the collective impact of the NimbleTools' universally designed interface and embedded accommodations on the validity of scores, without attempting to isolate the individual effects of either.

Finally, accommodations are assigned at the test level, but students may or may not make use of all accommodations for all items. Although most research assumes accommodations are being used consistently throughout testing, the actual accommodation applied may differ from item to item. A student may only use their read aloud for some, but not all items, for example. However, detailed data on accommodation use by item is often not available. This was also the case for this study. Actual use of accommodations by item could not be explored.

Table 3.2 lists standard test accommodations available during the 2009 science NECAP (New Hampshire Department of Education, Rhode Island Department of Education, & Vermont Department of Education, 2009).

Table 3.2. NECAP Test Accommodations

A. Alternative Settings

- Administer the test individually in a separate location
- Administer the test to a small group in a separate location
- Administer the test in locations with minimal distractions (e.g., study carrel or different room from rest of class)
- Preferential seating (e.g. front of room)
- Provide special acoustics
- Provide special lighting or furniture
- Administer the test with special education personnel
- Administer the test with other school personnel known to the student
- Administer the test with school personnel at a non-school setting

B. Scheduling and Timing

- Administer the test at the time of day that takes into account the student's medical needs or learning style
- Allow short supervised breaks during testing
- Allow extended time, beyond recommended time until in the administrator's judgment the student can no longer sustain the activity

C. Presentation Formats

- Braille
- Large-print version
- Sign directions to student
- Test and directions read aloud to student
- Student reads test and directions aloud to self
- Translate directions into other language
- Underlining key information in directions
- Visual magnification devices
- Reduction of visual print by blocking or other techniques
- Acetate shield
- Auditory amplification device or noise buffers
- Word-to-word translation dictionary, non-electronic with no definitions
- Abacus use for student with severe visual impairment or blindness

D. Response Formats

- School personnel transcribes student response exactly as written, indicated or dictated into the Student Answer Booklet
- Student writes using word processor, typewriter, or computer (spell and grammar checks must be turned off)
- Student hand writes responses on separate paper
- Student writes using brailler
- Student indicates responses to multiple-choice items
- Student dictates constructed responses or observations to school personnel
- Student dictates constructed responses or observations using assistive technology

INSTRUMENTATION

11th Grade NECAP Science Assessment

According to the test administrator manual for the 11th grade NECAP science assessment (New Hampshire Department of Education, Rhode Island Department of Education, & Vermont Department of Education, 2009), assessment targets were developed and adopted by each of the departments of education in the NECAP consortium. Four domains were measured: (1) science process skills (28%), (2) earth space science (24%), (3) life science (24%), and (4) physical science (24%).

Four forms of the assessment were administered. Each consisted of 65 items. This included 14 constructed response items and 51 multiple-choice items. For constructed response items, students were presented with a prompt that included text and for some items, a visual stimulus (e.g. figures, graphs, maps, pictures etc.). Multiple-choice items generally included an item stem and four response choices for students to select from. Among these items were common items, equating items, and embedded field test items. Common items, including 33 multiple-choice items, appeared on every form of the test and were used to determine a student's test score. Only common items were analyzed during this study.

Overall student results were reported as a scaled score and performance level. Scaled scores for the grade 11 NECAP science assessment ranged from 1100 through 1180. Performance was also described in terms of performance level: (1) substantially below proficient, (2) partially proficient, (3) proficient and (4) proficient with distinction.

Paper and Pencil Form

According to the test administrator manual for the 11th grade NECAP science test (New Hampshire Department of Education, Rhode Island Department of Education, & Vermont Department of Education, 2009), "the three NECAP states are equally committed to supporting the inclusion of all students in assessment by using elements of universal design in the NECAP tests" (p. 3). Reflective of this commitment, efforts were made to integrate principles of UDL during development of items and to other design features to the extent possible for the paper and pencil form. During development, among other criteria, items were evaluated by considering the following questions:

- Is the item language clear and grade appropriate?
- Is the item language accurate (syntax, grammar, and conventions)?
- Is there an appropriate use of simplified language? (Is language that interferes with the assessment target avoided)?
- Are charts, tables, and diagrams easy to read and understandable?
- Are charts, tables, and diagrams necessary to the item?
- Are instructions easy to follow?
- Is the item amenable to accommodations - read aloud, signed, or braille? (New Hampshire Department of Education, Rhode Island Department of Education, & Vermont Department of Education, 2009, p.8).

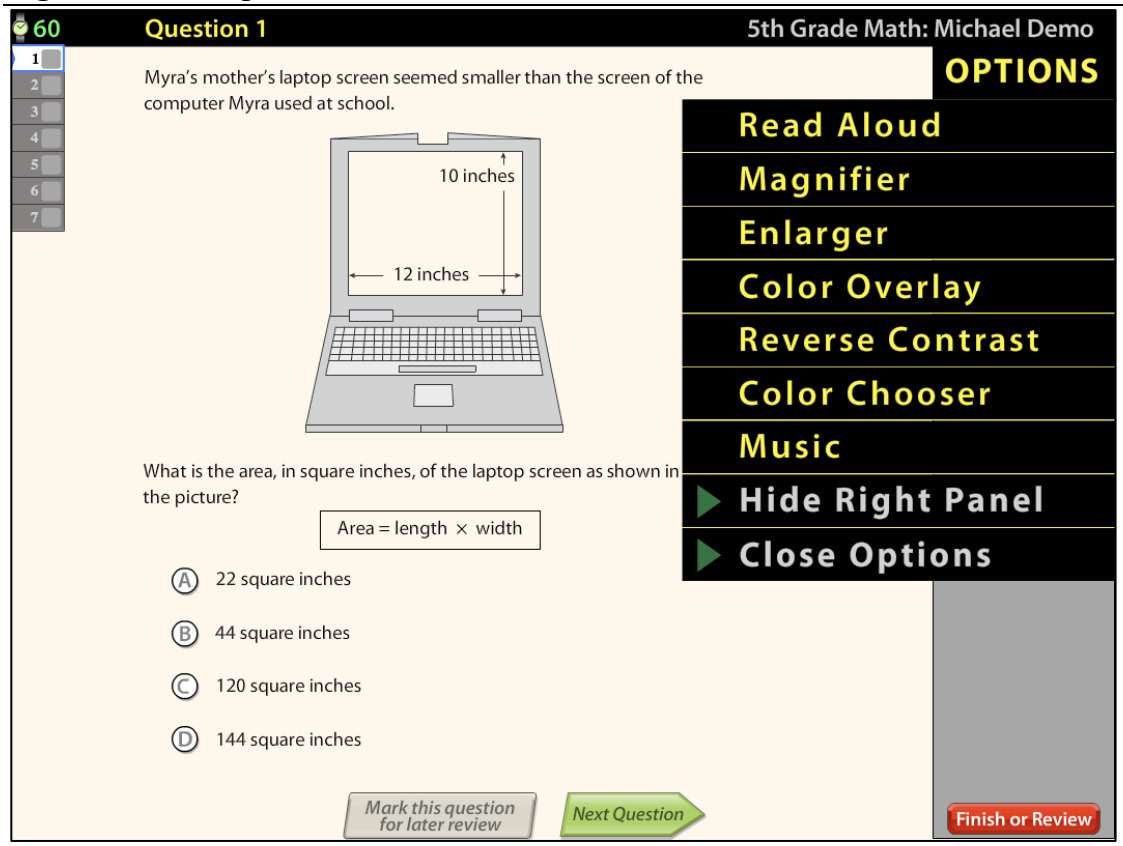
Each student completing the paper and pencil form of the science test received scratch paper, a test booklet, answer booklet, and science reference sheets. In some cases, additional forms of the paper and pencil test (e.g. braille, large print) were necessary to accommodate students with certain special needs (e.g. low or no vision). Students may have also been assigned several other accommodations, such as read aloud, underlining of key information in directions, use of visual magnification devices or reduction of visual print by blocking or other techniques. Table 3.2 summarizes the standard accommodations assigned to students who completed testing with a paper form and also NimbleTools.

Computer-Based Form: Nimble Tools Testing Interface

As described in the previous chapter, NimbleTools was a "universally designed application" with embedded accommodation tools that could be activated or deactivated for each individual student, "creating a customized test delivery interface that meets the specific needs of each student" (Nimble Assessment Systems, 2008). NimbleTools was not a test itself, but rather a delivery system on which testing programs could insert their own multiple-choice and constructed response items.

According to Nimble Assessment Systems (2008a), the standard NimbleTools delivery interface was divided into five sections (See Figure 3.1). Items (along with answer choices) were presented one at a time and located in the center of screen. Most, students responded to multiple-choice items by selecting a radio button located next to their desired answer choice. Located at the top of the screen was the name of the test, current question number and current user name in large, high contrast lettering. Located on the bottom of the screen were basic navigation tools ("next item," "previous item" and "mark item for review" buttons). To the left of the item were "status" buttons. These indicated the item number and whether or not each item within that particular section of the test had been answered. Also displayed here was whether a student had marked an item for review. To the right of the current item was the "options panel." Here students could access accommodations tools and other features, such as a calculator, formula sheet or scratch pad. The listed accommodations in the options panel should have reflected the customized accessibility profile for each individual user. These were configured ahead of testing by test administrators.

Figure 3.1. Example of NimbleTools Interface



At the time of the 2009 administration, NimbleTools offered 18 access and test accommodation tools. However, for the NECAP science assessment, only eight accessibility features were available for assignment. The following are descriptions of each accessibility feature according to Nimble Assessment Systems (2009):

- **Read Aloud:** NimbleTools links pre-recorded human voice recordings to test items presented to students. Students benefit by listening to a fully approved, standardized human voice, assuring correct pronunciation of words, symbols, and equations. NimbleTools empowers students to decide when they want to hear the text read to them, and allows them to play sound clips repeatedly. All buttons and directions have human-read sound clips associated with them. The Low-Vision version describes graphics and diagrams.
- **Auditory Calming (Background Music)**: For students who focus better when receiving auditory input, music or sounds can be provided during testing. These sounds are embedded into the system, so no extra hardware is needed, and there are no concerns about monitoring the content. The player is simple to use, without distracting visuals of many commercial computer MP3 players.

- **Magnifier:** The Magnifier Tool allows students to enlarge the entire test interface. Students have control over when and where they use this tool. The tool options are shown in enlarged, high contrast text. This should only be assigned to students who need the entire test enlarged throughout the test, as it can be disorienting for students who are not used to working this way."
- **Magnifying Glass:** Enlarge any part of the test by using the Magnifying Glass Tool. This tool is intended to be used occasionally by some students who may have difficulty seeing/reading very small details. It is included by default for all students using NimbleTools.
- **Color Overlay:** Students can choose from a variety of color tints which are placed over the questions and directions of the test. Many students find their reading accuracy and speed increases with the use of color overlays.
- **Reverse Contrast:** Students can choose to reverse the colors for the entire test interface. You also have the option of adding a color tint to the question text using the Color Overlay Tool (see above), which is automatically included when you choose Reverse Contrast.
- **Color Chooser**: Students can change the font and background colors for the test questions and direction. Students pick the font and background color combinations from a palette of colors proven to help students. This differs from Color Overlay in that only the text and background colors change. Lines and graphics are not affected by the color changes.
- **Custom & Answer Masking**: A common technique for focusing a student's attention on a specific part of a test item is provided by the Masking Tools. Two masking tools are currently available: Answer Masking and Custom Masking. Answer Masking hides the answers until students have an opportunity to solve the problem and then allows students to reveal answer choices individually or all at once. Custom Masking allows students to create and place 'sheets' on top of any part of the test question, masking those parts of the question they don't want to focus on. Both Masking options increase students' focus on the test question by temporarily hiding all other test elements. Masking can be turned on and off at the student's discretion.

All students also had the option of receiving accommodations related to changes in setting and timing including: secure and supervised breaks and extended time. Nimble Assessment Systems (2009) provided the following description of each:

- **Secure and Supervised Breaks**: Students are allowed to log out of a test session, take a break, then continue with the same session by logging back in. Answers will be saved and restored when they return. This should only be assigned to students who specifically require breaks, as their test session will require more careful supervision (to ensure test security)."
- **Extended Time**: "All students participating will be allowed extended time when taking the tests.

For the 11th grade NECAP science test, both constructed response and multiple-choice items were displayed on the computer monitor for NimbleTools users, but only responses to multiple-choice items were entered directly on the computer. Test administration procedures established by the NECAP states required test takers to hand write responses to constructed responses using paper and pencil answer booklets. Students were able to use access tools related to item presentation for these items (e.g. magnifier, read aloud, colored overlay). Since the experience of responding to constructed responses directly through NimbleTools was not permitted during this administration, access barriers may have still been present for some students. Therefore, this dissertation considered student response data for the 33 common multiple-choice items only.

PROCEDURES

All schools within the NECAP consortium were required to administer the 11th grade science assessment during the testing window of May 11-28, 2009. The test was generally administered across three test sessions over three days, with approximately one and half hours of testing each day. With the exception of make-up sessions, test sections were administered to all students in the same order. The order in which test items were presented within each section varied across students in an effort to deter cheating. A "script" of material to be read aloud to students (i.e. test directions) during test administration was provided to test proctors for each session to ensure consistent and accurate test administration.

All students received scratch paper, a test booklet, an answer booklet, and science reference sheets, which were collected at the end of each test session. Both test formats (paper and pencil and computer-based) included identical items, but the exact manner in which items were presented was not uniform. Items were presented one at a time for NimbleTools users while multiple items were included on a single page on the paper form. Sessions one and two of testing included multiple-choice items (scored as either correct or incorrect) and constructed-response items, which required students to respond using words, pictures, diagrams, charts, or tables. The third test session included short-answer items and constructed-response items, which required application of "inquiry skills to a scientific situation" and a response "using words, pictures, diagrams, charts, or tables to show their thinking and explain their response" (New Hampshire Department of Education, Rhode Island Department of Elementary and Secondary Education, & Vermont Department of Education 2009, p. 1).

Testing was not strictly timed and students were permitted an additional 45 minutes for sessions one and two as long they worked productively. Students who needed additional time for session three had to have been assigned an extended time accommodation prior to testing. Table 3.3 presents the expected and required scheduled time for each test session.

Table 3.3. Breakdown of Testing Time Across Sessions

Test Session	Test Activity	Expected Completion Time (Minutes)	Scheduled Time (Minutes)
General Instructions	Completing Student Information	5	5
Session 1	Directions	5	5
	Testing (25 multiple-choice & 3 constructed response)	45	90
Session 2	Directions	5	5
	Testing (26 multiple-choice & 3 constructed response)	45	90
Session 3	Directions	5	5
	Testing (8 questions)	55	55

(Adapted from table presented on p. 6 of *Test Administrator Manual – Grade 11 Science*, New Hampshire Department of Education, Rhode Island Department of Elementary and Secondary Education, & Vermont Department of Education, 2009)

Paper and Pencil Administrations

For paper and pencil administration, each student received scratch paper, a test booklet, an answer booklet, and science reference sheets. Students in this testing condition were asked to record their answers to all items in answer booklets (unless receiving an accommodation that dictated otherwise). Students were allowed to answer items in any order they preferred within a testing session, but were not permitted to answer items from other test sessions. All test materials were collected at the end of the last session for that day. Any accommodations used during any test session were recorded on each student's answer booklet.

NimbleTools Administration

All schools with students participating using NimbleTools received a memo, CD and NimbleTools manual. Prior to actual testing, students completed an interactive orientation to learn about the test interface and tools. During this orientation, students

were given an opportunity to practice with accessibility features. Students were also strongly encouraged to perform two practice tests. Generally, student feedback collected during this process, teacher recommendations and state policies were used to determine which accommodations would be assigned during actual testing.

Students were then pre-registered with their own "accessibility profile" for NimbleTools to participate with accommodations. This accessibility profile specified which accommodations each test taker would have access to during testing. Each student was then assigned three unique "ticket" numbers for each of the three test sessions. All students who participated with NimbleTools were registered for at least one accommodation. Any accommodations used during any test session were recorded on each student's answer booklet.

It was highly recommended that students using NimbleTools take the test in a separate location and be provided with extended time and individual proctoring accommodations because of timing issues and the potential distraction to test takers who were not using NimbleTools. Since this was not always possible, two scripts were provided to test proctors, one for test administration in a separate location and another for "mixed group" settings.

During the operational test, test proctors were responsible for logging students in and out of NimbleTools. Once logged in, students were asked to enter the litho code found on the bottom right hand corner of their assigned answer booklet. This was to ensure each student was administered the proper form. NimbleTools users also received scratch paper, a test booklet, and answer booklet. Online versions of science reference sheets were available on NimbleTools, but students could also request paper versions.

Although all items were presented online to students, NimbleTools users answered only multiple-choice items directly on the computer and answered constructed response items in separate answer booklets (unless receiving an accommodation that dictated otherwise). Students were allowed to answer items in any order they preferred within a test session, but were not permitted to answer items from another test session. When students completed a test session, they were shown a summary of their answers and asked to confirm they wanted to exit. The test proctor then logged the student out of the system. All other test materials were collected at the end of each test session.

ANALYSES

It was hypothesized that NimbleTools would address accessibility issues comprehensively and effectively for students with diverse needs. It was expected that this approach would limit or eliminate construct irrelevant variance related to accessibility barriers for those with special access needs. These conditions would allow the test to function as it does for students who did not have unique access needs and did not receive accommodations. An expected indicator of this outcome was that psychometric properties of scores collected using NimbleTools would mimic those collected under the non-accommodated condition. To determine the extent to which this held true, two sets of analyses were conducted: confirmatory factor analysis (CFA) and differential item functioning (DIF). The aim of the first analysis was to provide information regarding the underlying construct or dimensions being measured under different testing conditions to determine the extent to which a similar construct was measured across the three target groups (No Accommodations, Accommodations - Paper,

and Accommodations - Nimble). The second analysis also considered each of the three target groups to determine whether items function consistently across accommodated and non-accommodated conditions. Results of these analyses are presented in the next chapter.

Initial Item Analysis and Reliability

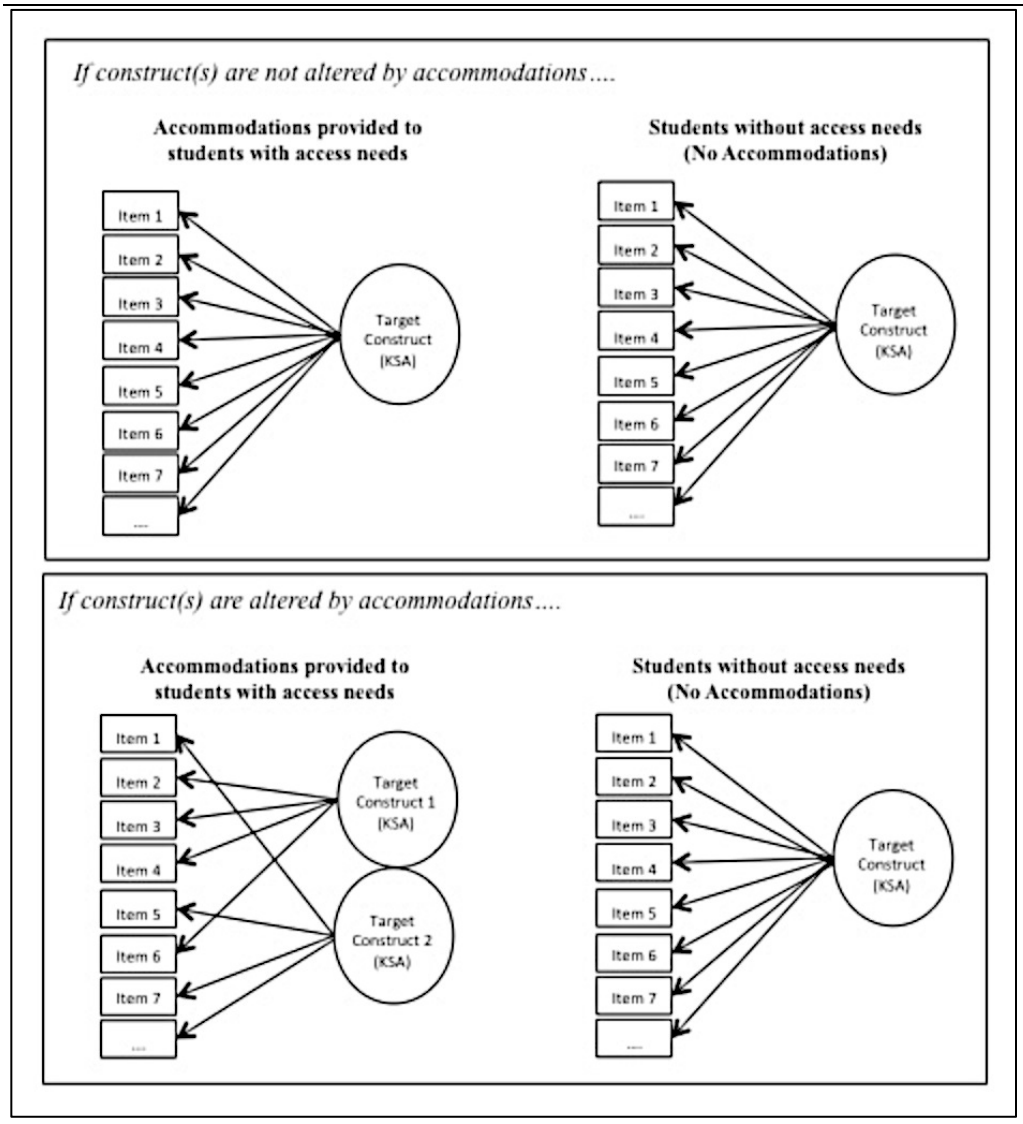
Descriptive statistics, including mean performance by item and by group were computed. Performance differences between groups were tested for statistical significance using t-tests and ANOVA. Cronbach's Alpha was also computed as an indicator of internal consistency or reliability among items for each testing condition. Cronbach's alpha normally ranges in value from zero to one, with values closer to one indicating a high degree of reliability. George and Mallery (2003) suggest the following rules of thumb for interpreting values: greater than .9 is excellent, between .89 and .8 is good, between .79 and .7 is acceptable, between .69 and .6 is questionable, between .59 and .5 is poor, and less than .49 is unacceptable (p. 231).

Confirmatory Factor Analysis

Factor analytic procedures can be useful in evaluating construct validity by determining the underlying dimensions or constructs measured by an assessment (Thurlow et al., 2000). If items appear to group together, we infer that they share a common factor that accounts for variation across items, ideally representing the construct of interest. When we observe similar structure across groups, we infer items and the assessment measure the same construct across groups.

In accommodations research, these techniques can be useful in determining whether the underlying construct or dimensions (in this case, 11th grade science achievement as defined by the NECAP standards) appear consistent under non-accommodated and accommodated conditions (Thurlow et al., 2000). As depicted in Figure 3.2, if the use of an accommodation does not alter the target construct, we would expect similar factor structures across accommodated and non-accommodated conditions. If the use of test accommodations or accessibility supports does alter the intended construct, then differences in the underlying factor structure or how items function together between non-accommodated and accommodated conditions should be evident (Bechard, Almond & Cameto, 2011).

Figure 3.2. Example of Underlying Factor Model For Accommodated and Non-Accommodated Conditions



To address the first research question, a confirmatory factory analysis was used to identify and compare underlying factor structures among items administered under non-accommodated and accommodated conditions using SPSS and Lisrel software. Four domains were purportedly measured by the 11th grade NECAP science assessment: (1) science process skills (28%), (2) earth space science (24%), (3) life science (24%), and (4) physical science (24%). However, it was hypothesized that the assessment provides a general measure of scientific knowledge and skills, represented by a single factor.

To test this hypothesis and determine the degree of invariance across groups, similar procedures described in Cook et al. (2010) were used (see Table 3.4). First, a series of single group Exploratory Factor Analyses (EFA) was conducted to determine the underlying factor structure individually for each test condition (No Accommodations, Accommodations - Paper and Accommodations - Nimble). This was followed by a series of single group confirmatory factor analyses and a series of multi-group confirmatory analysis to compare factor structure across groups. To test for factor invariance across conditions, a chi-square difference statistic was computed. This indicates whether a constrained model, where factor loadings are specified as equal for each group, had significantly worse fit than an unconstrained model, where factor loadings are permitted to vary across groups (Garson, 2011).

Table 3.4. Summary of Factor Analyses

Type	Question to be Answered	Number of Factors Hypothesize	Level of Analysis
Exploratory FA	Number of factors	-	Single Group
Single Group CFA	Confirm Single Factor	1	Single Group
Multi-Group CFA	Baseline Model	1	Multiple Groups
Multi-Group CFA	Equality of Factor Loadings	1	Multiple Groups
Multi-Group CFA	Equality of Factor Loadings &Variances	1	Multiple Groups
Multi-Group CFA	Equality of Factor Loadings, Variances & Residuals	1	Multiple Groups

(Cook et al., 2010)

The Exploratory Factor Analysis began with a principal components analysis (PCA), an exploratory technique used to specify one or more components that capture

most of the information contained in a complete set of items (Devillis, 2003). According to Garson (2011), PCA will account for the total variance of variables (in this case, items) and the resulting components will reflect both the common and unique variance of items. In extracting components, PCA first creates a linear equation that extracts the maximum total variance. A second linear equation is created that extracts the maximum remaining variance. This process continues until all the common and unique variance for a set of items is explained by the extracted factors.

Generally, factor analysis is approached under the assumption that one or a few big categories or concepts can be used to describe information gathered from all items. During analysis, results are assessed to determine the extent to which this assumption has held up. If it appears that one concept or category (i.e. latent variable/factor) has not done an adequate job of explaining covariation among items, a second concept is identified to explain the remaining covariation among items. This will continue until the amount of covariation remaining is acceptably small.

Statistically based methods of identifying factors will seek an exhaustive account of the factors underlying a set of items. Usually, however, the goal is to identify a small number of only the most influential factors. To determine the number of factors to be retained, a scree test (Cattell, 1966) was used. A scree test involves a plot of the eigenvalues associated with each of the extracted factors. Ideally, there is a point on the scree plot at which there is a sudden transition from vertical to horizontal points. Factors to be retained should lie on the vertical, well above horizontal points. One disadvantage of this technique is that the plot can be difficult to interpret if there is no abrupt drop in points.

In the event that multiple factors are identified, the raw, unrotated factors can often be rather meaningless abstractions (Devillis, 2003). Factor rotations can be used to present data in a way that is easier to understand by identifying clusters of variables (in this case items) that can be described in terms of some latent variable. To determine what rotation to use, one should look to the theory behind the instrument. Devillis (2003) advises that, "If theory strongly suggests correlated concepts, it probably makes sense for the factor analytic procedure to follow suit" (p. 124). When we are dealing with factors that are believed to correlate somewhat with one another, an oblique rotation may be appropriate. However, "what is lost when factors are rotated obliquely is the elegance and simplicity of uncorrelated dimensions" (Devillis, 2003, p. 123). Therefore an orthogonal rotation may be preferable. This type of rotation may produce results that are somewhat easier to interpret as factors are forced to be independent of one another.

In this case, science process skills were measured in the context of content from the three other domains (earth space science, life science, and physical science). Therefore, it was possible that a separate factor representing science process skills might be highly correlated with factors representing the other three domains. A single factor, representing general scientific knowledge and skills, encompassing all domains was also possible. Since there was no prior analysis of the underlying factor structure for the 11th grade NECAP science test available, the magnitude of the correlations between factors was used as a guide. An oblique rotation was employed first and the correlations among factors examined. If correlations are small, (< .15) an orthogonal rotation can be used to create a simpler model. If an oblique rotation reveals two highly correlated factors and items have substantial loadings on both, it may be worthwhile to extract one factor to see

if the two highly correlated factors merge into one. Using criteria applied by Cook et al. (2006) factor loadings of at least .30 or above were used to confirm that variables were represented by a factor.

Following exploratory analyses, CFA was conducted to validate findings for each group. Finally, a series of multi-group CFA was used to evaluate the extent of factor invariance among testing conditions. This involved simultaneous estimation of confirmatory factor models for data collected under each testing condition and testing a hypothesis that the factor structures were similar across groups (Long, 1983).

Assessing Model Fit Across Groups

To assess the fit of all confirmatory models (i.e. single group and multi-group analyses) to the hypothesized factor structure, a chi-square fit statistic was computed. In this case, the null hypothesis that the specified model provides an acceptable fit on the observed data was tested (Long, 1983). The null hypothesis is rejected when the chi-square statistic is larger than the critical value and we conclude the data do not confirm the hypothesized model (Long, 1983).

However, many have noted that the chi square statistic is sensitive to sample size and it is therefore advisable to also use other goodness of fit indicators (Cheung & Rensvold, 2002). Additional fit statistics used were the Root Mean Square Error of Approximation (RMSEA), Comparative Fit Index (CFI), Goodness of Fit (GFI) and Non-Normed Fit Index (NNFI). The RMSEA is the averaged squared difference between variable loadings for each group on each factor (Garson, 2011). RMSEA values generally fall between zero and two. According to Garson (2011), a RMSEA of zero

indicates a perfect match in terms of pattern and magnitude of loading while a value of two indicates loadings are at unity, but differ in sign between groups. RMSEA values less than .8 are considered reasonable while values less than .5 indicate a closer fit (Kline, 2005). Other empirically established criterion values to evaluate model fit are as follows: CFI > .90 (Bentler, 1990), GFI > .90 (Hoyle & Panter, 1995), and NNFI > .90 (RayKov & Marcoulides, 2000). Chi-square difference tests were used to evaluate nested models for multi-group analysis.

When possible, attempts should be made to explore and identify constructs represented by factors, especially when differences in factor structure are observed. Interpretation of factors was based upon observations of which items load heavily on a factor, which items load more modestly, and which items show no loading or negative loadings (Kane, 2006). Identification of the construct represented by a factor was based on the shared characteristics on items loading and not loading on a factor and the potential interaction these characteristics may have had with accommodations.

If accommodations, provided either through NimbleTools or other means, do not alter the intended construct, then factor analytic procedures should generally yield the same factor structure for data collected under non-accommodated and accommodated conditions (Thurlow et al., 2000). In the event that accommodations have altered the construct, then different factor structures are expected to emerge (Thurlow et al., 2000).

Differential Item Functioning

When item and test characteristics, including inaccessible test design, differentially impact a test taker's ability to demonstrate their true abilities related to the

target construct, construct irrelevant variance is introduced and measurement is said to be biased. The presence of item bias will impact one's ability to make valid inferences from test scores across different students. One of the most common methods used to explore potential item bias across groups is differential item functioning (DIF) analysis. As described in the previous chapter, DIF can be considered a number of different ways. In this case, Item Response Theory (IRT) is used to match groups on latent ability estimates and estimate item parameters separately for each group before they are compared. Items are said to demonstrate DIF when individuals or groups vary in the probability of answering a question correct after estimated ability or proficiency is held constant (Hambleton, Swaminthan, & Rogers, 1991; Lord, 1980). Items exhibiting DIF for one or more groups may indicate an item is not accessible and barriers may be present that prevent certain students from demonstrating what they know and can do (Johnstone, Thurlow, Moore & Altman, 2006).

If a large number of items exhibit DIF favoring a particular group, this may suggest an assessment, as a whole, does not function consistently across groups of test-takers and indicate differences in the underlying measurement scale for groups. Using student response data to multiple-choice items, this study examined differential functioning across three groups: (1) students who completed the science assessment using NimbleTools and received at least one embedded accommodation (Accommodations - Nimble; n = 656); (2) students who completed the paper-based form who received at least one accommodation traditionally offered by states (Accommodations - Paper; n = 2,343); and (3) students who completed the paper-based form without any accommodations (No Accommodations; n = 2,000).

IRT DIF

With item response theory, it is possible to construct an item characteristic curve (ICC) for each item (See Figures 3.3 and 3.4 from Thurlow et al., 2000).

Figure 3.3. Similar Item Characteristic Curves for Non-Accommodated and Accommodated Administrations (No DIF)

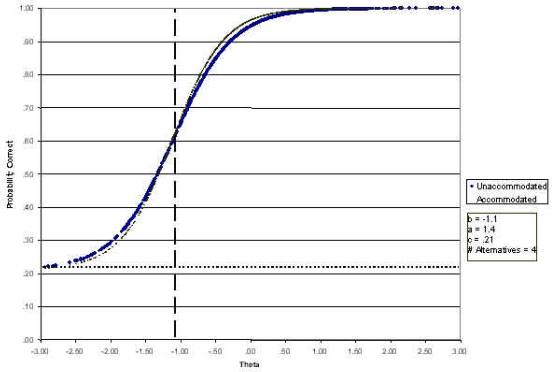
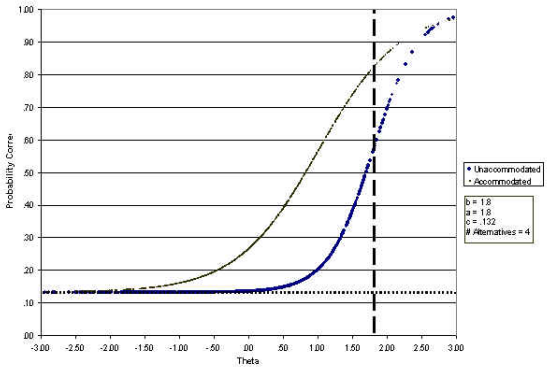


Figure 3.4. Different Item Characteristic Curves for Non-Accommodated and Accommodated Administrations



The vertical axis of each graph represents the probability of success on an item while the horizontal axis represents the ability or trait being measured by the item. Typically, the ICC for a properly functioning item should show that as ability or trait increases the probability of answering an item correctly should also increase. In this case, as the student achievement in science increases, the probability of answering an item correctly

should also increase. It is possible to construct separate ICCs for items administered under non-accommodated and accommodated conditions. When the two sets of ICCs appear almost identical (see Figure 3.1) this suggests an item "behaves" similarly under both test conditions and can be placed on the same measurement scale (Thurlow et al., 2000). When ICCs generated for an item administered under non-accommodated and accommodated conditions appear dramatically different (See Figure 3.2), then an item is said to show differential functioning, meaning an item does not function the same across subgroups. When a test contains a large number of items exhibiting DIF, this suggests measurement is inconsistent for different groups of examinees (Bielinski, Thurlow, Ysseldyke, Freidebach, & Freidebach, 2001).

Item parameter estimates for multiple-choice items were estimated using the 2-parameter logistic model, which includes item difficulty and discrimination (represented in Figure 3.5.).

Figure 3.5. 2-Parameter Logistic Model

$$P_{j1}(\theta_k) = \frac{1}{1 + \exp(-1.7a_j(\theta_k - b_j))}$$

Where: θ = underlying latent trait
a_i = discrimination of item i
b_i = difficulty of item i

This model permits item difficulty and discrimination to vary across items.

To conduct DIF analysis, BILOG-MG3 software was used, following procedures described by Bielinski and colleagues (2001). According to Bielinski et al. (2001), an advantage to this software is its ability to produce item difficulty estimates across multiple groups at once. To do so, one must define one group as the reference group and the remaining groups as focal groups. In this case, the No Accommodations group was

used as the reference group, while the Accommodations - Nimble and Accommodations - Paper were designated as focal groups. The purpose of defining the reference group is to set the mean and standard deviation of the item difficulty scale. Students with no identified special needs who took the paper and pencil form without accommodations were expected to be less sensitive to any accessibility barriers this test form may have presented. Item difficulty estimates for focal groups are then placed on the scale defined by the reference group while the sum of the item difficulty estimates in the focal group are set to equal the sum of the item difficulty estimates in the reference group.

Once items have been rescaled to this common scale, BILOG-MG3 then calculates the item difficulty difference across groups and the standard error of this difference. The standard error for these parameters was used to evaluate differences in item difficulty between groups, represented with the null hypotheses below:

$$H_o: b_{No\ Accommodations} = b_{Accommodations-Nimble}$$

$$H_o: b_{No\ Accommodations} = b_{Accommodations-Paper}$$

When the ratio of the difficulty difference to the standard error exceeds 2.0, this can be indicative of differential item functioning (Bielinski et al., 2001). The number of items and pattern of DIF (i.e. which items exhibit DIF? Do these items share certain characteristics?) was then considered for each comparison.

If accommodations are effective at "leveling the playing field" without altering the construct applied by students, then score interpretations are expected to be valid across all test takers, and there should be no evidence of DIF between accommodated and non-accommodated groups (Finch, Barton & Meyer, 2009). The presence of a large number of DIF items suggests that a test functions differently for different groups of

students (Bielinski et al., 2001) and intended score interpretation may not be valid for all test takers. Currently there is no research-based standard for determining what constitutes a "large" proportion of DIF items. Bolt and Ysseldyke (2007) observed that published tests usually include less than 15% of DIF items and offer this as a reference point.

Given this information, the next step is to re-examine test items that display differential functioning to determine the specific reasons items function differentially for groups of students. It may be the case that an item displays differential functioning, but closer examination reveals that the likely reason is relevant to the construct being measured and may be due to factors outside the test and accommodations, such as the quality of instruction received by different groups of students. A doctoral student with training in universal design and large-scale assessment conducted item review for this study. With the lens of accessibility in mind, the following item features and their interaction with accommodations and accessibility supports were considered: (1) item content (e.g. text length and complexity, graphics, target domain), (2) task demands (e.g. reasoning, calculation, or interpretation of a table or graph), and (3) presentation format (e.g. location on page, text formatting)

CHAPTER SUMMARY

This dissertation explored whether a technology-based approach to testing and accommodations was effective at addressing accessibility needs for test takers with special needs and resulted in scores that had similar psychometric qualities as those collected under non-accommodated conditions. It was hypothesized that the use of

technology would allow for more thoughtful and comprehensive application of UDL principles and higher quality and better standardized accommodations. It was expected that NimbleTools comprehensively addressed students' access needs and that measurement of 11th grade science therefore contained less contamination related to accessibility barriers. If this was the case, both item functioning and underlying factor structure should be similar across accommodated and non-accommodated groups. To determine the extent to which these hypotheses held true, two sets of analyses were conducted: confirmatory factor analysis and differential item functioning.

Using factor analytic techniques, the aim of the first analyses was to provide information regarding the underlying construct or dimensions being measured under accommodated and non-accommodation conditions to determine the extent to which a similar construct was measured under each (research question #1). If accommodations, provided either via NimbleTools or otherwise, did not alter the intended construct, then factor analytic procedures should generally yield the same factor structure for data collected under accommodated and non-accommodated conditions. This would suggest the condition under which the test was performed did not alter how items worked together to measure the target construct. This would also suggest that similar inferences regarding the intended construct could be made for students completing the assessment under different testing conditions. In the event that accommodations have altered the construct, then different factor structures were expected to emerge.

The second analysis, DIF, was used to explore the extent to which items function consistently across the three target groups. If a large number of items exhibit DIF and favor one group, this could suggest that overall the test did not function consistently

across accommodated and non-accommodated conditions (research question #2). DIF favoring students who receive accommodations via NimbleTools and/or on a paper assessment could suggest that accommodations provided inappropriate support and the construct was violated. The accommodation may have given test takers some unfair advantage or modified the construct in some way that altered the difficulty of an item or items. This pattern of DIF could also indicate that students taking the test under non-accommodated conditions may have needed an accommodation (Finch, Barton & Meyer, 2009) and were not properly identified. If there was DIF favoring students who took the test without accommodations, items were still differentially difficult for students with special needs in spite of accommodations. Through the lens of accessibility theory, this would suggest that the provided accommodations did not address accessibility concerns (Beddow, 2011). This may be the result of inappropriate accommodation assignment that either failed to provide the needed support or actually hindered student performance. It is also possible that accommodations were not effectively provided (e.g. reader mispronounced words during read aloud). The next chapter will present the results of these analyses.

CHAPTER 4: RESULTS

To answer questions related to the validity of student scores collected under accommodated and non-accommodated conditions, a secondary data analysis of operational data collected during the 2009 administration of the 11th grade NECAP science assessment was conducted. Results from two types of analyses are presented in this chapter: differential item functioning (DIF) and confirmatory factor analysis (CFA). To explore item functioning across accommodated and non-accommodated test conditions, DIF analyses were used, comparing item difficulty and discrimination for groups of test takers. If a large number of items exhibit DIF favoring one group over another, this could suggest inconsistent test functioning overall. CFA was used to examine underlying factor structure of items to determine if scores gathered under each test condition appeared to measure the same constructs. Both analyses were based on assessment data collected from three groups of students (1) students who received traditional accommodations on a paper-based form (Accommodations - Paper; n = 2,343), (2) students who completed computer-based testing and received technology-based supports using NimbleTools (Accommodations - Nimble, n = 656), and (3) students who did not receive accommodations and completed a paper-based form (No Accommodations, n = 2000).

ACCOMMODATION ASSIGNMENT

Table 4.1 summarizes accommodation assignment across test takers. The most frequently assigned accommodation was small group in a separate location. Sixty-nine percent of students assigned one or more accommodations were assigned to a small group

in a separate location. This was the case for both students who were assessed with the paper form and students who completed testing with NimbleTools. Other frequently assigned accommodations included administration with special education personnel (28.3% overall, 27.4% Accommodation - Paper, 31.3% Accommodations - Nimble) and extended time (28.1% overall, 29.9% Accommodation - Paper, 21.6% Accommodations - Nimble). Sixteen percent of accommodated students who completed the paper form received a read aloud. Nearly all NimbleTools users assigned a read aloud received it through the computer-based test delivery system as an embedded support. Among NimbleTools test takers, read aloud was the most frequently assigned support (87.3%).

Traditionally offered accommodations for which there was an equivalent NimbleTools support were not assigned to students at all or in very low rates. This included reduction of visual print by blocking or other techniques (0.2% overall), use of a visual magnification device (0%) and use of an acetate shield (0%). The rates of use for the equivalent NimbleTools supports were as follows: masking (37.2%), magnifier (9.1%) and colored overlay (17.4%). The most frequently assigned support among NimbleTools users was read aloud (87.3%). Other frequently assigned supports were allowed breaks (57.7%), masking (37.2%) and auditory calming (31.7%) (see Chapter 3 for a description of each NimbleTools support).

Table 4.1. Accommodations and Accessibility Supports by Test Condition

	Accommodations - Paper (n = 2343)		Accommodations - Nimble (n = 656)		Total (n = 2999)	
	#	%	#	%	#	%
Setting						
Separate location – individual	214	9.1	29	4.4	243	8.1
Separate location – small group	1702	72.6	357	54.4	2059	68.7
Location with minimal distraction	249	10.6	37	5.6	286	9.5
Preferential seating	41	1.7	5	0.8	46	1.5
Special acoustics	1	0.0	0	0.0	1	0.0
Special lighting or furniture	2	0.1	0	0.0	2	0.1
Administer with special education personnel	643	27.4	205	31.3	848	28.3
Administer with other school personnel known to the student	195	8.3	17	2.6	212	7.1
Administer with other school personnel at non-school setting	168	7.2	0	0.0	168	5.6
Auditory Calming via NimbleTools	0	0.0	208	31.7	208	31.7
Timing					0	0.0
Time of day change	26	1.1	5	0.8	31	1.0
Supervised breaks	362	15.5	74	11.3	436	14.5
Extended time	701	29.9	142	21.6	843	28.1
Allow Break via NimbleTools	0	0.0	377	57.5	377	57.5
Presentation						
Braille	0	0.0	0	0.0	0	0.0
Large print	9	0.4	1	0.2	10	0.3
Sign directions	10	0.4	0	0.0	10	0.3
Read aloud via proctor	376	16.0	1	0.2	377	12.6
Read aloud via NimbleTools	0	0.0	573	87.3	573	87.3
Student read aloud to self	8	0.3	1	0.2	9	0.3
Translate direction to other language	12	0.5	0	0.0	12	0.4
Underlining key info in directions	68	2.9	6	0.9	74	2.5
Visual magnification device	0	0.0	0	0.0	0	0.0
Magnification via NimbleTools	0	0.0	60	9.1	60	9.1
Reduction or visual print by blocking or other techniques	5	0.2	1	0.2	6	0.2
Masking via NimbleTools	0	0.0	244	37.2	244	37.2
Acetate shield	0	0.0	0	0.0	0	0.0
Overlay via NimbleTools	0	0.0	114	17.4	114	17.4
Auditory amplification or noise buffers	2	0.1	0	0.0	2	0.1
Word to work translation dictionary	12	0.5	0	0.0	12	0.4
Abacus for severe visual impairments or blindness	0	0.0	0	0.0	0	0.0
Reverse Contrast via NimbleTools	0	0.0	67	10.2	67	10.2
Color Choice via NimbleTools	0	0.0	86	13.1	86	13.1
Response						
Writes with word processor	34	1.5	16	2.4	50	1.7
Handwrites responses on separate sheet	13	0.6	1	0.2	14	0.5
Writes using brailler	0	0.0	1	0.2	1	0.0
Student indicates response to multiple-choice items	17	1.7	2	0.3	19	0.6
Student dictates constructed response to school personnel	39	1.7	11	1.7	50	1.7

	Accommodations - Paper (n = 2343)		Accommodations - Nimble (n = 656)		Total (n = 2999)	
	#	%	#	%	#	%
Student dictates constructed response using assistive technology device	4	0.2	0	0.0	4	0.1
Other						
Use Calculator	34	1.5	18	2.7	0	0.0
Other	2	0.1	0	0.0	0	0.0

Table 4.2 and 4.3 summarizes the total number of supports received by individual students. On average, accommodated students assessed with the paper form (Mean = 2.1, SD = 1.28) received less supports than students assessed with Nimble Tools (Mean = 4.0, SD = 2.6).

Table 4.2. Total Count of Assigned Accommodations and Accessibility Supports by Test Condition

	Accommodations - Paper (n =2343 students)		Accommodations - Nimble (n = 656 students)	
	%	#	%	#
1	41.0	960	91	13.9
2	29.3	687	137	20.9
3	17.1	401	130	19.8
4	8.2	193	56	8.5
5	1.9	44	67	10.2
6	1.6	38	63	9.6
7	0.5	12	36	5.5
8 or more	0.3	8	76	11.6

Table 4.3. Count of Assigned NimbleTools Features

	Accommodations - Nimble (n = 656 students)	
	%	#
1	30.3	199
2	33.8	22
3	11.4	75
4	12.7	84
5	1.7	11
6	2.3	15
7	3.0	20
8	4.6	30

OVERALL PERFORMANCE AND CLASSICAL ITEM STATISTICS

Scaled scores for the grade 11 NECAP science assessment ranged from 1100 to 1180. Performance was also described in terms of performance level. Table 4.4 and 4.5 summarize achievement by IEP status and test condition. A t-test and three-way ANOVA was used to determine if performance differences between groups were statistically significant. Performance differences were significant between students with an IEP and students who did not have an IEP ($t(4997) = 37.572$, $p < .05$). Students on IEPs performed lower. Performance differences between testing conditions were also significant ($F(2, 4996) = 488.33$, $p < .05$). Post hoc comparisons using the Tukey HSD test indicated significant differences between all three test conditions. Test takers who did not receive accommodations performed the highest ($M = 1134.65$, $SD = 8.23$). Accommodated students completing the paper form ($M = 11296.97$, $SD = 9.70$) performed higher than those who were assessed through NimbleTools ($M = 1125.67$, $SD = 7.61$). These performance differences are mirrored in performance level results (Table 4.5).

Table 4.4 Summary of Overall Mean Performance

	n	Mean Scaled Score	SD
IEP Status			
Yes	2140	1124.61	8.73
No	2859	1133.81	8.45
Test Condition			
No Accommodations	2000	1134.65	8.23
Accommodations - Paper	2343	1126.97	9.70
Accommodations - Nimble	656	1125.67	7.61

Table 4.5. Overall Results by Performance Level

	n	% Substantially Below Proficient	Partially Proficient	Proficient	Proficient with Distinction
IEP Status					
Yes	2140	28.8	47.1	23.4	0.7
No	2859	73.6	24.2	2.2	0.0
Test Condition					
No Accommodations	2000	24.8	48.1	26.3	0.9
Accommodations - Paper	2343	61.2	30.7	7.9	0.1
Accommodations - Nimble	656	71.2	27.9	0.9	0.0

Table 4.6 and 4.7 summarize item level performance by IEP status and test condition. T-tests were used to determine if performance differences between groups were statistically significant. To reduce the risk of Type I error, the Bonferroni correction was used to adjust p-values to account for multiple statistical tests being performed simultaneously on a single data set. A critical p-value of .05 was divided by the number of comparisons made (33). Using an adjusted $p < .0015$, item level performance was significantly higher on all items for students without disabilities.

Table 4.6. Item Means by Disability Status

	Student without Disabilities (SWOD) (n =2859)		Student with Disabilities (SWD) (n =2140)		SWOD – SWD
Item	Mean	SD	Mean	SD	Mean Diff.
1	.83	.374	.56	.496	**.27**
2	.64	.481	.53	.499	**.11**
3	.54	.499	.35	.478	**.18**
4	.52	.500	.42	.493	**.10**
5	.54	.498	.33	.469	**.22**
6	.46	.498	.38	.486	**.08**
7	.53	.499	.42	.493	**.11**
8	.50	.500	.34	.475	**.16**
9	.64	.479	.48	.500	**.17**
10	.65	.477	.44	.496	**.21**
11	.41	.491	.29	.453	**.12**
12	.59	.491	.39	.487	**.21**
13	.38	.485	.31	.461	**.07**
14	.55	.497	.41	.491	**.14**
15	.73	.443	.56	.497	**.17**
16	.48	.500	.40	.490	**.09**
17	.48	.500	.27	.445	**.21**
18	.47	.499	.36	.481	**.11**
19	.83	.376	.70	.458	**.13**
20	.60	.490	.52	.500	**.08**
21	.61	.488	.47	.499	**.14**
22	.43	.495	.28	.449	**.15**
23	.63	.484	.49	.500	**.14**
24	.59	.492	.39	.487	**.20**
25	.49	.500	.35	.477	**.14**
26	.40	.490	.30	.457	**.11**
27	.81	.396	.56	.496	**.24**
28	.46	.499	.37	.483	**.09**
29	.61	.487	.39	.488	**.22**
30	.54	.499	.37	.483	**.17**
31	.49	.500	.35	.477	**.14**
32	.72	.450	.48	.500	**.23**
33	.54	.499	.35	.476	**.19**

***Bold values indicate statistically significant group differences, p < .0015)**

Table 4.7 summarizes item level performance by test condition. Three-way ANOVAs were used to determine if performance differences between groups were statistically significant. Of particular interest was the difference between non-

accommodated and accommodated groups. Examining differences between non-accommodated students and students who were accommodated with a paper form, performance was significantly higher on all items for non-accommodated students. Similarly, performance was significantly higher for nearly all items for non-accommodated students compared to those accommodated using NimbleTools. Exceptions included items 16, 20 and 28. Performance was equivalent between groups on these items.

Items 16 and 28 are described later in this chapter and were also identified as exhibiting DIF. Item 20 was similar to both of these items in terms of length, general formatting and page location on the paper form. All three were relatively brief items and contained simplified language aside from scientific vocabulary. These items did not share characteristics in terms of item content and required test taker knowledge and skills. It is not readily clear based on review of items why on average NimbleTools test takers would perform similarly to non-accommodated students, while consistently performing lower on all other items.

Table 4.7. Item Means by Test Condition

	No Accommodations (n = 2000)		Accommodations - Paper (n = 2343)		NoAcc - AccPP	Accommodations - Nimble (n = 656)		NoAcc - AccNT
Item	Mean	SD	Mean	SD	Mean Diff	Mean	SD	Mean Diff
1	.86	.346	.64	.482	**.23**	.57	.495	**.29**
2	.65	.477	.55	.498	**.10**	.57	.495	**.08**
3	.55	.497	.40	.491	**.15**	.36	.479	**.20**
4	.54	.499	.44	.496	**.10**	.43	.496	**.11**
5	.56	.497	.39	.488	**.17**	.32	.469	**.23**
6	.47	.499	.40	.491	**.06**	.38	.486	**.08**
7	.52	.500	.46	.499	**.06**	.41	.492	**.11**
8	.53	.499	.39	.487	**.15**	.31	.464	**.22**
9	.67	.472	.51	.500	**.15**	.49	.500	**.17**
10	.68	.467	.49	.500	**.19**	.45	.498	**.23**
11	.42	.493	.34	.473	**.08**	.25	.431	**.17**
12	.61	.488	.45	.497	**.16**	.39	.488	**.22**
13	.38	.486	.33	.470	**.05**	.30	.457	**.08**
14	.57	.496	.45	.497	**.12**	.42	.493	**.15**
15	.75	.433	.59	.492	**.16**	.63	.482	**.12**
16	.49	.500	.40	.490	**.09**	.50	.500	-.01
17	.50	.500	.34	.475	**.16**	.21	.406	**.30**
18	.47	.499	.41	.492	**.06**	.36	.481	**.11**
19	.85	.362	.72	.450	**.13**	.76	.429	**.09**
20	.60	.491	.54	.499	**.06**	.55	.498	.04
21	.61	.487	.51	.500	**.10**	.47	.500	**.14**
22	.44	.497	.34	.474	**.10**	.24	.426	**.20**
23	.63	.482	.52	.500	**.11**	.52	.500	**.11**
24	.60	.489	.44	.496	**.16**	.42	.494	**.18**
25	.52	.500	.39	.487	**.13**	.31	.463	**.21**
26	.41	.491	.33	.469	**.08**	.31	.462	**.10**
27	.82	.387	.63	.484	**.19**	.62	.485	**.19**
28	.46	.499	.39	.487	**.08**	.44	.496	.02
29	.65	.479	.45	.497	**.20**	.38	.485	**.27**
30	.56	.497	.42	.493	**.14**	.36	.482	**.20**
31	.49	.500	.39	.487	**.10**	.40	.490	**.09**
32	.74	.438	.54	.498	**.20**	.51	.500	**.24**
33	.57	.495	.40	.489	**.17**	.33	.471	**.24**

***Bold values indicate statistically significant group differences between accommodated and non-accommodated conditions (p < .05)**

Although these results may offer a baseline and some context for which to interpret more complex analyses presented later in this chapter, it's important to note the limitations of classical test statistics. This approach does not control for ability differences and is also sensitive to large sample size. This may lead to statistical significance based on p-values for even small group differences that failed to have practical significance.

RESEARCH QUESTION #1: IS THE UNDERLYING FACTOR STRUCTURE CONSISTENT FOR SCORES GATHERED UNDER ACCOMMODATED AND NON-ACCOMMODATED CONDITIONS?

Confirmatory factor analysis was used to explore measurement invariance across groups. SPSS v22 was used to conduct initial exploratory factor analysis. Then Lisrel v9.2 was used to conduct single and multi-group confirmatory factor analyses.

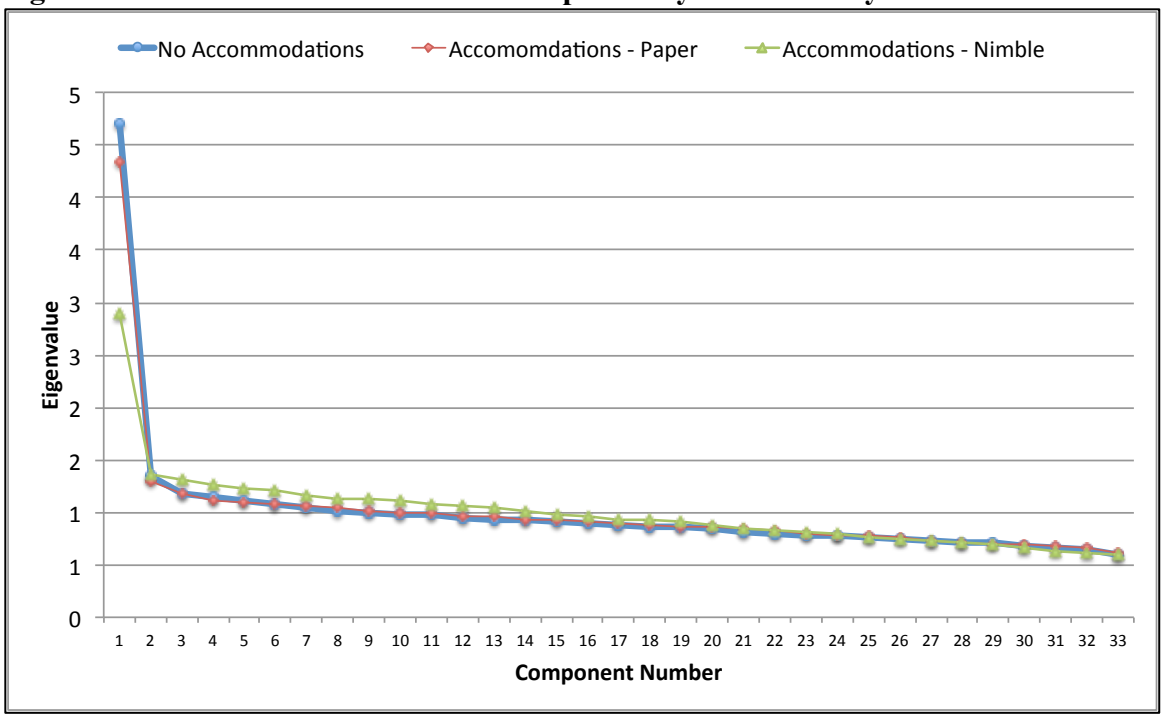
Single Group Exploratory Factor Analysis (EFA)

Single group EFA was conducted to determine underlying factor structure individually for each group. Specifically, principal components analysis (PCA) was used to specify one or more components that capture most of the information contained in the complete set of items.

To determine the number of factors to be retained, a scree test was used. Visual examination of scree plots indicated that for all groups, the assessment measured a single factor (Figure 4.1). However, the percentage of total variance accounted for by the largest eigenvalue for each group was low. Variance accounted for was 14.26% for No

Accommodations, 13.15% for Accommodations - Paper and 8.74% for Accommodations - Nimble.

Figure 4.1. Scree Plots Obtained from Exploratory Factor Analysis



In addition to the scree plot, factor loadings were also examined. Factors loadings by item for rotated and unrotated solutions as well factor correlations for each testing condition can be found in Appendix A. Factor loadings of .30 or above were used to identify salient factor loadings. Examining the unrotated solutions, approximately two-thirds of items loaded most highly on the first factor and/or had a factor loading of .30 or higher for No Accommodations (25 items) and Accommodations - Paper conditions (22 items). Factor structure was less clear for Accommodations - Nimble. Only 17 out of 33 items had factor loadings .30 or higher on the first factor.

Oblique (promax) and orthogonal (varimax) rotations were applied to explore factors and potential correlations further. When a promax rotation was applied, there was

evidence of correlated factors among all three groups, with several factor correlations greater than .15. However general factor structure was not improved. For example, examining the rotated solution for No Accommodations, the most items loaded on any one factor was .4. Similar results were found for Accommodations - Paper and Accommodations - Nimble. Factor structure was similarly poor when a varimax rotation was applied, with only a few items loading on each potential factor. It should also be noted that rotation cannot improve the basic attributes of an analysis, including the amount of variance extracted from items (Costello & Osbourne, 2005). Costello and Osbourne also warn that if factor structure does not become clearer after multiple runs, this could indicate a problem with item construction or the hypothesized construct.

Reliability of Factors

Pursuing the possibility of a single factor model, Table 4.8 lists Cronbach's alpha for the complete sample and subgroups. Overall reliability for the 33 multiple-choice items was .80. Applying criteria described in the previous chapter, this indicated overall reliability was just within the good range. Reliability was also just within the good range for non-accommodated test takers ($\alpha = .80$). Reliability was acceptable and lower for Accommodations - Paper ($\alpha = .78$). Reliability was questionable among students with IEPs ($\alpha = .68$) and Accommodations - Nimble ($\alpha = .63$).). Table 4.9 summarizes the resulting Cronbach's alpha if an item was removed. Reliability did not appear to be meaningfully impacted by any one particular item.

Table 4.8. Reliability Statistics by IEP Status and Test Condition

	Cronbach's Alpha	n
All	.80	4999
IEP Status		
No IEP	.80	2859
IEP	.68	2140
Test Condition		
No Accommodations	.80	2000
Accommodations - Paper	.78	2343
Accommodations - Nimble	.63	656

Table 4.9. Cronbach's Alpha if Deleted by IEP Status and Test Condition

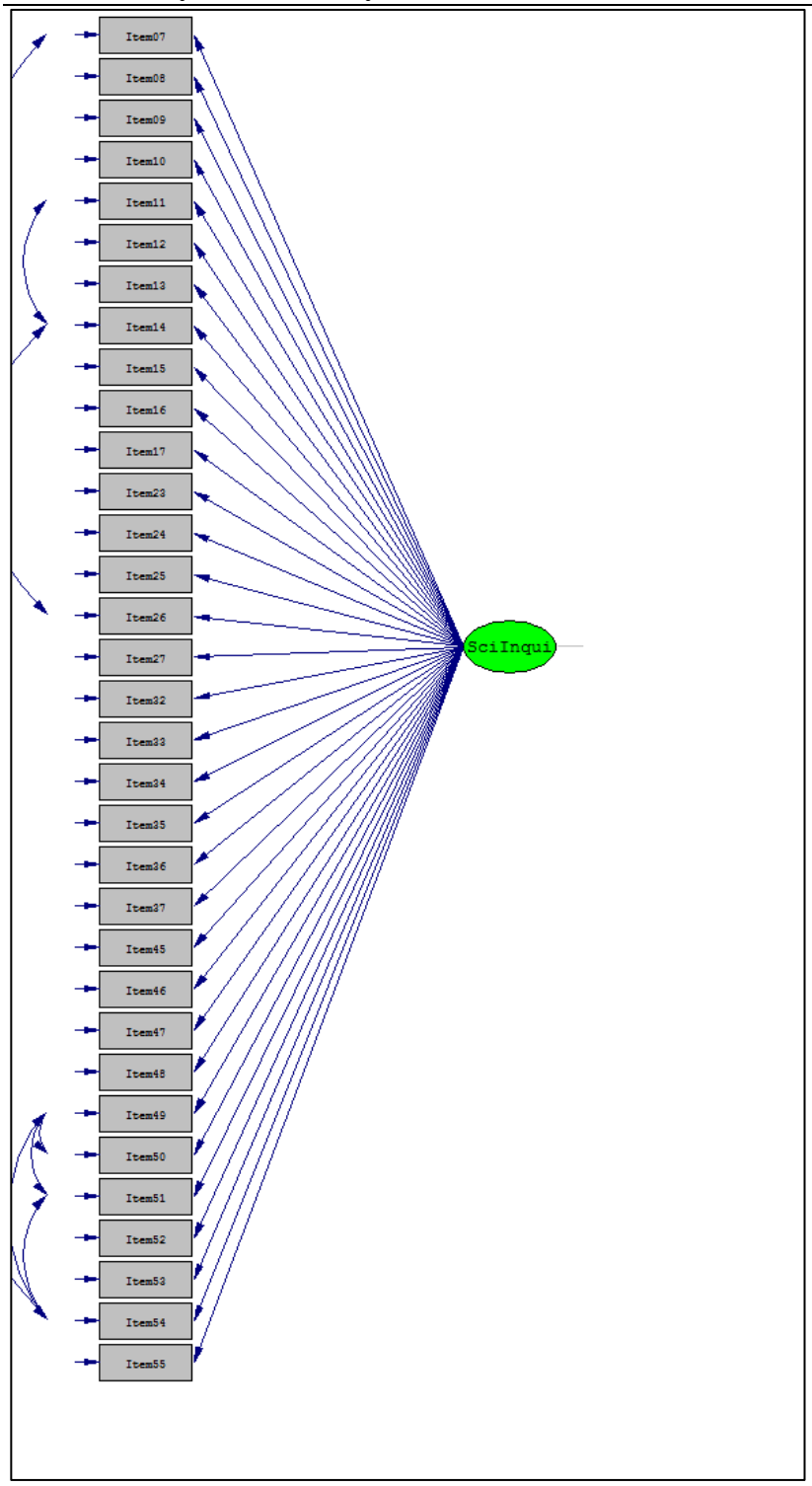
Item	Total Sample (n = 4999)	No IEP (n = 2859)	IEP (n = 2140)	No Accommodations (n = 2000)	Accommodations - Paper (n = 2343)	Accommodations - Nimble (n = 656)
1	.792	.790	.666	.788	.765	.608
2	.801	.799	.679	.797	.774	.625
3	.795	.791	.675	.788	.770	.615
4	.802	.798	.685	.795	.778	.626
5	.796	.789	.684	.785	.773	.630
6	.802	.798	.684	.795	.778	.624
7	.801	.796	.683	.793	.775	.628
8	.796	.789	.681	.786	.773	.620
9	.796	.793	.670	.790	.770	.612
10	.794	.790	.671	.787	.769	.611
11	.799	.793	.684	.790	.774	.631
12	.795	.791	.676	.788	.770	.616
13	.803	.798	.689	.796	.779	.630
14	.797	.793	.675	.791	.770	.616
15	.797	.795	.671	.791	.771	.618
16	.802	.797	.687	.793	.778	.631
17	.794	.790	.674	.788	.769	.616
18	.799	.794	.678	.791	.772	.624
19	.796	.792	.670	.789	.770	.615
20	.803	.80	.685	.797	.778	.625
21	.797	.793	.674	.789	.771	.618
22	.796	.791	.677	.788	.771	.620
23	.796	.791	.677	.788	.770	.623
24	.798	.793	.682	.790	.773	.628
25	.796	.791	.676	.788	.772	.619
26	.80	.796	.682	.794	.776	.617
27	.791	.788	.665	.784	.764	.605
28	.799	.794	.680	.789	.775	.624
29	.793	.788	.671	.786	.767	.614
30	.797	.792	.679	.789	.773	.623
31	.799	.792	.687	.788	.774	.635

Item	Total Sample (n = 4999)	No IEP (n = 2859)	IEP (n = 2140)	No Accommodations (n = 2000)	Accommodations - Paper (n = 2343)	Accommodations - Nimble (n = 656)
32	.793	.790	.669	.788	.766	.615
33	.794	.789	.675	.786	.769	.616

Single Group Confirmatory Factor Analysis

Though results from initial exploratory analysis were unclear and perhaps signaled flaws with the test itself, the fit of a single factor model was evaluated for each test condition. Figure 4.2 depicts the relationship between items and a single, general factor. It was hypothesized that the assessment measured a single underlying construct, labeled scientific knowledge and skills. This single factor solution was verified for each test condition separately.

Figure 4.2. Single Factor Model Used for Single Group and Multi-Group Confirmatory Factor Analysis



Results from single group confirmatory analyses are shown in Table 4.10 and Table 4.11. Table 4.10 provides factor loadings by item for each test condition, including statistical significance. Factor loadings on a single global factor were significant at $p < .05$ for nearly all items for all three test conditions. Loadings were not significant for five items for Accommodations - Nimble (Items 7, 11, 13, 16, and 31).

Table 4.10. Summary of Factor Loadings for Single Group Analysis

	No Accommodations		Accommodations - Paper		Accommodations - Nimble	
	Unstandardized Factor Loadings	SE	Unstandardized Factor Loadings	SE	Unstandardized Factor Loadings	SE
1	**.14**	.10	**.24**	.17	**.21**	.20
2	**.06**	.22	**.12**	.23	**.08**	.24
3	**.19**	.21	**.17**	.21	**.14**	.21
4	**.09**	.24	**.07**	.24	**.06**	.24
5	**.22**	.20	**.14**	.22	**.05**	.22
6	**.09**	.24	**.08**	.23	**.08**	.23
7	**.11**	.24	**.11**	.24	.04	.24
8	**.20**	.21	**.14**	.22	**.09**	.21
9	**.15**	.20	**.18**	.22	**.17**	.22
10	**.19**	.18	**.19**	.21	**.17**	.22
11	**.15**	.22	**.12**	.21	.01	.19
12	**.19**	.20	**.17**	.22	**.13**	.22
13	**.08**	.23	**.05**	.22	.03	.21
14	**.15**	.22	**.18**	.22	**.14**	.22
15	**.13**	.17	**.17**	.21	**.14**	.21
16	**.12**	.24	**.07**	.23	.03	.25
17	**.19**	.21	**.18**	.19	**.12**	.15
18	**.15**	.23	**.16**	.22	**.08**	.22
19	**.14**	.11	**.17**	.17	**.12**	.17
20	**.07**	.24	**.08**	.24	**.08**	.24
21	**.17**	.21	**.18**	.22	**.13**	.23
22	**.19**	.21	**.17**	.20	**.10**	.17
23	**.18**	.20	**.18**	.22	**.08**	.24
24	**.16**	.21	**.14**	.23	**.06**	.24
25	**.19**	.21	**.15**	.22	**.12**	.20
26	**.11**	.23	**.10**	.21	**.13**	.20
27	**.20**	.11	**.25**	.17	**.22**	.19
28	**.11**	.22	**.18**	.22	**.08**	.24
29	**.21**	.19	**.21**	.20	**.17**	.21
30	**.18**	.22	**.15**	.22	**.09**	.22
31	**.18**	.22	**.11**	.22	-.01	.24
32	**.16**	.16	**.23**	.20	**.16**	.23
33	**.21**	.20	**.19**	.20	**.13**	.20

***Bolded values are significant at p < .05**

Table 4.11 summarizes indices used to evaluate model fit. Results indicated that the one-factor model was a reasonable fit for all test conditions. The majority of fit

indices, with the exception of the Chi-square values, indicated acceptable fit for a single factor model for each group. RMSEA values were well below the .05 criterion. This was especially true for Accommodations - Nimble, with a RMSEA of .009 indicating excellent fit. All GFI, CFI and NNFI values were above .90.

Table 4.11. Summary of Fit Statistics for Single Group Confirmatory Factor Analysis (One Factor)

Group	DF	Normal Theory Chi-Sq	P-Value	RMSEA	GFI	CFI	NNFI
No Accommodations	488	901.02	.000	.021	.973	.924	.917
Accommodations – Paper	488	967.00	.000	.021	.975	.913	.906
Accommodations - Nimble	488	512.76	.212	.009	.956	.959	.956

Chi-square values were significant at the $p < .05$ level for No Accommodations and Accommodations - Paper groups, but not for Accommodations - Nimble. A significant value contradicts other fit indices and indicates a single factor model did not show acceptable fit to the data for these groups. However, the Chi-square is known to be sensitive to sample size (Jöreskog, 1969). As the number of cases increases, it becomes more difficult to retain the null hypothesis. With that in mind, although fit indices were slightly contradictory, it was concluded that the single factor model showed acceptable fit to proceed with testing for multi-group analyses.

Multi-Group Confirmatory Factor Analysis

Following single group analysis, multi-group analyses were completed to test for factor invariance across groups. If factor invariance is demonstrated, it could be

concluded that accommodated conditions did not change the underlying construct measured by the NECAP. Paired comparisons were conducted between No Accommodations and Accommodations - Paper and No Accommodations and Accommodations - Nimble. As described in chapter three, factorial invariance was explored through four steps:

(1) Establish a baseline model

(2) Test equality of factor loadings

(3) Test equality of factor loadings and factor variances

(4) Test equality of factor loadings, factor variances and errors of measurement

Results of multi-group analyses for the comparison between No Accommodations and Accommodations - Paper are provided in Table 4.12. Based on these results, acceptance or rejection of the hypothesis of factorial invariance was not clear cut. In the first step beyond the baseline model a significant Chi-square difference test (based on the Normal Theory Chi-square) suggested the hypothesis of equality of factor loadings was to be rejected. However the remaining goodness of fit indices met the criteria described in the previous chapter, suggesting model fit. Cook and colleagues (2010) encountered a similar result. Given the Chi-square sensitivity to sample size, they elected to give more weight to the RMSEA, CFI, NNFI and GFI. Proceeding in the same fashion, the RMSEA values were well below the .05 criterion and GFI, NNFI and CFI were above .90. Interpretation of the remaining fit indices indicated that the NECAP assessment was measuring a single factor and that this was similar across accommodated and non-accommodated paper-based conditions.

Results of multi-group analyses for the comparison between No Accommodations and Accommodations - Nimble are provided in Table 4.13. Results generally mirrored those found in the first paired comparison. Again, a significant Chi-square difference test (based on the Normal Theory Chi-square) suggested the hypothesis of equality of factor loadings was to be rejected. However, RMSEA values were well below the .05 criterion and nearly all GFI, NNFI and CFI values were above .90. An exception was the GFI associated with the last model, which examined invariance of factor loadings, factor variances and errors of measurement. The GFI fell slightly below the .90 criteria though other fit indicators (i.e. NNFI and CFI) pointed toward good model fit. If more weight is given to the RMSEA, CFI, and NNFI, results indicated that the NECAP assessment was measuring a single factor and that this was similar across NimbleTools and non-accommodated conditions.

Table 4.12. Summary of Multi-Group CFA for No Accommodations vs. Accommodations - Paper

		Goodness of Fit							
Group	DF	Normal Theory Chi- Sq	p – Value	RMSEA	GFI	CFI	NNFI	Chi Sq Difference	DF	Chi- Sq Diff p-value
Baseline	900	2063.52	.000	.022	.973	.962	.960	-	-	-
Invariance of factor loadings	1022	2276.46	.000	.024	.970	.956	.954	212.94	32	.000
Invariance of factor loadings and factor variances	1023	2276.58	.000	.024	.970	.956	.955	.12	1	.729
Invariance of factor loadings, factor variances and errors of measurement	1056	2785.95	.000	.028	.938	.939	.939	509.37	33	.000

Table 4.13. Summary of Multi-Group CFA for No Accommodations vs. Accommodations - Nimble

		Goodness of Fit							
Group	DF	Normal Theory Chi- Sq	p – Value	RMSEA	GFI	CFI	NNFI	Chi Sq Difference	DF	Chi- Sq Diff p-value
Baseline	990	1541.74	.000	.021	.955	.965	.962	-	-	-
Invariance of factor loadings	1022	1684.24	.000	.022	.943	.958	.956	142.50	32	.000
Invariance of factor loadings and factor variance	1023	1753.60	.000	.023	.950	.953	.952	69.36	1	.000
Invariance of factor loadings, factor variances and errors of measurement	1056	2190.16	.000	.028	.866	.928	.928	436.56	33	.000

RESEARCH QUESTION #2: DO ITEMS FUNCTION SIMILARLY UNDER ACCOMMODATED AND NON-ACCOMMODATED CONDITIONS? SPECIFICALLY, HOLDING ABILITY CONSTANT, ARE ITEM DIFFICULTY AND DISCRIMINATION EQUIVALENT FOR ACCOMMODATED STUDENTS AND NON-ACCOMMODATED STUDENTS?

As described in the previous chapter, the complete differential item functioning (DIF) analysis involved four steps. During the first two steps, item parameters were calculated by fitting the model to each group separately. Initial item parameter estimates for discrimination (a) and difficulty (b) can be found in Table 4.14.

Table 4.14. Single Group 2PL Item Parameter Estimates and Standard Errors

	No Accommodations				Accommodations - Paper				Accommodations - Nimble			
	(a)	SE	(b)	SE	(a)	SE	(b)	SE	(a)	SE	(b)	SE
1	.823	.058	-1.731	.090	.895	.055	-.550	.037	.572	.074	-.356	.096
2	.194	.028	-1.919	.301	.326	.029	-.362	.083	.251	.047	-.692	.225
3	.520	.038	-.297	.058	.462	.033	.548	.069	.411	.061	.945	.176
4	.241	.029	-.389	.120	.203	.026	.722	.151	.236	.046	.716	.240
5	.657	.043	-.272	.046	.380	.031	.755	.091	.211	.044	2.100	.484
6	.247	.029	.332	.116	.219	.026	1.076	.169	.259	.048	1.137	.272
7	.290	.030	-.200	.096	.294	.028	.322	.091	.204	.042	1.100	.319
8	.594	.040	-.168	.050	.372	.031	.794	.095	.310	.055	1.579	.303
9	.441	.036	-1.039	.096	.502	.035	-.094	.053	.453	.063	.040	.110
10	.616	.043	-.878	.066	.512	.034	.037	.053	.465	.065	.287	.113
11	.415	.033	.538	.080	.343	.029	1.244	.127	.187	.042	3.609	.843
12	.532	.038	-.587	.062	.457	.033	.292	.061	.370	.057	.783	.171
13	.232	.028	1.263	.190	.179	.025	2.381	.359	.195	.043	2.665	.624
14	.404	.033	-.427	.075	.468	.032	.295	.060	.373	.056	.581	.154
15	.453	.037	-1.611	.130	.506	.033	-.496	.059	.378	.058	-.930	.180
16	.307	.031	.087	.089	.200	.026	1.248	.199	.181	.039	.040	.255
17	.514	.036	-.028	.056	.529	.035	.835	.072	.474	.070	1.886	.253
18	.397	.033	.216	.073	.412	.031	.567	.075	.271	.051	1.300	.291
19	.735	.056	-1.731	.099	.643	.044	-1.078	.067	.486	.077	-1.579	.225
20	.200	.027	-1.183	.207	.214	.026	-.458	.127	.253	.047	-.522	.209
21	.472	.035	-.664	.073	.469	.033	-.096	.056	.350	.055	.186	.141
22	.529	.037	.310	.059	.471	.033	.934	.082	.382	.062	1.958	.314
23	.528	.039	-.726	.068	.488	.035	-.137	.054	.266	.049	-.229	.181
24	.439	.035	-.635	.077	.365	.030	.419	.078	.218	.044	.876	.276
25	.522	.037	-.097	.056	.393	.031	.754	.086	.368	.059	1.396	.241
26	.301	.031	.785	.117	.291	.028	1.537	.167	.399	.062	1.315	.216
27	1.218	.083	-1.194	.046	.937	.058	-.507	.035	.702	.091	-.552	.089
28	.458	.036	.218	.065	.309	.029	.939	.117	.278	.050	.559	.193
29	.688	.044	-.660	.052	.609	.038	.225	.049	.469	.063	.709	.136

	No Accommodations				Accommodations - Paper				Accommodations - Nimble			
	(a)	SE	(b)	SE	(a)	SE	(b)	SE	(a)	SE	(b)	SE
30	.483	.037	-.341	.062	.383	.031	.558	.081	.286	.051	1.209	.261
31	.492	.036	.049	.059	.307	.029	.929	.119	.144	.034	1.664	.504
32	.589	.039	-1.273	.086	.709	.041	-.213	.040	.419	.058	-.042	.118
33	.596	.039	-.352	.052	.516	.034	.545	.063	.398	.059	1.135	.195

Next item parameters were calculated fitting the model to both groups at the same time. The results were compared to determine which items appeared to exhibit differential functioning. Items that did not exhibit DIF were selected as anchor items. In this case, five items (items 6, 7, 13, 14, and 15) were selected for the No Accommodations and Accommodations - Nimble comparison and for No Accommodations and Accommodations - Paper (Item 8, 13, 18, 19 and 20). Although in many large-scale settings, it is customary to use an anchor section consisting of a larger number of anchor items, Thissen, Steinberg and Wainer (1993) demonstrated that IRT DIF procedures can be performed acceptably with relatively few anchor items. In the final step, the parameters for anchor items were set equal for the two groups and item parameters were estimated again for each group and compared for differential functioning.

Tables 4.14 and 4.15 provide adjusted difficulty parameters and standard error estimates provided by BILOG-MG3. Adjusted difficulty parameters are the result of rescaling that places item difficulty estimates for focal groups (i.e. accommodated groups) on the scale defined by the reference group (i.e. non-accommodated group), which allows for more direct comparison. When the ratio of the difficulty difference to the standard error exceeds 2.0, this can be indicative of differential item functioning (Bielinski et al., 2001). Applying this criterion, nine items showed evidence of DIF

between No Accommodations and Accommodations - Paper (Items 1, 11, 15, 21, 22, 23, 24, 28, and 32; See Table 4.14).

Table 4.15. Adjusted Difficulty Difference: No Accommodations vs. Accommodations - Paper

	No Accommodations		Accommodations - Paper		NoAcc - AccPaper		
Item	(b)	SE	(b)	SE	(b) diff.	SE	Diff:SE Ratio
1	-1.550	.059	-1.277	.031	.273	.067	**4.103**
2	-1.516	.170	-1.211	.107	.305	.201	1.520
3	-.298	.00	-.323	.00	-.025	.00	**Anchor**
4	-.432	.131	-.086	.132	.346	.186	1.860
5	-.315	.056	-.228	.057	.088	.080	1.103
6	.356	.123	.270	.146	-.086	.191	.451
7	-.202	.094	-.444	.089	-.242	.130	1.867
8	-.207	.00	-.224	.00	-.018	.00	**Anchor**
9	-.947	.072	-.863	.053	.084	.089	.945
10	-.893	.060	-.753	.044	.141	.074	1.893
11	.547	.078	.298	.090	-.249	.119	**2.096**
12	-.600	.060	-.523	.052	.077	.080	.966
13	1.414	.00	1.344	.00	-.070	.00	**Anchor**
14	-.393	.064	-.476	.060	-.083	.087	.950
15	-1.503	.092	-1.259	.057	.244	.109	**2.247**
16	.103	.109	.241	.131	.138	.171	.807
17	-.036	.052	.004	.061	.040	.080	.497
18	.198	.067	-.226	.069	-.423	.096	**Anchor**
19	-1.903	.00	-1.772	.00	.132	.00	**Anchor**
20	-1.177	.00	-1.246	.00	-.069	.00	**Anchor**
21	-.644	.064	-.863	.053	-.218	.083	**2.618**
22	.305	.058	.067	.066	-.238	.088	**2.713**
23	-.714	.061	-.900	.048	-.186	.078	**2.398**
24	-.660	.075	-.409	.065	.251	.099	**2.543**
25	-.109	.061	-.146	.064	-.037	.088	.420
26	.773	.105	.697	.134	-.076	.170	.447
27	-1.183	.040	-1.219	.026	-.036	.048	.754
28	.241	.073	-.024	.079	-.265	.108	**2.460**
29	-.654	.048	-.597	.041	.056	.063	.891
30	-.363	.065	-.300	.064	.063	.091	.695
31	.053	.069	-.067	.075	-.120	.102	1.176
32	-1.136	.060	-.985	.040	.151	.072	**2.099**
33	-.360	.051	-.309	.051	.051	.072	.708

***Bold values indicate ratio exceeds 2.0 criteria**

Using the same criteria, eight items (Items 1, 16, 17, 23, 28, 29, 31 and 32) showed evidence of DIF between No Accommodations and Accommodations - Nimble (Table 4.15).

Table 4.16. Adjusted Difficulty Difference: No Accommodations vs. Accommodations - Nimble

	No Accommodations		Accommodations - Nimble		NoAcc. - AccNimble		
Item	(b)	SE	(b)	SE	(b) diff.	SE	Diff:SE Ratio
1	-1.655	.078	-1.162	.060	.493	.098	**5.031**
2	-1.945	.308	-1.796	.273	.149	.412	.362
3	-.300	.056	-.231	.105	.069	.119	.580
4	-.398	.122	-.238	.214	.160	.246	.650
5	-.295	.050	-.160	.095	.135	.108	1.250
6	.286	.00	.219	.00	-.067	.00	**Anchor**
7	-.221	.00	-.179	.00	.043	.00	**Anchor**
8	-.179	.051	-.080	.102	.098	.114	.860
9	-.985	.085	-.894	.102	.091	.133	.684
10	-.865	.062	-.733	.079	.132	.100	1.320
11	.564	.086	.825	.197	.262	.215	1.219
12	-.585	.060	-.410	.097	.175	.114	1.535
13	1.284	.00	1.411	.00	.127	.00	**Anchor**
14	-.414	.00	-.432	.00	-.018	.00	**Anchor**
15	-1.622	.00	-1.670	.00	-.048	.00	**Anchor**
16	.089	.095	-.905	.160	-.994	.186	**5.344**
17	-.036	.054	.640	.144	.676	.154	**4.390**
18	.213	.074	-.044	.143	-.257	.161	1.596
19	-1.676	.091	-1.912	.092	-.236	.129	1.829
20	-1.199	.211	-1.568	.251	-.369	.328	1.125
21	-.662	.071	-.799	.101	-.138	.123	1.122
22	.300	.059	.438	.134	.138	.146	.945
23	-.752	.070	-1.054	.094	-.303	.117	**2.590**
24	-.672	.082	-.468	.120	.204	.145	1.407
25	-.105	.056	.030	.116	.135	.129	1.047
26	.738	.107	.588	.208	-.150	.234	.641
27	-1.176	.043	-1.255	.046	-.080	.062	1.290
28	.216	.067	-.584	.110	-.801	.128	**6.258**
29	-.658	.050	-.475	.076	.184	.091	**2.022**
30	-.353	.063	-.203	.116	.150	.132	1.136
31	.050	.067	-.360	.116	-.411	.134	**3.067**
32	-1.254	.080	-.962	.081	.292	.114	**2.561**
33	-.357	.051	-.198	.094	.159	.107	1.486

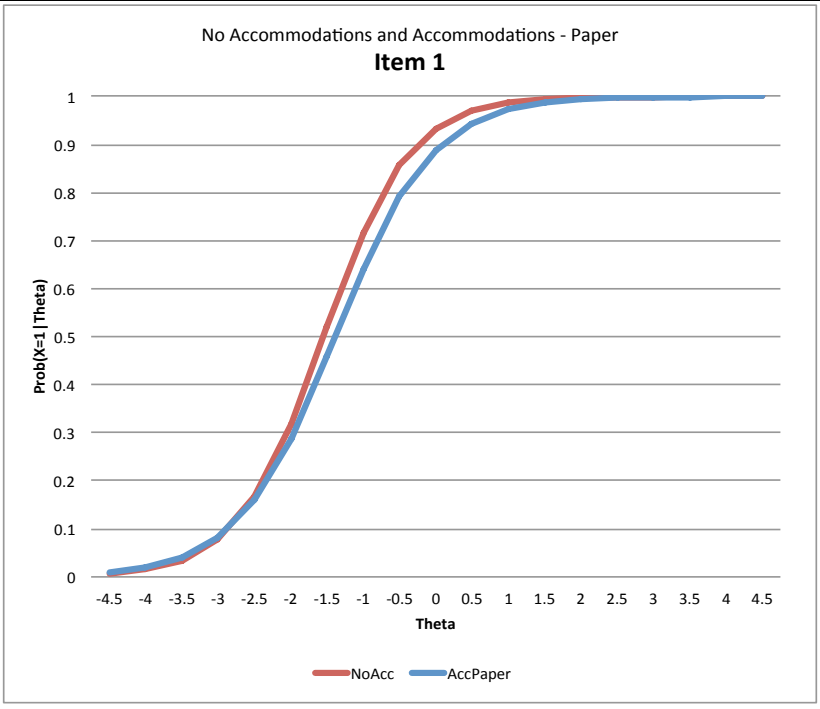
***Bold values indicate ratio exceeds 2.0 criteria**

Separate Item Characteristic Curves (ICCs) were constructed for items administered under non-accommodated and accommodated conditions and were visually examined for large differences as additional evidence of DIF. The vertical axis of each graph represents the probability of success of on an item while the horizontal axis represents the ability or trait being measured by the item. Visual patterns of DIF are described as uniform and non-uniform. Uniform DIF suggests that an item is

systematically more difficult for members of one group, even after ability matching. This is caused by a shift in the difficulty (b) parameter. Non-uniform DIF refers to a shift in item difficulty that is not consistent across ability level. This is generally caused by a shift in the discrimination (a) parameter, but may also involve a shift in difficulty (b).

ICCs for items identified as showing evidence of DIF between No - Accommodations and Accommodations - Paper (Items 1, 11, 15, 21, 22, 23, 24, 28, and 32) are found in Appendix B. In general ICCs for these items showed signs of slight DIF or none at all. Rather ICCs for accommodated and non-accommodated students appeared quite similar. Osterlind and Everson (2010) advised that the criterion of a 2.0 or higher ratio difference might result in over identifying items as DIF. A more conservative criterion of a ratio greater than 3 may be applied. Under this stricter criterion, 1 item was found to exhibit DIF (Item 1). However, again the uniform DIF exhibited appears small (Figure 4.3).

Figure 4.3. ICC for Item 1: No Accommodations vs. Accommodations - Paper



ICCs for items identified as showing evidence of DIF between No Accommodations and Accommodations - Nimble (Items 1, 16, 11, 23, 28, 29, 31 and 32) are also found in Appendix B. In general the magnitude of DIF appeared larger than observed for comparisons with Accommodations - Paper. Even when the stricter 3.0 ratio difference is applied, five items still exhibited DIF (Items 1, 16, 17, 28 and 31). Visual inspection of ICCs aligned with this result. ICCs among these items generally did show signs of uniform DIF, though the degree did vary. Items 16, 28, and 31 were found to be more difficult for No Accommodations test takers, while items 1 and 17 were more difficult for Accommodations - Nimble test takers. DIF was most apparent for items 1, 16, and 28 (see Figure 4.4, 4.5 and 4.6).

Figure 4.4. ICC for Item 1: No Accommodations vs. Accommodations - Nimble

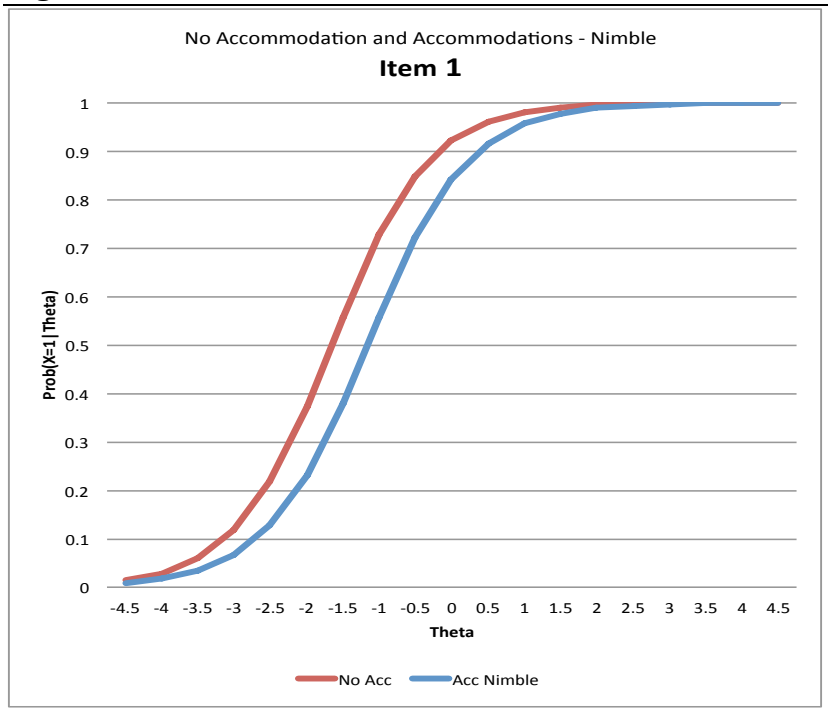


Figure 4.5. ICC for Item 16: No Accommodations vs. Accommodations - Nimble

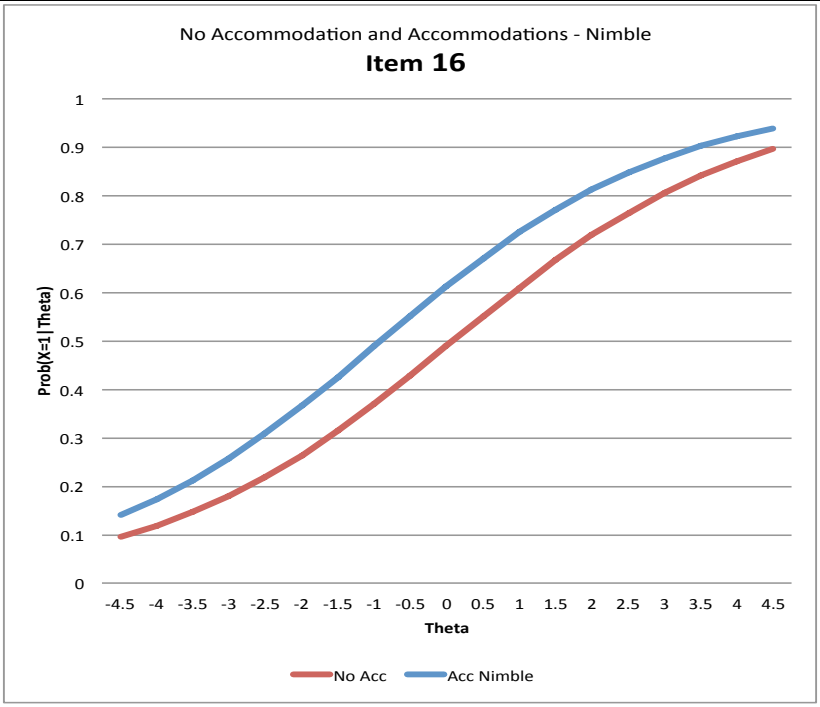
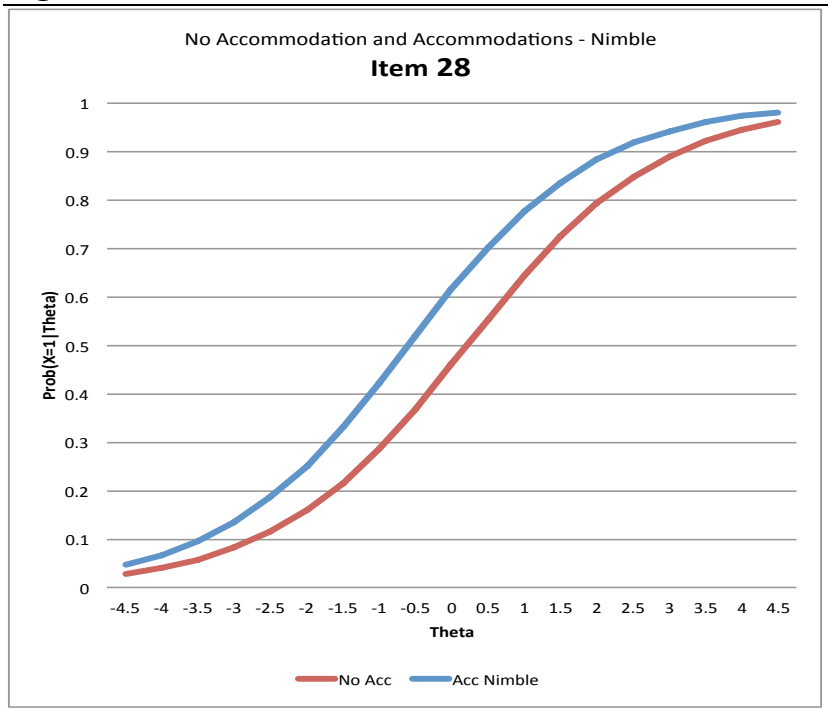


Figure 4.6. ICC for Item 28: No Accommodations vs. Accommodations - Nimble



Overall, visual inspection of ICCs suggested the more stringent criterion of 3.0 should be used to evaluate the difficulty difference to standard error ratios. When this criterion was applied, one item appeared to exhibit DIF for Accommodations - Paper test takers. This represented 3% of all multiple-choice test items. Similarly five items were found to exhibit DIF for Accommodations - Nimble test takers. This represented 15.2% of all multiple-choice test items administered. Bolt and Ysseldyke (2007) point out that published tests usually include less than 15% of DIF items and suggested this as a reference point. The proportion of items exhibiting DIF was well below this level for comparisons with Accommodations - Paper. Although 15% of items were found to exhibit DIF for comparisons with Accommodations - Nimble, DIF did not consistently favor one group over another and likely balanced out across the test. Some items were differentially difficult for No Accommodations test takers, while others were differentially difficulty for Accommodations - Nimble test takers.

RESEARCH QUESTION #3: IF DIFFERENTIAL ITEM FUNCTIONING IS EXHIBITED, DO PATTERNS OF DIF AND ITEM CHARACTERISTICS SUGGEST THAT ACCOMMODATIONS OR USE OF ACCESSIBILITY SUPPORTS MAY BE RELATED TO DIF?

Given information provided from DIF analysis, individual test items were examined to investigate why items might function differentially for groups of test takers. With the lens of accessibility in mind, the following item features and their interaction with accommodations and/or accessibility supports were considered: (1) item content (e.g. text length and complexity, target domain), (2) task demands (e.g. reasoning,

calculation, or interpretation of a table or graph), and (3) presentation features (e.g. presence of visuals or graphics, location on page/screen, text formatting). Table 4.17 summarizes characteristics of items that exhibited DIF.

Table 4.17. Summary of Item Characteristics Among Items Exhibiting DIF

	DIF Exhibited for:					
Item	Acc Paper v. No Acc.	Acc Nimble v. No Acc.	Domain	Word Count	Response Choices	Contains Visual Elements?
1	Yes (-)	Yes (-)	Physical Science	22	Text (2 words each)	Yes – Line Graph
16	No	Yes (+)	Earth and Space Science	31	Numeric (single and double digit with unit of measure	None
17	No	Yes (-)	Earth and Space Science	10	Text (1 word each)	None
28	No	Yes (+)	Life Science	28	Text (3 words each)	None
31	No	Yes (+)	Life Science	29	Numeric (fractions)	Yes – Figure with legend

(+) indicates an item was easier for accommodated students
(-) indicates an item was more difficult for accommodated students

Item 1 was found to exhibit DIF for both Accommodations - Paper and Accommodations - Nimble. Magnitude of DIF was greater for the latter group. This item was found to be more difficult for Accommodations – Nimble and Accommodations - Paper test takers. Aside from scientific vocabulary, in general language used throughout the NECAP assessment and for this item appeared appropriately simplified. This physical science item required the test taker to interpret a line graph. The graph included several visual details, such as grid lines, multiple graphed lines, arrows pointing to graphed lines from value labels and axis labels. Above the graph was a brief description. The item stem was in the form of a question and two word answer choices were presented

below. This item was positioned on the top left side of the paper form, along with two other items on the same page. One of these items also includes graphical elements.

One could speculate the visual features of this item and on this page may have proved distracting for test takers. However, there were other items with similar visual features and layout in later portions of the assessment that did not show this pattern of DIF. Using NimbleTools, all items were presented one at a time. Presumably this may have minimized distractibility, but in fact this item was found to be more difficult for NimbleTools test takers. Depending on the specific technology used for testing, screen resolution or size could have interfered with a test taker's ability to clearly view the graph presented along with the item stem and answer choices.

Item 17 was also found to exhibit DIF and was more difficult for Accommodations - Nimble test takers. No DIF was present for Accommodations - Paper students. This earth and space science item required understanding of scientific vocabulary related to stars. The item contained relatively limited text (10 words included in the item stem and one word answer choices) and no graphical components. This item was located on the top right of the paper form. Five other items were placed on the same page of the paper form. The specific source of DIF is unclear. There were no particular items features that one would expect to negatively interact with the NimbleTools computer-based testing interface or accessibility supports.

Items 16, 28, and 31 were also found to exhibit DIF for Accommodations - Nimble. These items were found to be easier for Accommodations – Nimble test takers. These items varied in terms of content, task demand and layout. They also did not appear to have features expected to negatively interact with NimbleTools features.

Item 16 related to earth space science and required the test taker to understand and perform a half-life calculation. The problem presented included numbers, scientific notation and scientific vocabulary. Item length was relatively brief (31 words across two sentences). Answer choices included single and double-digit numbers along with an abbreviation for a unit of measurement. This item was located on the top right of the paper form with two other relatively short items presented on the same page, resulting in ample white space on the remainder of the page.

Focusing on the domain of life science, item 28 required test takers to have knowledge of the parts of a cell, their function and relationship to genetics. Item length was relatively brief (28 words across two sentences). This included two bolded words. Students were presented with 3 word answer choices. This item was placed on the top right of the page for the paper form. Four other items are presented on this page. One of those items included a large table, taking up most of the right side of the page.

Lastly, item 31 was a life science item with a similar layout as item 1. Item 31 included a figure and legend. This item required the test taker to understand concepts specific to genetics, interpret the figure and identify the probability of an event in the form of a fraction. A brief text description containing information needed to solve the problem was found above the figure. Item stem and answer choices were placed below. For the paper form, this item took up the entire left side of the page. Two other items were placed on the right side of the page.

CHAPTER SUMMARY

The chapter presented the result of analyses, including confirmatory factor analysis (DIF) and differential item function (DIF). Results suggest this assessment generally did function similarly across NimbleTools and the non-accommodated condition. There was also evidence of similar underlying factor structure. Test function and underlying factor structure were also similar for accommodated students completing a paper-form. These results offer evidence of consistent measurement across accommodated and non-accommodated conditions

Overall, review of items exhibiting DIF did not suggest the systematic presence of item or test features expected to interact with NimbleTools supports in such a way that would alter item functioning or constructs measured. The few items found to exhibit DIF did not consistently favor any one group, nor did they seem to share many common item characteristics. There was no clear connection to DIF and NimbleTools features and item characteristics. Sources of item-level DIF seemed to vary from item to item. There was no evidence that the test as a whole functioned differently for accommodated students whether they completed testing with NimbleTools or a paper form. The final chapter of this dissertation will provide further discussion of these findings as well as limitations, considerations and directions for future research,

CHAPTER 5: DISCUSSION

The aim of this research was to explore the impact of implementing principles of UDL and accommodations through a technology-based format on test validity. This study analyzed test data for students attending high schools in New Hampshire, Vermont and Rhode Island who participated in the 2009 11th grade New England Common Assessment Program (NECAP) science assessment. Student data were collected during operational testing and used for another study examining the feasibility, effect and capacity to deliver state achievement tests using NimbleTools, a computer-based test delivery system with embedded testing accommodations designed using principles of UDL. Three test conditions were of interest: (1) no accommodations with a paper-based form (No Accommodations), (2) accommodated test administration with a paper-based form (Accommodations - Paper) and (3) accommodated test administration using a universally designed computer-based test delivery system with embedded accessibility supports (Accommodations - Nimble). This chapter reviews results and discusses implications, limitations, and recommendations for future research.

SUMMARY OF RESULTS

Research questions posed in this study centered on the validity of scores collected under accommodated and non-accommodated conditions. To explore these questions, descriptive analyses were conducted to provide context and a baseline for interpreting more complex results. Underlying factor structure was examined to determine the extent to which scores had consistent factor structures across accommodated and non-accommodated test conditions. Item difficulty and discrimination were examined for

equivalence between test conditions, after holding ability constant. Results from these analyses and implications are discussed below.

Accommodation Assignment

The type of accommodations assigned varied between students who used NimbleTools and those who completed the paper form. It may be the case that students with different access needs were assigned to NimbleTools, given the specific access features available and format of testing. For example, among accommodated students who completed a paper form, a very small number used reduction or visual print by blocking or other techniques (0.2%) and none used an acetate shield or visual magnification device. There were, however, some students who utilized comparable supports offered through NimbleTools, such as overlay, magnifier or masking. A number of students also used access features not available or offered for paper-based testing, such as reverse contrast, auditory calming or color choice.

In the specific instance of read aloud, 16% of accommodated students who completed the paper form received a read aloud. Among the NimbleTools features available to test takers, the most frequently assigned support was the read aloud (87.3%). For this latter group of students, this meant read aloud traditionally provided by a human access assistant was provided instead using an embedded technology-based read aloud. A significant benefit for schools is the reduced need for staff to provide support during testing. This type of read aloud may also be delivered through head phones and students may participate in group settings, instead of separate spaces. This means of delivery also avoids the need for potentially time-consuming and complex training to ensure consistent

provision of read aloud by staff. At the time of the study NimbleTools offered 18 accessibility tools. Only eight features were available for assignment during this study. This was perhaps a missed opportunity to provide a wider range of standardized supports for test takers.

Comparability of Underlying Constructs

The first research question posed concerned the consistency of the underlying factor structure as evidence of constructs measured across scores gathered under accommodated and non-accommodated conditions. Results from initial exploratory analysis were unclear and perhaps signaled flaws in the test overall. Scree plots suggested the possibility of a single factor model for all three groups, but total variance accounted for was low, ranging between 8.74% and 14.26%. When a single factor model was explored further, overall reliability across student groups was good at .80. Reliability was also found to be good for non-accommodated students ($\alpha = .80$). Reliability was slightly lower for accommodated students who completed paper-based testing ($\alpha = .78$). In contrast, reliability for Accommodations - Nimble was questionable ($\alpha = .70$).

Continuing to pursue the possibility of a single factor model likely representing a general measure of scientific knowledge and skills, a series of confirmatory factor analysis were conducted. Ultimately, a single factor model was confirmed for both single group and multi-group comparisons, though a small number of items did not have significant loadings for Accommodations - NimbleTools. Results across nested models generally suggest that factor structure is invariant across accommodated and non-

accommodated test conditions. This was the case for both comparisons made with accommodated students completing a paper form and those receiving supports through NimbleTools. These results were similar to previous studies that found evidence of factor invariance among accommodated and non-accommodated conditions, though previous research appeared limited to evaluations of read aloud accommodations (Cook et. al, 2006; Harris, 2008; Cook et al. 2010, Huynh & Barton, 2006; Kim, Schneider & Siskind 2009; Pomplun & Omar 2000).

Overall, these results suggested that a similar construct was being measured even when students were accommodated through embedded technology-based supports in the place of or in addition to traditional accommodations. However, findings from exploratory analyses suggested that the specific construct measured might have not been well defined, regardless of test condition. Lower reliability for NimbleTools may not signal differences in construct, but may indicate increased measurement error. For example, lack of familiarity among test takers in how to effectively use supports during testing or navigate the testing interface could have contributed error to the measurement process. The quality of technology used for testing may also introduce error, lowering reliability. For example, low screen resolution could interfere with a test taker's ability to clearly view graphs and other visuals. Poor audio either through speakers or headphones could interfere with a read aloud, diminishing the test takers ability to fully access test content.

Consistency in Item Functioning

The second research question posed concerns item functioning across accommodated and non-accommodated conditions. Specifically, holding ability constant, was item difficulty and discrimination equivalent for accommodated students and non-accommodated students? The presence of a large number of items exhibiting differential functioning favoring one group over another could suggest differences in overall test functioning and perhaps differences in the underlying scale.

In this case, the majority of items appeared to function similarly across test conditions. A small number of items did show signs of DIF. This was the case for slightly more items when comparisons where made with students receiving accommodations and testing with NimbleTools. Analysis found one item exhibiting DIF for Accommodations - Paper and five items for Accommodations - Nimble. However, items did not consistently favor one test condition. Although previous studies have used different criteria for identifying DIF, findings from this study are similar to previous research.

Few previous studies examined items to identify potential causes of DIF. This was likely a consequence of test security concerns and restricted access to actual items. However, examination of flagged items is needed to make more conclusive statements about the possible reasons for differential functioning and the role accommodations may play in item functioning. In this study, flagged items were available for review. This examination did not suggest the presence of specific features or item characteristics related to accessibility or the provisions of accommodations as a clear source of DIF.

OVERVIEW OF FINDINGS AND IMPLICATIONS

It was hypothesized that NimbleTools would address accessibility issues comprehensively and effectively for students with diverse needs. Principles of UDL applied during test development along with the use of accommodations and other supports were intended to minimize sources of construct irrelevant variance related to accessibility barriers for these students. If these efforts were successful and measurement was consistent for all students, we would expect that scores for test takers using NimbleTools would show psychometric properties similar to non-accommodated conditions. Overall, results suggest this assessment generally did function similarly across NimbleTools and non-accommodated conditions. There was also evidence of similar underlying factor structure. Test function and underlying factor structure were also similar for accommodated students completing a paper-form. These results offer evidence of consistent measurement across accommodated and non-accommodated conditions

These results also support the viability of using technology-based assessments as a valid means of assessing students and offering embedded, standardized supports to address access needs. This approach offers the advantage of multiple accessibility features within a single interface, eliminating the need for multiple versions and in some cases, human access assistants. This offers the opportunity for potential cost saving for states and other large-scale test programs in terms of physical test materials, staffing and training. In terms of measurement, technology-based accommodations can be offered with more consistency across test takers.

These results also have implications for large-scale assessment systems that are considering a hybrid approach to testing, administering both technology-based and paper-based versions of the same assessments. This approach may be taken if it is not feasible to universally implement technology–based test delivery across a consortium, state or school. For example, the state of Massachusetts is currently phasing in computer-based testing, replacing a paper assessment. During this transition, districts and schools are permitted to assess some grades using technology-based tests and others using paper assessments. The NECAP states are undergoing a similar transition, phasing in technology-based mathematics and reading assessments as part of the Smarter Balanced Assessment Consortium (SBAC). This presents a challenge for test developers who must design items that will be valid for both testing formats. These results suggest it is possible.

These findings are offered with a small caveat with regards to specific item level differences found during analyses. There was evidence of slight differences for NimbleTools testing takers in the form of lower reliability and a few items with non-significant factor loadings and evidence of DIF. This suggests there is perhaps room for improvement in terms of test and item design to ensure consistent measurement for all items. Individual items must be carefully evaluated to ensure consistency across technology-based and paper-based forms, both with and without accommodations applied. Previous research does suggest that computer-based testing imposes different demands than paper-based tests (e.g. typing, scrolling through multiple screens, recalling information not currently displayed on screen, reading from a screen) (Hollenbeck et al, 1999; Ommerborn & Schuemer, 2001). Items administered with NimbleTools during the

2009 11th grade NECAP science assessment were originally designed for a paper-based assessment. It is possible these particular items did not transfer as expected to the technology-based form. A review of these items did not readily suggest why this would be the case. Other research methods, such as cognitive interviews, may offer better insights of the potential interaction between computer-based test delivery and specific item design.

Item level differences do have implications for item level reporting, a common practice for state assessment programs. Although generally not conducted under high stakes conditions, item analysis is another tool for educators and decision makers to both reflect on their practice and identify specific needs among students. Educators are often encouraged to review item level results for these purposes from state tests similar to those analyzed in this study. In cases where item level differences exist among scores collected under technology-based and accommodated conditions, this could lead to misleading interpretations.

To improve the overall quality of assessments, one contribution of this study is to encourage large-scale test programs to employ similar analytic techniques to explore the impact of test accommodations and to ensure problematic items are identified. The methods employed for this study serve as an informative complement to other approaches, such as cognitive interviews, which can provide more detailed information about how test takers interact with items and the assessment as a whole. Used together, test developers would be better equipped with information to inform development of instruments that consider accessibility, the design of specific access supports and their interaction with specific items. It is important to note that results from CFA and DIF

contradicted results of classical test statistics. Sole reliance on the latter could lead to misleading conclusions about assessments and their functioning across different test takers.

Finally, it is common practice for states to report sub-score or subdomain results on student level reports. As described in chapter three, four domains were purportedly measured by the 11th grade NECAP science assessment: science process skills, earth space science, life science and physical science. Although analysis excluded constructed response items, no clear factor structure emerged suggesting measurement of these domains during exploratory or confirmatory analysis. Absence of this structure suggests caution should be taken with regards to the practice of reporting sub-scores intended to describe student performance in these areas.

LIMITATIONS AND CONSIDERATIONS

Although experimental designs are generally preferable in accommodations research, random assignment to test conditions was not possible for this study, which used operational test data. There are some benefits to non-experimental accommodations research, including large sample sizes and real world test conditions. The benefit of large sample sizes in this case permits greater generalizability and increased power. This is especially important for IRT DIF and CFA as they require relatively large samples to achieve stable estimates. Further, under these conditions, student test performance will have real consequences for test takers and schools. This better represents normal test conditions to which researchers aim to generalize. Finally, test takers may also be more likely to be provided with accommodation packages tailored to their individual needs.

Accommodation assignment in this case was based on the decision of instructional teams who work closely with students throughout the year.

In spite of the benefits just described, there are also a number of limitations. First, the use of a non-experimental design limits the ability to make causal statements about the impact of accommodations and the manner in which they are provided. It is also difficult to disentangle the impact of a particular accommodation or technology-based support among students who received multiple accommodations. However, attempting to isolate the effect of single accommodations by assigning participants only one type of accommodation for research may leave students' access needs unmet and mask the effect of accommodations. Students may face access challenges if denied all the necessary supports, interfering with their ability to demonstrate what they know and can do. In this study, both accommodated students completing the paper form and those using NimbleTools often received two or more accommodations

Analysis in this study did not differentiate between categories of disabilities or types of accommodations. This is a common characteristic of accommodations research due to the relatively small size of the population who use accommodations and the even smaller size of subsets of students by disability, accommodation or combination of accommodations. In addition, NECAP policy allowed supports to be assigned to a student regardless of disability status. A number of students who had not been identified with a disability were assigned one or more supports in this study. Among test takers with identified disabilities, no information was provided about specific disability type. This contributes to the difficulty of examining the effect of the supports by sub-groups. Nevertheless, differentiation within the population is important because the group in

question is not heterogeneous, having a diverse set of access needs and often uses a diverse combination of supports. Analyzing data for specific disability groups or accommodation combinations may lead to additional insights.

There is also the danger that non-experimental data can be more easily confounded by other variables. That may include variation in accommodation assignment procedures between schools and the exact provision of accommodations. Although NECAP states were provided with general guidance on how to make accommodations decisions, actual implementation can vary by school. Another consideration was the challenge of ensuring appropriate and consistent administration and provision of accommodations. A flawed read aloud, for example, provided by a poorly trained access assistant could impact a student's performance. This would have confounded analysis in instances where read aloud was provided by a proctor for those completing the paper assessment, This was less of a concern among NimbleTools test takers as this delivery mode offered the benefit of more standardized supports.

Another limitation was a possible self-selection bias present at the school level. All schools in this study volunteered to use NimbleTools. Schools may have differed in meaningful ways (e.g. staff attitudes toward technology or available school infrastructure to support computer-based administration) that may have directly and/or indirectly impacted student performance. In general, it is known that schools opted to use or not use NimbleTools for a variety of reasons. School staff may not have felt prepared to administer a technology-based assessment due to poor infrastructure. Some schools may have been unaware of the option. In other cases, school staff may not have seen this option as a benefit to their students. Among schools that did opt to use NimbleTools,

staff may have seen it as an opportunity to provide support to students not easily provided through traditional accommodations. Unfortunately school level data was not available for this analysis. So this could not be fully explored.

Based on accommodation assignment, there seemed to be differences between students assigned to use NimbleTools and those accommodated with a paper form. It was expected that only those students whose access needs would have been best met with NimbleTools and its features would have been assigned. Specifically, the percentage of students with an IEP assigned to NimbleTools was noticeably larger than those assigned paper-accommodations. In addition, a much higher percentage of students using NimbleTools were assigned the read aloud support compared to students completing the paper form. NimbleTools may have made it more feasible to provide certain accommodations, including read aloud. It is possible that NimbleTools made it more feasible for schools to provide the read aloud support since it was delivered by computer software rather than by school staff. Given the resource intensive nature of a proctored read aloud, it is also possible that some schools that opted out of NimbleTools may not have had the resources to provide a read aloud to all students who might benefit. Although this could have implications for individual scores, there was no evidence to suggest this systematically occurred or there was a significant impact on overall test functioning or underlying factor structure.

In contrast, given NECAP's policy of allowing students to receive an accommodation based on need rather than IEP status, it is possible that the ease with which NimbleTools provided read aloud support led some schools to assign this support more often. There were also differences observed for visual supports, such as colored

overlay and use of a magnifier. There were some supports, such as auditory calming available with NimbleTools, but not available to students who completed the paper-form. Conversely, NimbleTools did not support some accommodations and accessibility supports (e.g. student dictates response, ability to underline or otherwise mark up text, non-English translation). Also some NimbleTools features were not available during this study (e.g. sign directions, electronic braille display interactivity*)*. Therefore students requiring these accommodations would not have been assigned to NimbleTools. Again, this could have implications for individual scores, but results do not suggest this had a significant impact on overall test functioning or underlying factor structure.

Among schools that elected to use NimbleTools, variation in the availability of reliable technology and its daily use during instruction may have also impacted a students' ability to complete the online assessment and effectively use technology-based supports. In an effort to minimize this potential impact, all students assigned to Nimble Tools were asked to complete online practice tests and training modules before testing. This was both used to identify specific supports that might improve test accessibility, but to also familiarize students with the general technology-based interface and experience of using accessibility tools.

Finally, researchers and test developers should consider how accommodations were operationally defined for this study. How accommodations are defined and provided can vary from state-to-state and among studies. For example, a read aloud accommodation may refer to a read aloud of directions only, but in other instances refer to a read aloud of directions and item stems, or directions, item stems and answer

choices. Operational definitions of accommodations should be considered with respect to the generalizability of results.

DIRECTIONS FOR FUTURE RESEARCH

Validity must be considered from many perspectives and therefore complimentary studies (e.g. cognitive lab studies, experimental research) are needed to further investigate the impact of technology-based accommodations. This will be critical as the number of large-scale technology-based assessment programs increases. Both the Smarter Balanced Assessment Consortium (SBAC) and the Partnership for Assessment of Readiness for College and Careers (PARCC) have developed and piloted computer-based assessments with embedded accessibility tools as well as locally provided accommodations (PARCC, 2014, SBAC, 2014). Both also reference universal design principles in their test design frameworks (Measured Progress/ETS Collaborative, 2012; PARCC, 2016).

Though not available for this study, the use of technology-based assessment systems may offer the opportunity to collect and analyze more fine grained information about the actual use of supports during testing. Accommodations are assigned at the test level, but students may or may not make use of all accommodations for all items. For example, a student may only use read aloud for some, but not all items. Most research assumes accommodations are being used consistently throughout testing, but the actual accommodation applied may differ from item to item. While administering technology-based assessment tools, it may be possible to capture test taker use of specific tools throughout testing. This could help answer questions about how often an accommodation

or accessibility feature is activated during testing, on which items or portions of an assessment are features being used, what features are assigned but not used during testing, etc.

Furthermore, other details of technology-based assessment and the provision of accessibility related supports needs to be explored. What role does device type play? How do device characteristics (e.g. peripherals, screen size, resolution, speaker or headphone quality) interact with accessibility features? What kind of device may be problematic or do well to address which type of access needs?

Another area for further explorations concerns the basic assumption that technology is being used increasingly during instruction and that students have enough experience to use technology during high-stakes testing. It is assumed that integration of the same technology-based supports on large-scale assessments would lead to better continuity between instruction and statewide assessments (Dolan et al, 2005). This would also result in better alignment with student preferences for technology-based testing over traditional paper-and-pencil assessment (Thompson, Thurlow & Moore, 2003). To what extent does this hold true and how does use of technology during instruction relate to use of technology during assessment? Does this look the same for students with special needs?

A related area in need of further study concerns technology infrastructure. Recent attempts to implement state-wide technology-based assessment systems have raised questions about the adequacy of technology infrastructure in schools. As mentioned earlier both PARCC and SBAC test consortiums have piloted online technology-based assessments. This approach assumes schools have the needed infrastructure to administer

online assessments with fidelity and consistency. SBAC (2014) reported that state readiness for computer-based testing varied significantly across states and schools. They also noted variation in the "readiness of adults to administer them." Since the formation of these consortia a number of states have dropped out, citing challenges with technology infrastructure.

Further, although PARCC and SBAC both reported the absence of system-wide issues during pilot testing, local technology issues were noted. Internet connectivity, particularly for devices using wireless connections was cited as a specific example during the SBAC field test (SBAC, 2014). From the perspective of a multi-state assessment program, this kind of technological challenge may not be viewed as a system-wide issue. However, this can disrupt or distract the test taker and may ultimately impact validity. This potential negative impact should be ruled out for test takers both with and without access needs.

Those systems and states that are proceeding with implementation of large-scale technology-based assessments may offer researchers a wealth of data and new opportunities to explore the impact of technology based supports and validity across test takers with diverse access needs. Both the Smarter Balanced Assessment Consortium (SBAC) and Partnership for Assessment of Readiness for College and Careers (PARCC) offer valuable opportunities to increase research that could guide states' and districts in moving forward with technology-based assessments and the use of embedded supports. Given their scale, this could also present opportunities to differentiate between categories of disabilities or types of accommodations. This analysis is often hindered by the relatively small size of the population of students who are assigned accommodations.

Further discussion is also needed around the concept of accessibility features and how this relates to accommodations. What should be available to all? What is restricted or selectively used? These decisions should be driven by the relationship with the intended construct. In some cases, what is feasible to implement during operational testing may also play a role. These questions were raised by researchers from CAST in their review of PARCC's draft accommodations and accessibility framework. CAST (Hall et al. 2013) researchers raised the question, "if students' IEP/504 teams are to determine which accessibility features these students will use, how do accessibility features differ from other accommodations?" When framed this way, this is potentially a legal, philosophical and procedural question. Further research is needed to guide decision makers in developing policy and for practitioners making decisions for individual students.

Finally on the matter of consequential validity, as systems proceed down the path towards increased use of technology-based assessments and presumably increased use of technology-based accommodations, the intended and unintended impacts on curriculum and instruction should be monitored, including the supports offered to students outside of testing. Again researchers from CAST caution:

> We see a danger of assessment policy and procedures driving instructional practices, including materials and tools, such as accessible instructional materials, used for students in the classroom. While it is very clear and we agree that accommodations may interfere with a construct being measured at the item level, we are concerned that schools and/or teachers may not allow accommodations for instruction because they may not be allowed on the assessment (Hall et al. 2013a).

CONCLUSIONS

Although research examining the impact of test accommodations on validity has grown, results have been mixed. Among the available research, science assessments have been less frequently studied. There has also been less research on accommodation use among high school students, with middle school and elementary populations studied more frequently. More research is also needed to explore application of UDL to assessment and the impact of technology-based accommodations. This dissertation attempts to begin to address these gaps. Results from this study offered evidence that overall item functioning and underlying factor structure was consistent across accommodated and unaccommodated conditions, regardless of whether accommodations were provided with a paper form or within a universally designed computer-based test delivery system. These results support the viability of using technology-based assessment as a valid means of assessing students and offering embedded, standardized supports to address access needs.

REFERENCES

Abedi, J., Hofsetter, C., Baker, E. & Lord, C. (2001). *NAEP math performance and test accommodations: Interactions with student language background* (Report No. 536). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Abedi, J., Leon, S., & Kao, J. C. (2008). *Examining differential distractor functioning in reading assessments for students with disabilities*. (Report No. 743). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Abedi, J., Leon, S., & Mirocha, J. (2003). *Impact of student language background on content-based performance: Analyses of extant data* (Report No. 603). Los Angeles: University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Altman, J., Thurlow, M., & Vang, M. (2010). *Annual performance report: 2007-2008 state assessment data.* Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

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

American Education Research Association, American Psychological Association, & National Council on Measurement in Education (2014). *The standards for educational and psychological testing.* Washington, D.C.: American Educational Research Association.

American Recovery and Reinvestment Act (ARRA) of 2009, Pub. L. No. 111-5, 123 Stat. 115, 516 (2009).

Assistive Technology Act of 2004 (Brief Title: ATA 2004). (P.L.108-364).

Bechard, S. (2000). *Students with disabilities and standards-based reform*. Aurora, CO: Mid-continent Research for Education and Learning.

Bechard, S., Almond, P., & Cameto, R. (2011). Item and test alterations: Designing and developing alternate assessments with modified achievement standards. In M. Russell & M. Kavanaugh (Eds.), *Assessing students in the margins: Challenges, strategies and Techniques.* Charlotte, NC: Information Age Publishing.

Beddow, P.A. (2011). Beyond universal design: Accessibility theory to advance testing for all students. In M. Russell & M. Kavanaugh (Eds.), *Assessing students in the*

margins: Challenges, strategies and Techniques. Charlotte, NC: Information Age Publishing.

Beddow, P. A., Elliott, S. N., & Kettler, R. J. (2009). *TAMI Accessibility Rating Matrix.* Nashville, TN: Vanderbilt University.

Beddow, P. A., Elliott, S. N., & Kettler, R. J. (2013). Test accessibility: Item reviews and lessons learned from four state assessments. *Education Research International*, 2013, 1–12.

Bennett, R. E. (1998). *Reinventing assessment: Speculations on the future of large-scale educational testing*. Princeton , NJ : Educational Testing Service, Policy Information Center. Available: http://www.ets.org/research/pic/bennett.html.

Bentler, P.M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, 107 (2), 238-46.

Bielinski, J., Thurlow, M., Ysseldyke, J., Freidebach, J., & Freidebach, M. (2001). *Read-aloud accommodations: Effects on multiple-choice reading and math items* (Technical Report 31). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved March 2, 2010, from the World Wide Web: http://education.umn.edu/NCEO/OnlinePubs/Technical31.ht.

Bolt, S.E. (2004). *Using DIF Analyses to Examine Several Commonly-Held Beliefs about Testing Accommodations for Students with Disabilities*. Paper Presented at the annual conference of the National Council on Measurement in Education.

Bolt, S. E., & Ysseldyke, J. (2007). Accommodating students with disabilities in large-scale testing: A comparison of differential item functioning identified across disability types. *Journal of Psychoeducational Assessment,* 26(121), 121-138.

Brown, P. B., & Augustine, A. (2000). *Findings of the 1999-2000 screen reading field test*. Delaware Department of Education.

Burk, M. (1999). *Computerized test accommodations: A new approach for inclusion and success for students with disabilities*. Washington, DC: A.U. Software, Inc.

Calhoon, M. B., Fuchs, L. S., & Hamlett, C. L. (2000). Effects of computer-based test accommodations on mathematics performance assessments for secondary students with learning disabilities. *Learning Disability Quarterly, 23*(4), 271-281.

CAST. (2016). *UDL at a Glance*. Retrieved from http://www.cast.org/our-work/about-udl.html#.WNwv-P21uAw.

Cattell, R.B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research,* 1, 245-276.

Center for Universal Design. (1997). *About UD: Universal design principles*. Retrieved February 20, 2010, from http://www.design.ncsu.edu/cud/about_ud/udprincipleshtmlformat.html. Archived at http://www.webcitation.org/5eZBa9RhJ.

Cheung, G. & Rensvold, R. (2002). Evaluating goodness of fit indexes for testing measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal,* 9 233-245.

Choi, S. W., & Tinker, T. (2002). *Evaluating comparability of paper-and-pencil and computer based assessment in a K-12 setting.* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Chudowsky, N., Chudowsky, V., & Kober, N. (2009). *State test score trends through 2007-2008, part 3: Are achievement gaps closing and is achievement rising for all?* Retrieved from Center on Education Policy website: http://www.cep-dc.org/publications/index.cfm?selectedYear=2009

Clapper, A. T., Morse, A. B., Thompson, S. J., Thurlow, M. L. (2005). *Access assistants for state assessments: A study of state guidelines for scribes, readers, and sign language interpreters* (Synthesis Report 58). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Cohen, A. S., Gregg, N., & Deng, M. (2005). The Role of Extended Time and Item Content on a High-Stakes Mathematics Test. *Learning Disabilities Research and Practice*, 20 (4), 225 -233.

Cook, L. Eignor, D. Steinberg, J. Sawaki, Y. & Cline, F. (2006). *Using factor analysis to investigate the impact of accommodations on the scores of students with disabilities on a reading comprehension assessment*. Educational Testing Service.

Cook, L., Eignor, D., Sawaki, Y., Steinberg, J., & Cline, F. (2010). Using factor analysis to investigate accommodations used by students with disabilities on an English-Language arts assessment. *Applied Measurement in Education*, *23*(2), 187-208.

Costello, A. B. & Osborne, J. (2005). Best practices in exploratory factor analysis: four recommendations for getting the most from your analysis. *Practical Assessment Research & Evaluation*, 10(7). Available online: http://pareonline.net/getvn.asp?v=10&n=7

Crawford, L. (2007). *State testing accommodations: A look at their value and validity.* New York, NY: National Center for Learning Disabilities.

Devillis, R.F. (2003). *Scale development: Theory and applications (2nd ed.)*. Thousand Oaks, CA: Sage Publications, Inc.

Dolan, R. P. & Hall, T. E. (2001). Universal design for learning: Implications for large-scale assessment. *IDA Perspectives* 27(4), 22-25.

Dolan, R. P., Hall, T. E., Banerjee, M., Chun, E., & Strangman, N. (2005). Applying principles of universal design to test delivery: The effect of computer-based read-aloud on test performance of high school students with learning disabilities. *Journal of Technology, Learning, and Assessment, 3*(7). Available from http://www.jtla.org.

Dorans, N.J., & Holland, P.W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P.W. Holland, & H. Wainer (Eds.), *Differential item functioning* (p. 35-66). Hillsdale, NJ: Lawrence Earlbaum.

Driscoll, D. P. (2007). *Requirements for the participation of students with disabilities in MCAS*. Malden, MA: Massachusetts Department of Education.

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 assessment. *School Psychology Quarterly*, 24(4), 224 - 239.

Elliot, S. N., Kratochwill, T.R., & Schulte, A. G. (1999). *Assessment accommodations checklist.* Monterey, CA: CTB/McGraw Hill.

Every Student Succeeds Act of 2015, Pub. L. 114-95 § 114 Stat. 1177 (2015).

Finch, H., Barton, K., & Meyer, P. (2009). Differential item functioning analysis for accommodated versus nonaccommodated students. *Educational Assessment,* 14, 28-56.

Fletcher, J.M., Francis, D.J., Boudosquie, A., Copeland, K., Young, V., Kalinowski, S. & Vaughn, S. (2006). Effects of accommodation on high-stakes testing for students with reading disabilities. *Exceptional Children*, 72(2), 136-150.

Garson, G. D. (2011). Factor analysis. From *Statnotes: Topics in Multivariate Analysis.* Retrieved March 10, 2010 from http://faculty.chass.ncsu.edu/garson/pa765/statnote.htm.

George, D., & Mallery, P. (2003). *SPSS for Windows step by step: A simple guide and reference*. 11.0 update (4th ed.). Boston: Allyn & Bacon

Hafner, A. (2001). *Evaluating the impact of test accommodations on test scores of LEP students & non LEP students*. Paper presented at the annual meeting of the American Educational Research Association.

Haladyna, T.M. & Downing, S.M. (2004). Construct-irrelevant variance in high stakes testing. *Educational Measurement: Issues and Practices,* 23(1), 17-27.

Hall, T. E., Hitchcock, C., Jackson, R., Karger, J., Ralabate, P., Rose, D. H. & Zabala, J. (2013, May 12). CAST's Response to the PARCC Accommodations Manual [Letter to Partnership for Assessment of Readiness for College and Careers (PARCC)]. CAST, Wakefield, MA.

Hall, T. E., Hitchcock, C., Jackson, R., Karger, J., Ralabate, P., Rose, D. & Zabala, J. (2013a, June 12). CAST's Response to the Smarter Balanced Assessment Consortium's Draft of Accessibility and Accommodations Guidelines [Letter to Smarter Balanced Assessment Consortium]. CAST , Wakefield, MA.

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

Harker, J.K. & Feldt, L.S. (1993). A comparison of achievement test performance of nondisabled students under silent reading and reading plus listening modes of administration. *Applied Measurement in Education*, 6 (4), 307 -320.

Harris, L. W. (2008). *Comparison of student performance between teacher read and CD-ROM delivered modes of test administration of English language arts tests*. Dissertation Abstracts International.

Helwig, R., Rozek-Tedesco, M.A., & Tindal, G. (2000). *An oral versus standard administration of a large-scale mathematics test*. Newark, DE: Delaware Education Research and Development Center, University of Delaware.

Helwig, R., Rozek-Tedesco, M.A., Tindal, G., Heath, B. & Almond, P. (1999). Reading as an access to mathematics problem solving on multiple-choice tests for sixth-grade students. *The Journal of Educational Research*, 93 (2), 113 -125.

Hollenbeck, K., Tindal, G., Harniss, M., & Almond, P. (1999). Reliability and decision consistency: An analysis of writing mode at two times on a statewide test. *Educational Assessment, 6* (1), 23-40.

Hoyle, R. H., & Panter, A. T. (1995). Writing about structural equation models. In R. H. Hoyle, *Structural equation modeling: concepts, issues, and applications*. Newbury Park, CA: Sage.

HumRRO. (2003). *CATS Online: Logistic and Construct Evaluation of Computer Administered Assessment* (No. FR-03-40). Alexandria, VA: Human Resources

Research Organization.

Huynh, H. & Barton, K. E. (2006). Performance of students with disabilities with requlars and oral administrations of high-stakes reading examination. *Applied Measurement in Education*, 19(1). 21-39.

Higher Education Amendments Act of 1998, Pub. L. 105-244. (1998).

Individuals with Disabilities Educational Improvement Act (Brief Title: IDEA 2004). (P.L. 108-446).

Johnson, E.S., Kimball, K., Brown, S.O. & Anderson, D. (2001). A statewide review of the use of accommodations in large-scale, high-staked assessments. *Exceptional Children*, 67(2), 251 -264.

Johnstone, C. J. (2003). *Improving validity of large-scale tests: Universal design and student performance* (No. Technical Report 37). Minneapolis, MN: University of Minnesota, National Center on Education Outcomes.

Johnstone, C.J., Anderson, M.E., & Thompson, S. J. (2006). Universally designed assessments for ELLs with disabilities: What we've learned so far. *Journal of Special Education Leadership, 19* (1), 27-33.

Johnstone, C., Thurlow, M., Moore, M., & Altman, J. (2006). *Using systematic item selection methods to improve universal design of assessments* (Policy Directions 18). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Jöreskog, K.G. (*1969*). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika*, 34, 183–202.

Kane, M.T. (2006). *Validation*. In R.L. Brennan (Ed.) *Educational Measurement (4th ed.)*.(pp.17-64). Westport, CT: The American Council on Education and Praeger Publishers.

Kavanaugh, M. & Russell, M.. (2011). Examining RFP requirements. In M. Russell & M. Kavanaugh (Eds.), *Assessing students in the margins: Challenges, strategies and Techniques.* Charlotte, NC: Information Age Publishing.

Kettler, R.J., Niebling, B.C., Mroch, A.A., Feldman, E.S., &Newell, M.L. (2003). *Effects of testing accommodations on math and reading scores: An experimental analysis of the performance of fourth and eighth grade students with and without disabilities*. Madison, Wisconsin: Wisconsin Center for Education Research, University of Wisconsin-Madison.

Ketterlin-Geller.L.R. (2003). *Establishing a validity argument for universally designed assessments.* Unpublished Doctoral Dissertation, University of Oregon, Eugene, OR.

Ketterlin-Geller, L. R. (2005). Knowing what all students know: Procedures for developing universal design for assessment. *Journal and Assessment*, 4(2). Available from http://www.jtla.org.

Ketterlin-Geller, L.R. (2008). Testing students with special needs: A model for understanding the interaction between assessment and student characteristics in a universally designed environment. *Educational Measurement*: *Issues and Practice*, 27(3), 3-16.

Kim, D., Schneider, C., & Siskind, T. (2009). Examining the underlying factor structure of a statewide science test under oral and standard administrations. *Journal of Psychoeducational Assessment*, 27(4), 232-333.

Kim, D., Schneider, C., & Siskind, T. (2009a). Examining equivalence of accommodations on a statewide elementary-level science test. *Applied Measurement in Education*, 22 (2), 144-163.

Kopriva, R., Emick, J., Hipolito-Delgado, C., & Cameron, C. (2007). Do proper accommodation assignments make a difference? Examining the impact of improved decision making on scores for English language learners. *Educational Measurement: Issues and Practice, 26*(3), 11-20.

Koretz, D. (1997). *The assessment of students with disabilities in Kentucky*. (Report No. 431). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Koretz, D. & Hamilton, L. (1999). *Assessing students with disabilities in Kentucky: The effects of accommodations, format, and subject*. (Report No. 498). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Koretz, D., & Hamilton, L. (2001). *The performance of students with disabilities on New York's Revised Regents Comprehensive Examination in English.* (Report No. 540). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Kosciolek, S. & Ysseldyke, J.E. (2000*). Effects of a reading accommodation on the validity of a reading test* (Report No. 28). Minneapolis, MN: National Center on Educational Outcomes, University of Minnesota.

Lang, S. C., Elliott, S. N., Bolt, D. M., & Kratchowill, T. R. (2008). The effects of testing accommodations on student's performances and reactions to testing. *School Psychology Quarterly*, 23(1), 107 -124.

Long, J. S. 1983. *Confirmatory factory analysis*. Beverly Hills, CA: Sage.

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

MacArthur, C. (1996). Using technology to enhance the writing processes of students with learning disabilities. *Journal of Learning Disabilities*, 29(4), 344-354.

Maihoff, N. (2000). *Effects of administering an ASL signed standardized test via DVD player/television and by paper-and-pencil: A pilot study*. Newark, DE: Delaware Education Research and Development Center, University of Delaware.

Messick, S. (1989). Validity. In Linn, R.L. (Ed). *Educational measurement* (3rd ed.). The American Council on Education/Macmillan series on higher education (pp.13-103). New York, NY, England: Macmillan Publishing Co, Inc; American Council on Education.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, 23(2), 13-23.

Messick, S. (1995). The validity of psychological assessment: Validation of inferences from persons' responses and performance as scientific inquiry into score meaning. *American Psychologist,* 50, 741-749.

McKevitt, B. C. (2001). The effects and consequences of using testing accommodations on a standardized reading test. (Doctoral dissertation, University of Wisconsin-Madison). *Dissertation Abstracts International*. (UMI 3020768)

Measured Progress & National Center on Educational Outcomes. (2014). *Smarter balanced assessment consortium: Accessibility and accommodations framework*. Retrieved from http://www.smarterbalanced.org/wp-content/uploads/2015/09/Accessibility-and-Accommodations-Framework.pdf.

Middleton, K. & Laitusis, C. C. (2007). *Examining test items for differential distractor functioning among students with learning disabilities*. Princeton, NJ : Educational Testing Service.

Miller, T.R., & Spray, J.A. (1993). Logistic discriminant function analysis for DIF identification of polytomously scored items. *Journal of Educational Measurement, 30.* 107-122.

Miranda H., Russell M., & Hoffman T. (2004). Examining the feasibility and effect of a computer-based read-aloud accommodation on mathematics test performance: Part of the New England Compact Enhanced Assessment Project. Retrieved from http://www.bc.edu/research/intasc/researchprojects/enhanced_assessment/PDF/EAP_VerbalResponse.pdf.

National Center for Educational Outcomes. (2009) *Accommodations for students with disabilities*. Retrieved February 9, 2010 from http://www.cehd.umn.edu/NCEO/TopicAreas/Accommodations/Accomtopic.htm.

National Center for Educational Outcomes. (2016). *APR Assessment Snapshot Data.* Retrieved December 1, 2016 from https://nceo.info/Resources/publications/APRsnapshot/data.

New Hampshire Department of Education, Rhode Island Department of Elementary and Secondary Education, & Vermont Department of Education. (2009) *New England common assessment program test administrator manual –science Grade 11*. Retrieved February 10, 2010, from http://www.education.nh.gov/instruction/assessment/necap/admin/documents/gr11science_manual09.pdf.

Nimble Assessment Systems. (2008). *The universal assessment system: Examining the validity of inferences about achievement for students with disabilities and special needs.* Unpublished manuscript.

Nimble Assessment Systems (2008a). *NimbleTools*. Retrieved March 10, 2010 from http://nimbletools.com/uas/testingGear.htm.

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

Office of Educational Technology. (2016). *Future ready learning: Reimagining the role of technology in education*. Washington D.C.: U.S. Department of Education.

Ommerborn, R., & Schuemer, R. (2001). *Using computers in distance study: Results of a survey amongst disabled students*. Hagen, Germany: FernUniversität.

PARCC. (2014). *PARCC field test: Lessons learned*. Retrieved November 1, 2016, from the World Wide Web: http://www.parcconline.org/files/81/Field%20Test/94/field-test-lessons-learned-final_0-2.pdf

PARCC. (2014). *PARCC accessibility features and accommodations manual*. Retrieved August 2, 2016, from the World Wide Web: http://avocet.pearson.com/PARCC/Home#10616

Phillips, S.E. (1994). High-stakes testing accommodations: Validity versus disabled rights. *Applied Measurement in Evaluation, 7(2),* 93-120.

Pomplun, M. & Omar, M. H. (2000). Score comparability of a state mathematics assessment across students with and without reading accommodations. *Journal of Applied Psychology*. 85(1), 21-29.

Raykov, T., & Marcoulides, G. A. (2000). A method for comparing completely standardized solutions in multiple groups. *Structural Equation Modeling*, 7, 292-308.

Rhode Island Department of Elementary and Secondary Education. (2003). *Rhode Island assessment accommodation study: Research summary.* Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved October 2, 2009, from the World Wide Web: http://education.umn.edu/NCEO/TopicAreas/Accommodations/RhodeIsland.htm

Rose, D. H. (2001). Universal design for learning: Deriving guiding principles from networks that learn. *Journal of Special Education Technology,* 16 (1), 66-70.

Rose, D., & A. Meyer. (2000). Universal design for learning, associate editor column. *Journal of Special Education Technology,* 15 (1), 66-67.

Roussos, L. A. & Stout, W. F. (1996). Simulation studies of the effects of small sample size and studied item parameters on SIBTEST and Mantel-Haenszel type I error performance. *Journal of Educational Measurement*, 33(2): 215–230.

Russell, M. Hoffmann, T. & Higgins, J. (2009). Meeting the needs of all students: A universal design approach to computer based testing. *Innovate,* 5(4). Retrieved March 29, 2010, From http://www.innovateonline.info/pdf/vol5_issue4/Meeting_the_Needs_of_All_Students-__A_Universal_Design_Approach_to_Computer-Based_Testing.pdf.

Russell, M., Kavanaugh, M., Masters, J., Higgins, J. & Hoffmann, T. (2009). Computer-based signing accommodations: comparing a recorded human with an avatar. *Journal of Applied Testing Technology, 10* (3), 21.

Russell, M. & Plati, T. (2001). Effects of computer versus paper administration of a state-mandated writing assessment. *Teachers College Record*.

Schulte, A. A., Elliott, S. N., & Kratochwill, T. R. (2001, March). Effects of testing accommodations on students' standardized mathematics test scores: An experimental analysis. *School Psychology Review,* 30, 527–547.

Shealy, R., & Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DIF as well as item bias/dif. *Psychometrika, 58*, 159–194.

Sireci, S. G. (2008). *Validity Issues in Accommodating Reading Tests.* Jurnal Pendidik dan Pendidikan, 23, 81-110.

Stone, E., Cook, L. L., Laitusis, C., & Cline, F. (2010). Using differential item functioning to investigate the impact of testing accommodations on an English language arts assessment for students who are blind or visually impaired. *Applied Measurement in Education, 23*(2), 132–152.

Thompson, S. J., Blount, A., & Thurlow, M. L. (2002). *A summary of research on the effects of test accommodations—1999 through 2001.* Minneapolis, MN: National Center on Educational Outcomes.

Thompson, S.J., Johnstone, C.J., Anderson, M. E., & Miller, N. A. (2005). *Considerations for the development and review of universally designed assessments* (Technical Report 42). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved March, 3, 2010, from the World Wide Web: http://education.umn.edu/NCEO/OnlinePubs/Technical42.htm

Thompson, S. J., Johnstone, C. J. & Thurlow, M. L.. (2002). *Universal design applied to large scale assessments.* Synthesis Report 44. Minneapolis: University of Minnesota, National Center on Educational Outcomes. http://cehd.umn.edu/NCEO/OnlinePubs/Synthesis44.html (accessed February 12, 2010. Archived at http://www.webcitation.org/5dsOUMJgr.

Thompson, S. & Thurlow, M. (2002). *Universally designed assessments: Better tests for everyone!* (Policy Directions No. 14). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Thompson, S., Thurlow, M., & Moore, M. (2003). *Using computer-based tests with students with disabilities* (Policy Directions No. 15). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Thompson, S. J., Thurlow, M. L., Quenemoen, R. F., & Lehr, C. A. (2002). *Access to computer-based testing for students with disabilities* (Synthesis Report 45). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Thurlow, M. L., Altman, J. R., Cormier, D., & Moen, R. (2008). *Annual performance reports: 2005-2006 state assessment data.* Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Available at www.nceo.info/OnlinePubs/APRreport2005-2006.pdf.

Thurlow, M. L., McGrew, K.S., Tindal, G., Thompson, S. L., Ysseldyke, J. E., & Elliott, J. L. (2000). *Assessment accommodations research: Considerations for design and analysis* (Technical Report 26). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Thurlow, M. L., Laitusis, C. C., Dillon, D. R., Cook, L. L., Moen, R. E., Abedi, J., & O'Brien, D. G. (2009). *Accessibility principles for reading assessments*. Minneapolis, MN: National Accessible Reading Assessment Projects.

Thurlow, M. L., McGrew, K.S., Tindal, G., Thompson, S. L., Ysseldyke, J. E., & Elliott, J. L. (2000). *Assessment accommodations research: Considerations for design and analysis* (Technical Report 26). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved February 10, 2010, from the World Wide Web: http://education.umn.edu/NCEO/OnlinePubs/Technical26.htm.

Tindal, G. (1998). *Determining when accommodated test administrations are comparable to standard test administrations.* Paper prepared for the Assessing Special Education Students (ASES) Work Group and Council of Chief State School Officers. Obtain from CCSSO, One Massachusetts Avenue NW, Suite 700, Washington, DC 20001-1431.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (p. 67-113). Hillsdale NJ: Erlbaum.

Trotter, A. (2001). Testing computerized exams. *Education Week*, 20(37), 30-35.

U.S. Department of Education. (2010). Overview information: Race to the top fund assessment program: Notice inviting applications for new awards for fiscal year (FY) 2010. *Federal Register,* 20(68), 18171 - 18185.

Wainer, H. (1993). Model-based standardized measurement of an item's differential impact. In P.W. Holland & H. Wainer (Eds.) *Differential item functioning.* (p. 123-135). Hillsdale, NJ: Lawrence Erlbaum Associates.

Ysseldyke, J. E., Thurlow, M.L., Langenfeld, K. L., Nelson, J. R., Teelucksingh, E., & Seyfarth, A. (1998). *Educational results for students with disabilities: What do the data tell us?* (Technical Report 23). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

APPENDIX A: EXPLORATORY FACTOR ANALYSIS RESULTS: FACTOR LOADINGS AND FACTOR CORRELATIONS

Table A.1. EFA for No Accommodations: Factor Loadings for Unrotated Solution

	Component							
Item	1	2	3	4	5	6	7	8
18	.333			-.323				
11	.337							
14	.345						.440	
15	.346	-.376						
9	.357					.333		
24	.367					-.314		
28	.376							
21	.389							
30	.392							
31	.393			.322				
23	.410							
17	.415							
22	.416			-.360				
3	.417		.307					
12	.419							
19	.419							
25	.421		-.305					
32	.428	-.345						
1	.438	-.345						
8	.452	.310						
10	.453							
33	.459							
5	.484							
29	.491							
27	.566							
2					.626	-.301		
4			.374		-.433		-.349	
6			.389				.312	.325
7					.416	.344		
13			-.355					.421
16							.475	
20		-.336		.330				
26			.312					

Extraction Method: Principal Component Analysis.
a. TestConditions = No Accommodations
b. 8 components extracted.

Table A.2. EFA for Accommodations - Paper: Factor Loadings for Unrotated Solution

	Component								
Item	1	2	3	4	5	6	7	8	9
8	.323		.342					-.315	
24	.325								
5	.332	.363		.308					
30	.336				.358	-.345			
25	.342								-.377
18	.358							-.332	
12	.391								
22	.392						-.347		
3	.393								
21	.393		-.354						.314
15	.399	-.431							
23	.399								
14	.399						-.351		
9	.411								
19	.421								
10	.422								-.310
33	.425								
17	.429								
29	.475								-.307
32	.508								
1	.541								
27	.552								
2		-.304	.336						
4		-.300							
6				.457					
7									
11					.388				
13						.519		.310	
16			.343				.523	.308	
20		-.397		.355					
26					-.317				
28					.351			-.362	
31				.483					

Extraction Method: Principal Component Analysis.
a. TestConditions = Accommodations - Paper
b. 9 components extracted.

Table A.3. EFA for Accommodations - Nimble: Factor Loadings for Unrotated Solution

	Component									
Item	1	2	3	4	5	6	7	8	9	10
22	.301			.334						
21	.311		-.353		.314					
25	.322					-.379				
15	.327							.311		
12	.330									
14	.338			-.311				.380		
26	.342									
33	.353							-.324		
19	.360								-.349	
3	.363							-.311		
17	.365									
32	.381									
9	.405									
10	.414									
29	.420	-.331								
1	.473								.329	
27	.527									
2						-.313	.316		.318	
4		.371								
5		-.342					-.309			
6										
7					.481					.477
8										-.330
11		.558								
13				.568						
16			.310		.426					-.365
18								.355	-.337	
20							.387			
23					.330		-.386			
24				-.352						
28			-.350	.348						
30						.475				
31			.435			.340				

Extraction Method: Principal Component Analysis.
a. TestConditions = Accommodations - Nimble
b. 14 components extracted.

Table A.3. EFA for Accommodations - Nimble: Factor Loadings for Unrotated Solution (continued)

	Component				
Item	10	11	12	13	14
22					
21					
25		-.315			
15					
12					-.385
14					
26					
33					
19					
3				.381	
17					
32					
9					
10					
29					
1					
27					
2				.397	
4				.393	
5			.394		
6			-.487		-.326
7	.477				
8	-.330				
11					
13					
16	-.365				
18					
20					
23					
24					
28					-.321
30					
31					

Extraction Method: Principal Component Analysis.
a. TestConditions = Accommodations - Nimble
b. 14 components extracted.

Table A.4. EFA for No Accommodations: Factor Loadings with Promax Rotation

	Component										
Item	1	2	3	4	5	6	7	8	9	10	11
12	.394										
27	.626										
29	.692										
32	.970										
31		.368									
8		.510							.312		
5		.690									
3		.831									
33			.614								
19			.620								
22			.840								
17				.477							
11				1.026							
1					.346						
2					.998						
28						.960					
25							.486				
15							.968				
4								.981			
9									.983		
23										.974	
30											.964
6											
13											
14											
18											
20											
26											
24											
16											
7											
21											
10											

Extraction Method: Principal Component Analysis.
Rotation Method: Promax with Kaiser Normalization.
a. TestConditions = No Accommodations
b. Rotation converged in 12 iterations.

Table A.4. EFA for No Accommodations: Factor Loadings with Promax Rotation (continued)

	Component									
Item	12	13	14	15	16	17	18	19	20	21
12										
27										
29										
32										
31										
8										
5										
3										
33										
19										
22										
17				.312						
11										
1										
2										
28										
25										
15										
4										
9										
23										
30										
6	1.002									
13		1.126								
14			.960							
18				.890						
20					1.037					
26						1.003				
24							.914			
16								1.013		
7									.980	
21										.959
10										

Extraction Method: Principal Component Analysis.
Rotation Method: Promax with Kaiser Normalization.
a. TestConditions = No Accommodations
b. Rotation converged in 12 iterations.

Table A.5. EFA for Accommodations - Paper: Factor Loadings with Promax Rotation

	Component										
Item	1	2	3	4	5	6	7	8	9	10	11
32	.677										
27	.761								-.302		
15	1.012										
33		.791									
22		.972									
18			.340								
21			.752								
19			.834								
17				.728							
5				.820				.340			
23					.301						
14					.528						
9					.926						
24						1.247					
10							.362				
3							1.138				
8								1.304			
28									1.038		
1										.361	
2										.959	
16											1.101
31											
29											
30											
4											
20											
11											
13											
25											
26											
7											
6											
12											

Extraction Method: Principal Component Analysis.
Rotation Method: Promax with Kaiser Normalization.
a. TestConditions = Accommodations - Paper
b. Rotation converged in 15 iterations.

Table A.5. EFA for Accommodations - Paper: Factor Loadings with Promax Rotation (continued)

	Component									
Item	12	13	14	15	16	17	18	19	20	21
32										
27										
15										
33										
22										
18										
21				-.313						
19		.336								
17	-.305									
5			-.305							
23										-.454
14										
9										
24										
10					-.404					
3										
8										
28										
1										
2										
16										
31	1.107									
29		-.339								
30		.924								
4			1.050							
20				1.016						
11					.962					
13						1.071				
25							-.344			
26							1.010			
7								.980		
6									.987	
12										.927

Extraction Method: Principal Component Analysis.
Rotation Method: Promax with Kaiser Normalization.
a. TestConditions = Accommodations - Paper
b. Rotation converged in 15 iterations.

Table A.6. EFA for Accommodations - Nimble: Factor Loadings with Promax Rotation

	Component										
Item	1	2	3	4	5	6	7	8	9	10	11
29	.432							-.321			
14	.653										
10	.900										
27		.316									
1		.351									
32		1.081									
28			-.308				.411				
22			.308								
26			.398						.314		
17			.991								
33				.412							
19				1.068							
25					.622						
12					.935						
3						-.315					
21						1.029					
13							1.050				
31								1.072			
15									.953		
9										-.317	
7										1.047	
24											1.002
23											
8											
18											
4											
30											
2											
6											
16											
20											
5											
11											

Extraction Method: Principal Component Analysis.
Rotation Method: Promax with Kaiser Normalization.
a. TestConditions = Accommodations - Nimble
b. Rotation converged in 58 iterations.

Table A.6. EFA for Accommodations - Nimble: Factor Loadings with Promax Rotation (continued)

	Component									
Item	12	13	14	15	16	17	18	19	20	21
29	-.338									
14										
10										
27										
1										
32										
28										
22										
26										
17										
33		.348								
19										
25										
12										
3			.355							
21										
13										
31										
15										
9										
7										
24										
23	1.107									
8		1.074								
18			1.048							
4				1.081						
30					1.084					
2						1.135				
6							1.026			
16								1.089		
20									1.011	
5										1.013
11										

Extraction Method: Principal Component Analysis.
Rotation Method: Promax with Kaiser Normalization.
a. TestConditions = Accommodations - Nimble
b. Rotation converged in 58 iterations.

Table A.7. EFA for No Accommodations: Factor Correlation with Promax Rotation

	Component										
Item	1	2	3	4	5	6	7	8	9	10	11
1	1.000	.412	.493	.400	.286	.144	.367	.251	.303	.154	.316
2	.412	1.000	.355	.410	.082	.222	.316	.127	.248	.330	.171
3	.493	.355	1.000	.239	.172	.175	.340	.110	.298	.248	.213
4	.400	.410	.239	1.000	.121	.062	.152	.204	.221	.057	.275
5	.286	.082	.172	.121	1.000	-.128	.027	.332	.053	-.018	.031
6	.144	.222	.175	.062	-.128	1.000	.240	-.113	-.011	.287	.073
7	.367	.316	.340	.152	.027	.240	1.000	.022	.217	.287	.196
8	.251	.127	.110	.204	.332	-.113	.022	1.000	.108	-.012	.101
9	.303	.248	.298	.221	.053	-.011	.217	.108	1.000	.097	.254
10	.154	.330	.248	.057	-.018	.287	.287	-.012	.097	1.000	-.034
11	.316	.171	.213	.275	.031	.073	.196	.101	.254	-.034	1.000
12	.200	.213	.237	.110	.132	-.075	.073	.119	.250	.088	.044
13	.339	.319	.248	.327	-.071	.218	.229	-.012	.246	.109	.288
14	.197	.109	.080	.270	.218	-.229	-.090	.184	.098	-.089	.020
15	-.021	.037	.110	-.017	.076	-.030	.001	.021	-.059	.173	-.108
16	-.067	-.131	-.023	-.200	.129	-.130	-.025	.111	-.061	.041	-.213
17	.251	.147	.253	.148	.107	-.048	.090	.076	.231	.014	.194
18	-.035	-.025	.003	-.144	.119	-.039	.032	.034	-.016	.100	-.077
19	-.088	.053	-.016	-.109	-.241	.247	.183	-.176	-.019	.131	.041
20	-.074	-.087	.019	-.078	.109	-.109	-.065	.064	.037	.058	-.100
21	-.088	-.004	.007	-.121	-.084	.087	.062	-.061	-.067	.146	-.175

Extraction Method: Principal Component Analysis.
Rotation Method: Promax with Kaiser Normalization.
a. TestConditions = No Accommodations

Table A.7. EFA for No Accommodations: Factor Correlation with Promax Rotation (continued)

	Component									
Item	12	13	14	15	16	17	18	19	20	21
1	.200	.339	.197	-.021	-.067	.251	-.035	-.088	-.074	-.088
2	.213	.319	.109	.037	-.131	.147	-.025	.053	-.087	-.004
3	.237	.248	.080	.110	-.023	.253	.003	-.016	.019	.007
4	.110	.327	.270	-.017	-.200	.148	-.144	-.109	-.078	-.121
5	.132	-.071	.218	.076	.129	.107	.119	-.241	.109	-.084
6	-.075	.218	-.229	-.030	-.130	-.048	-.039	.247	-.109	.087
7	.073	.229	-.090	.001	-.025	.090	.032	.183	-.065	.062
8	.119	-.012	.184	.021	.111	.076	.034	-.176	.064	-.061
9	.250	.246	.098	-.059	-.061	.231	-.016	-.019	.037	-.067
10	.088	.109	-.089	.173	.041	.014	.100	.131	.058	.146
11	.044	.288	.020	-.108	-.213	.194	-.077	.041	-.100	-.175
12	1.000	.116	.085	.060	-.028	.211	.052	-.147	.029	-.138
13	.116	1.000	-.034	-.129	-.268	.119	-.158	.075	-.163	-.157
14	.085	-.034	1.000	.080	.084	.097	.059	-.236	.094	.003
15	.060	-.129	.080	1.000	.094	-.019	.139	-.093	.081	.031
16	-.028	-.268	.084	.094	1.000	-.085	.162	-.005	.258	.248
17	.211	.119	.097	-.019	-.085	1.000	-.031	-.121	-.001	-.111
18	.052	-.158	.059	.139	.162	-.031	1.000	.032	.128	.112
19	-.147	.075	-.236	-.093	-.005	-.121	.032	1.000	-.030	.205
20	.029	-.163	.094	.081	.258	-.001	.128	-.030	1.000	.188
21	-.138	-.157	.003	.031	.248	-.111	.112	.205	.188	1.000

Extraction Method: Principal Component Analysis.
Rotation Method: Promax with Kaiser Normalization.
a. TestConditions = No Accommodations

Table A.8. EFA for Accommodations - Paper: Factor Correlation with Promax Rotation

	Component										
Item	1	2	3	4	5	6	7	8	9	10	11
1	1.000	.545	.066	-.029	.128	.580	.378	.515	.323	-.063	.449
2	.545	1.000	.175	.095	.112	.525	.363	.431	.283	-.083	.354
3	.066	.175	1.000	.153	.261	.054	.193	.034	.230	.069	.007
4	-.029	.095	.153	1.000	.346	-.081	-.235	-.244	-.017	.188	-.131
5	.128	.112	.261	.346	1.000	.008	-.174	-.107	.072	.093	-.017
6	.580	.525	.054	-.081	.008	1.000	.438	.591	.249	-.165	.403
7	.378	.363	.193	-.235	-.174	.438	1.000	.573	.312	-.009	.297
8	.515	.431	.034	-.244	-.107	.591	.573	1.000	.270	-.170	.425
9	.323	.283	.230	-.017	.072	.249	.312	.270	1.000	-.043	.241
10	-.063	-.083	.069	.188	.093	-.165	-.009	-.170	-.043	1.000	-.105
11	.449	.354	.007	-.131	-.017	.403	.297	.425	.241	-.105	1.000
12	-.344	-.314	-.001	.313	.242	-.427	-.504	-.502	-.152	.151	-.289
13	.175	-.010	-.261	.032	.040	.032	-.113	.023	-.122	.014	.058
14	-.029	-.020	-.012	.310	.269	-.156	-.308	-.302	-.102	.160	-.128
15	.244	.186	.241	-.048	.142	.235	.248	.223	.199	-.108	.123
16	.230	.229	.178	.003	.089	.227	.183	.216	.240	-.139	.184
17	.153	.219	.192	-.121	-.072	.229	.330	.278	.207	-.160	.212
18	-.188	-.034	.342	.156	.177	-.190	-.041	-.190	.092	.060	-.042
19	.038	.025	.195	.109	.156	-.002	.047	-.015	.075	.088	-.012
20	-.098	-.129	-.006	.091	.024	-.112	-.098	-.123	-.074	.092	-.116
21	-.060	-.076	-.305	.201	.044	-.119	-.288	-.200	-.286	.167	-.187

Extraction Method: Principal Component Analysis.
Rotation Method: Promax with Kaiser Normalization.
a. TestConditions = Accommodations - Paper

Table A.8. EFA for Accommodations - Paper: Factor Correlation with Promax Rotation (continued)

	Component									
Item	12	13	14	15	16	17	18	19	20	21
1	-.344	.175	-.029	.244	.230	.153	-.188	.038	-.098	-.060
2	-.314	-.010	-.020	.186	.229	.219	-.034	.025	-.129	-.076
3	-.001	-.261	-.012	.241	.178	.192	.342	.195	-.006	-.305
4	.313	.032	.310	-.048	.003	-.121	.156	.109	.091	.201
5	.242	.040	.269	.142	.089	-.072	.177	.156	.024	.044
6	-.427	.032	-.156	.235	.227	.229	-.190	-.002	-.112	-.119
7	-.504	-.113	-.308	.248	.183	.330	-.041	.047	-.098	-.288
8	-.502	.023	-.302	.223	.216	.278	-.190	-.015	-.123	-.200
9	-.152	-.122	-.102	.199	.240	.207	.092	.075	-.074	-.286
10	.151	.014	.160	-.108	-.139	-.160	.060	.088	.092	.167
11	-.289	.058	-.128	.123	.184	.212	-.042	-.012	-.116	-.187
12	1.000	.034	.231	-.194	-.197	-.299	.187	.075	.164	.239
13	.034	1.000	.114	-.066	.028	-.142	-.269	-.036	.062	.220
14	.231	.114	1.000	-.070	-.069	-.166	.046	.079	.036	.244
15	-.194	-.066	-.070	1.000	.233	.260	.042	.082	-.080	-.260
16	-.197	.028	-.069	.233	1.000	.238	.028	.009	-.070	-.236
17	-.299	-.142	-.166	.260	.238	1.000	.078	.009	-.095	-.336
18	.187	-.269	.046	.042	.028	.078	1.000	.086	-.006	-.205
19	.075	-.036	.079	.082	.009	.009	.086	1.000	.027	-.054
20	.164	.062	.036	-.080	-.070	-.095	-.006	.027	1.000	.078
21	.239	.220	.244	-.260	-.236	-.336	-.205	-.054	.078	1.000

Extraction Method: Principal Component Analysis.
Rotation Method: Promax with Kaiser Normalization.
a. TestConditions = Accommodations - Paper

Table A.9. EFA for Accommodations - Nimble: Factor Correlation with Promax Rotation

	Component										
Item	1	2	3	4	5	6	7	8	9	10	11
1	1.000	.340	.347	.359	.028	.008	.179	-.204	.101	-.218	.026
2	.340	1.000	.169	.399	.162	.093	.370	-.262	-.039	-.211	-.007
3	.347	.169	1.000	.057	.134	.008	.228	-.352	.077	-.058	.235
4	.359	.399	.057	1.000	-.105	-.087	.214	-.089	.031	-.281	-.118
5	.028	.162	.134	-.105	1.000	.303	-.022	-.155	.125	.035	.001
6	.008	.093	.008	-.087	.303	1.000	-.183	.128	-.012	-.313	-.234
7	.179	.370	.228	.214	-.022	-.183	1.000	-.357	-.061	-.069	.041
8	-.204	-.262	-.352	-.089	-.155	.128	-.357	1.000	-.021	-.148	-.026
9	.101	-.039	.077	.031	.125	-.012	-.061	-.021	1.000	-.045	.167
10	-.218	-.211	-.058	-.281	.035	-.313	-.069	-.148	-.045	1.000	.153
11	.026	-.007	.235	-.118	.001	-.234	.041	-.026	.167	.153	1.000
12	.117	-.126	.094	.222	-.121	-.056	-.175	.114	.304	-.130	.088
13	.191	.302	.098	.372	.047	-.150	.231	-.232	-.071	.000	-.149
14	.076	.110	.241	-.090	.200	-.135	.140	-.309	-.026	.281	.120
15	.074	-.047	.166	.064	-.141	-.037	.123	.017	.279	-.218	.147
16	-.162	.149	-.115	-.130	.277	.323	.059	.011	-.124	-.059	-.149
17	-.134	.169	-.079	-.148	.237	.204	.074	-.005	-.246	.029	-.106
18	.105	.218	-.145	.205	-.163	-.031	.055	.170	-.249	-.084	-.078
19	.352	.086	.340	.094	.061	.234	.043	-.156	.126	-.239	-.026
20	-.042	.050	-.016	.030	-.031	.110	.117	.036	.102	-.152	.080
21	-.165	-.206	-.081	-.231	-.057	-.034	-.161	.217	-.019	.162	.163

Extraction Method: Principal Component Analysis.
Rotation Method: Promax with Kaiser Normalization.
a. TestConditions = Accommodations - Nimble

Table A.9. EFA for Accommodations - Nimble: Factor Correlation with Promax Rotation (continued)

	Component									
Item	12	13	14	15	16	17	18	19	20	21
1	.117	.191	.076	.074	-.162	-.134	.105	.352	-.042	-.165
2	-.126	.302	.110	-.047	.149	.169	.218	.086	.050	-.206
3	.094	.098	.241	.166	-.115	-.079	-.145	.340	-.016	-.081
4	.222	.372	-.090	.064	-.130	-.148	.205	.094	.030	-.231
5	-.121	.047	.200	-.141	.277	.237	-.163	.061	-.031	-.057
6	-.056	-.150	-.135	-.037	.323	.204	-.031	.234	.110	-.034
7	-.175	.231	.140	.123	.059	.074	.055	.043	.117	-.161
8	.114	-.232	-.309	.017	.011	-.005	.170	-.156	.036	.217
9	.304	-.071	-.026	.279	-.124	-.246	-.249	.126	.102	-.019
10	-.130	.000	.281	-.218	-.059	.029	-.084	-.239	-.152	.162
11	.088	-.149	.120	.147	-.149	-.106	-.078	-.026	.080	.163
12	1.000	.053	-.160	.274	-.374	-.437	-.147	.217	-.054	-.058
13	.053	1.000	.167	-.099	-.005	.062	.074	.026	-.114	-.264
14	-.160	.167	1.000	-.138	.112	.167	-.067	-.066	-.129	-.039
15	.274	-.099	-.138	1.000	-.168	-.292	-.118	.210	.224	-.051
16	-.374	-.005	.112	-.168	1.000	.448	.085	-.141	.072	-.005
17	-.437	.062	.167	-.292	.448	1.000	.176	-.223	-.013	.048
18	-.147	.074	-.067	-.118	.085	.176	1.000	-.087	.007	.065
19	.217	.026	-.066	.210	-.141	-.223	-.087	1.000	.083	-.144
20	-.054	-.114	-.129	.224	.072	-.013	.007	.083	1.000	.047
21	-.058	-.264	-.039	-.051	-.005	.048	.065	-.144	.047	1.000

Extraction Method: Principal Component Analysis.
Rotation Method: Promax with Kaiser Normalization.
a. TestConditions = Accommodations - Nimble

Table A.10. EFA for No Accommodations: Factor Loadings with Varimax Rotation

	Component										
Item	1	2	3	4	5	6	7	8	9	10	11
12	.448			.306							
27	.574										
29	.584					.315					
32	.718										
31		.409				.346					
8		.515							.349		
5		.603									
3		.680									
19			.559								
33			.576								.339
22			.704								
17				.520							
11				.822							
10					.324						
1					.414						
2					.849						
25						.302	.534				
28						.843					
15							.820				
4								.886			
9									.885		
23										.860	
30											.859
6											
13											
14											
18											
20											
26											
24											
16											
7											
21											

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.
a. TestConditions = No Accommodations
b. Rotation converged in 12 iterations.

Table A.10. EFA for No Accommodations: Factor Loadings with Varimax Rotation (continued)

	Component									
Item	12	13	14	15	16	17	18	19	20	21
12										
27										
29										
32										
31										
8										
5										
3										
19										
33										
22										
17				.349						
11										
10	.341									
1										
2										
25										
28										
15										
4										
9										
23										
30										
6	.916									
13		.940								
14			.847							
18				.851						
20					.913					
26						.937				
24							.886			
16								.899		
7									.917	
21										.868

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.
a. TestConditions = No Accommodations
b. Rotation converged in 12 iterations.

Table A.11. EFA for Accommodations - Paper: Factor Loadings with Varimax Rotation

	Component										
Item	1	2	3	4	5	6	7	8	9	10	11
23	.303				.374						
29	.343	.317									
25	.370					.354					
32	.538										
27	.617										
15	.691										
33		.661									
22		.749									
18			.449				.389				
21			.634								
19			.664								
17				.649							
5				.700							
14					.560						
9					.770						
24						.870					
10							.396				
3							.775				
8								.859			
28									.920		
1										.422	
2										.897	
16											.919
31											
30											
4											
20											
11											
13											
26											
7											
6											
12											

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.
a. TestConditions = Accommodations - Paper
b. Rotation converged in 15 iterations.

Table A.11. EFA for Accommodations - Paper: Factor Loadings with Varimax Rotation (continued)

	Component									
Item	12	13	14	15	16	17	18	19	20	21
23										-.405
29										
25							-.425			
32										
27										
15										
33										
22										
18										-.303
21										
19		.320								
17										
5										
14										
9										
24										
10					-.360					
3										
8										
28										
1										
2										
16										
31	.815									
30		.833								
4			.915							
20				.907						
11					.881					
13						.925				
26							.846			
7								.957		
6									.966	
12										.727

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.
a. TestConditions = Accommodations - Paper
b. Rotation converged in 15 iterations.

Table A.12. EFA for Accommodations - Nimble: Factor Loadings with Varimax Rotation

	Component										
Item	1	2	3	4	5	6	7	8	9	10	11
22	.310		.386								
29	.469							-.377			
14	.579			.355							
10	.724										
27		.342				.323					
1		.416									
32		.835									
26			.436						.406		
17			.781								
33				.450							
19				.790							
25					.611						
12					.760						
11						-.359					.315
3						-.337					
9						.303				-.360	
21						.749					
28							.436			-.318	
13							.832				
31								.817			
15									.827		
7										.814	
24											.861
23											
8											
18											
4											
30											
2											
6											
16											
20											
5											

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.
a. TestConditions = Accommodations - Nimble
b. Rotation converged in 58 iterations.

Table A.12. EFA for Accommodations - Nimble: Factor Loadings with Varimax Rotation (continued)

	Component									
Item	12	13	14	15	16	17	18	19	20	21
22										
29										
14										
10										
27										
1										
32										
26							-.304			
17										
33		.451								
19										
25										
12										
11					-.318	-.351				
3			.442							
9										
21										
28										
13										
31										
15										
7										
24										
23	.854									
8		.881								
18			.879							
4				.911						
30					.851					
2						.859				
6							.871			
16								.875		
20									.941	
5										.910

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.
a. TestConditions = Accommodations - Nimble
b. Rotation converged in 58 iterations.

APPENDIX B: ITEM CHARACTERISTIC CURVES FOR ITEMS SHOWING DIF

DIF Results for Comparisons Between No Accommodations and Accommodations - Paper

Figure B.1. No Accommodations v. Accommodations - Paper: Item 1

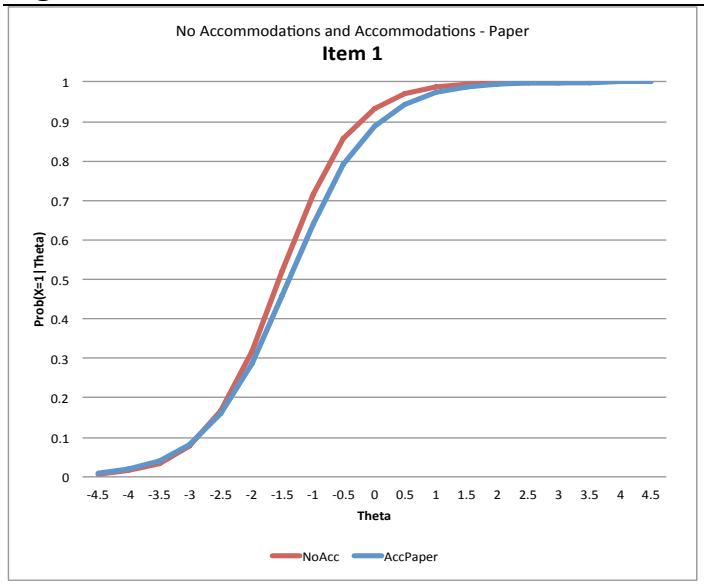


Figure B.2. No Accommodations v. Accommodations - Paper: Item 11

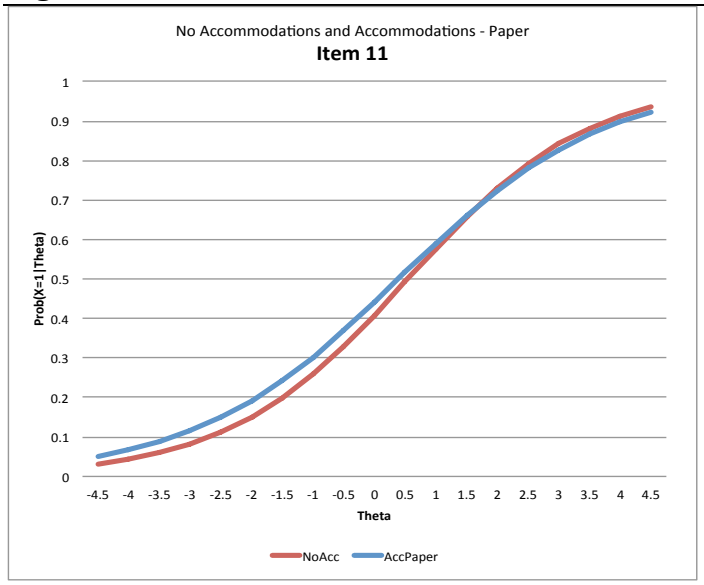


Figure B.3. No Accommodations v. Accommodations - Paper: Item 15

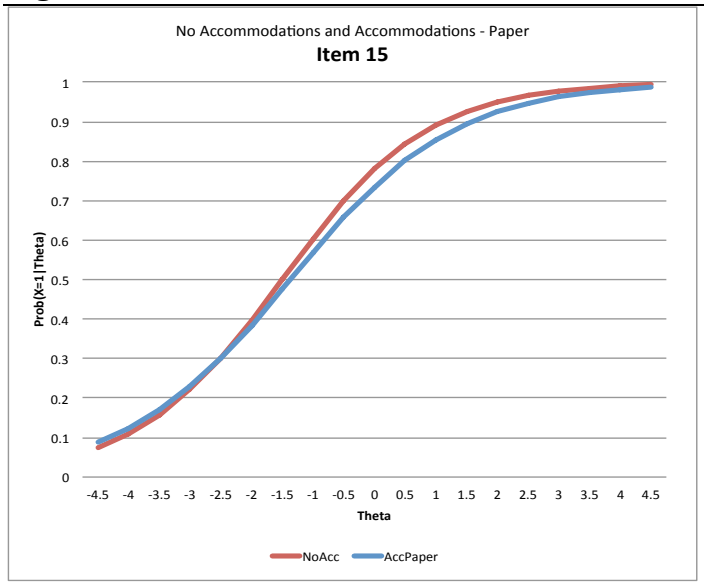


Figure B.4. No Accommodations v. Accommodations - Paper: Item 21

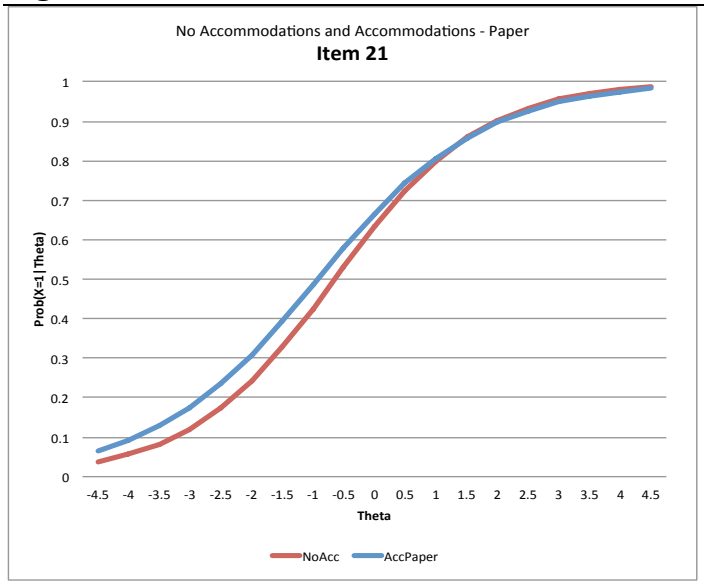


Figure B.5. No Accommodations v. Accommodations - Paper: Item 22

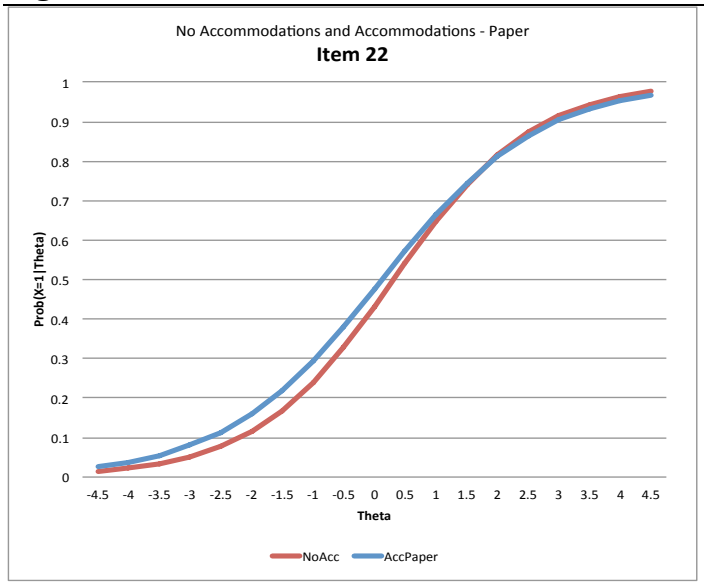


Figure B.6. No Accommodations v. Accommodations - Paper: Item 23

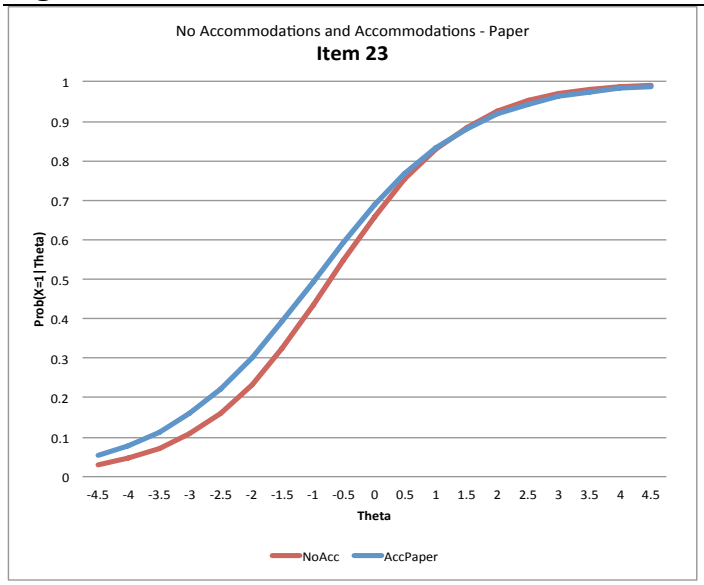


Figure B.7. No Accommodations v. Accommodations - Paper: Item 24

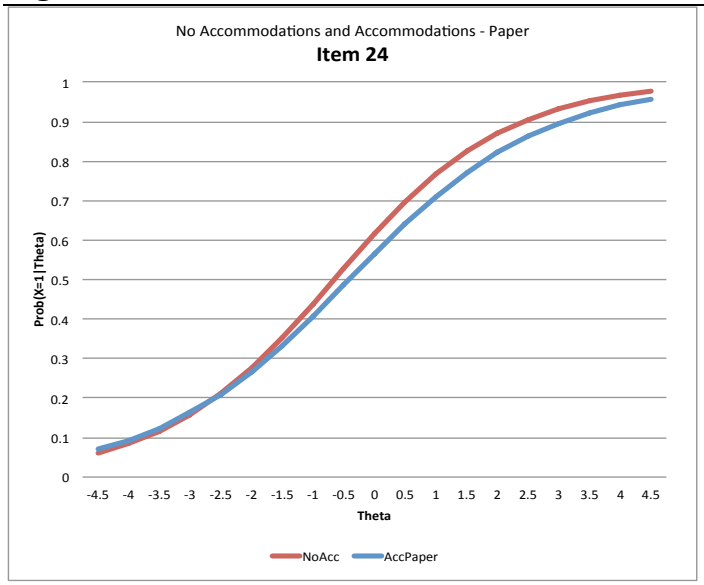


Figure B.8. No Accommodations v. Accommodations - Paper: Item 28

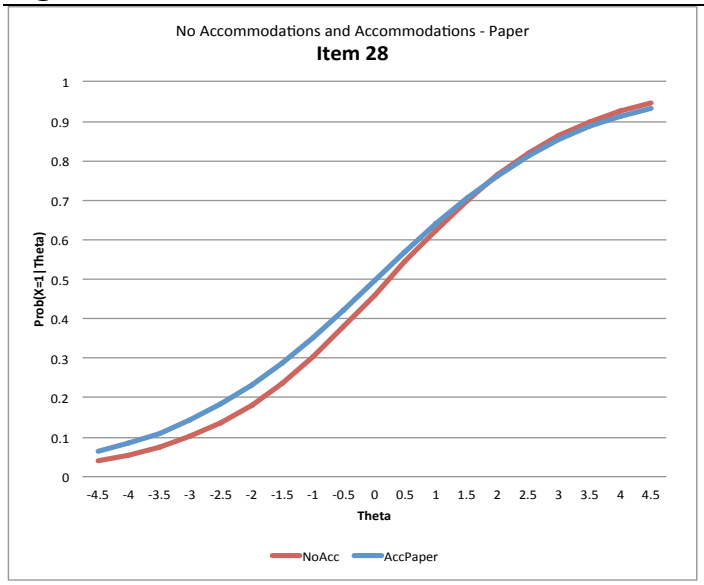
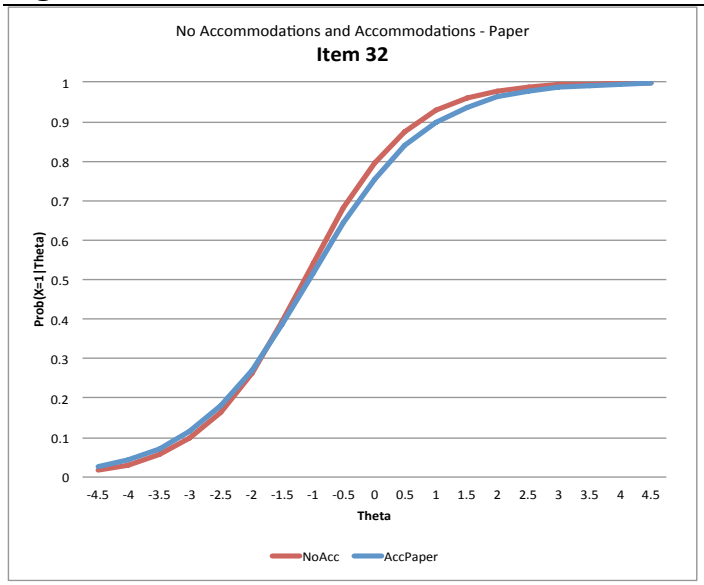


Figure B.9. No Accommodations v. Accommodations - Paper: Item 32



DIF Results for Comparisons Between No Accommodations and Accommodations - Nimble

Figure B.10. No Accommodations v. Accommodations - Nimble: Item 1

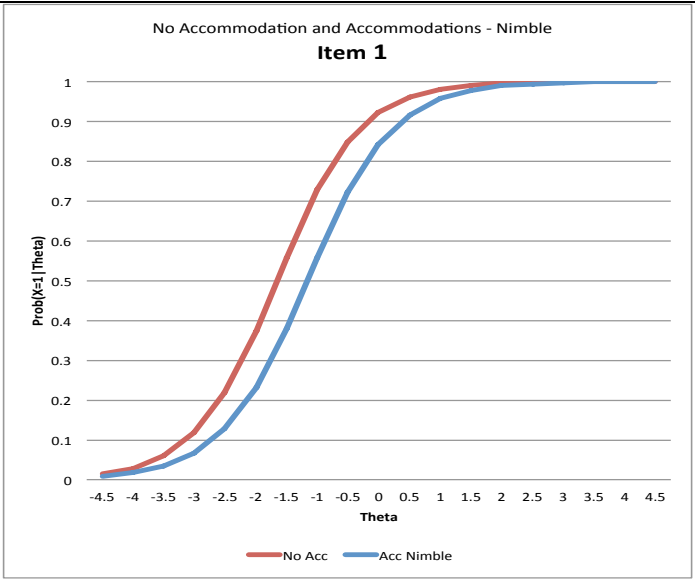


Figure B.11. No Accommodations v. Accommodations - Nimble: Item 16

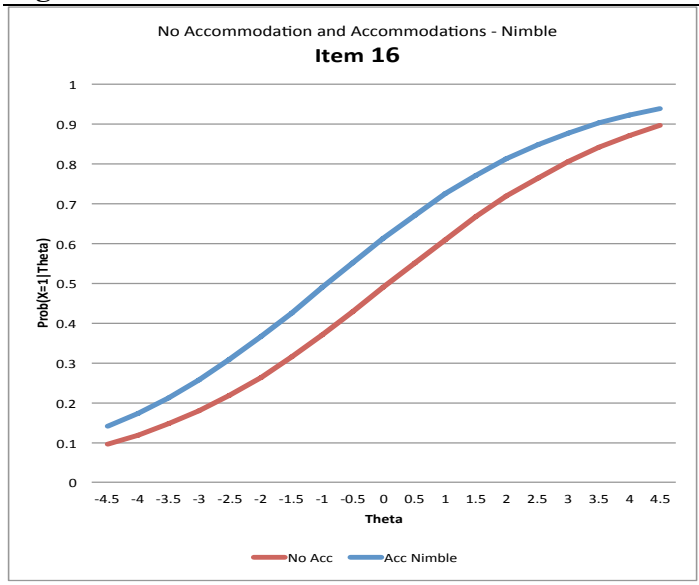


Figure B.12. No Accommodations v. Accommodations - Nimble: Item 17

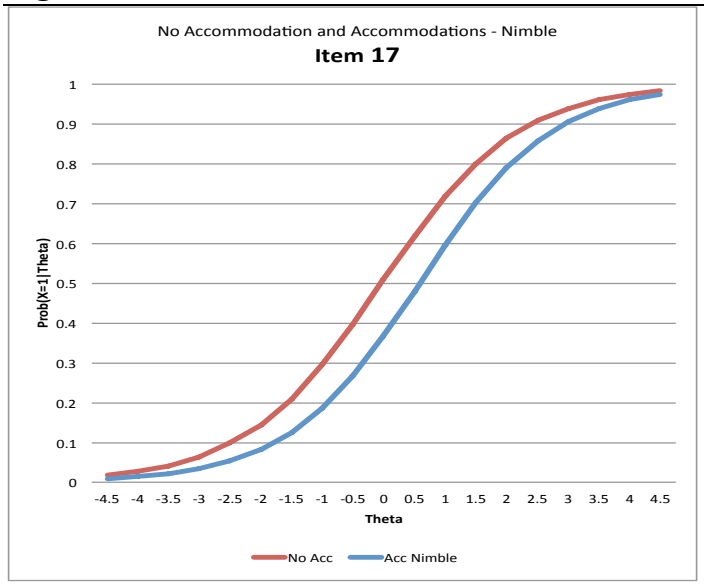


Figure B.13. No Accommodations v. Accommodations - Nimble: Item 23

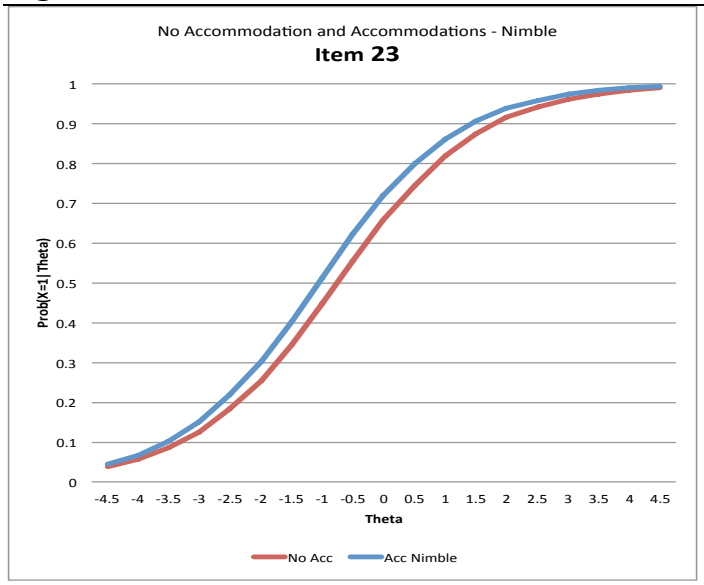


Figure B.14. No Accommodations v. Accommodations - Nimble: Item 28

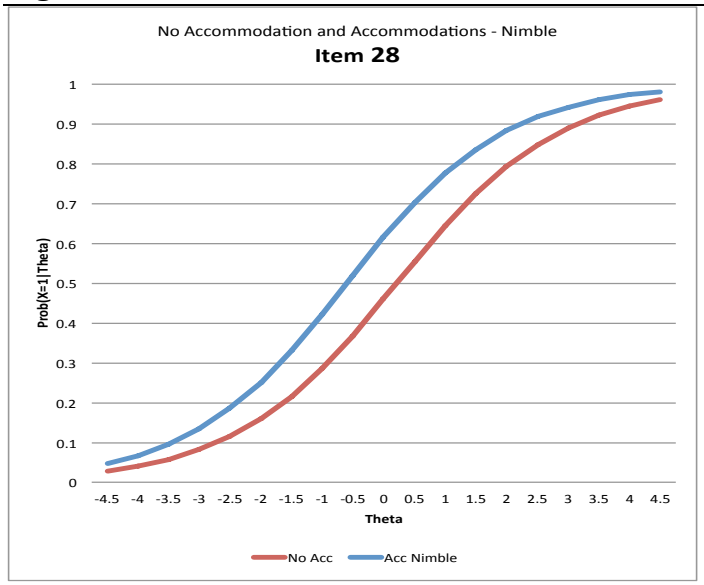


Figure B.15. No Accommodations v. Accommodations - Nimble: Item 29

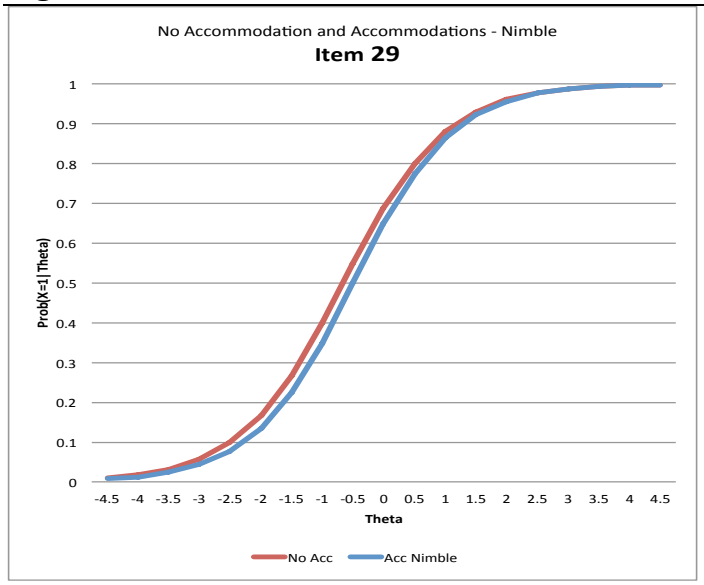


Figure B.16. No Accommodations v. Accommodations - Nimble: Item 31

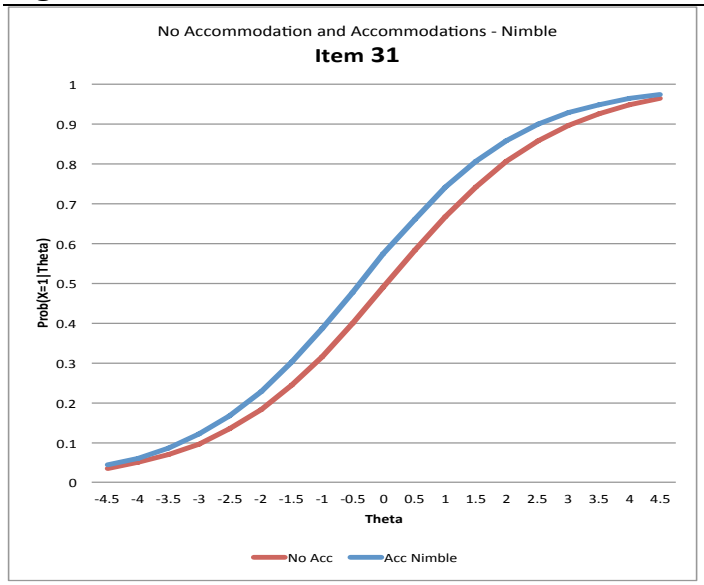


Figure B.17. No Accommodations v. Accommodations - Nimble: Item 32