CCSSO
Council of Chief State School Officers

Technical Report for Studies of the Validity of Test Results for Test Accommodations

Connecticut Enhanced Assessment Grant

Establishing the Validity of Test Accommodations and Score Interpretations for Students with Disabilities: A Collaboration of State-based Research

Report prepared by
John F. Olson, Principal Investigator and Project Coordinator
Mohamed Dirir, Project Analyst
The CTEAG Project Management Team

THE COUNCIL OF CHIEF STATE SCHOOL OFFICERS

The Council of Chief State School Officers (CCSSO) is a nonpartisan, nationwide, nonprofit organization of public officials who head departments of elementary and secondary education in the states, the District of Columbia, the Department of Defense Education Activity, and five U.S. extra-state jurisdictions. CCSSO provides leadership, advocacy, and technical assistance on major educational issues. The Council seeks member consensus on major educational issues and expresses their views to civic and professional organizations, federal agencies, Congress, and the public.

Technical Report
for Studies of the Validity of Test Results for Test Accommodations

Establishing the Validity of Test Accommodations and Score Interpretations for Students with Disabilities: A Collaboration of State-based Research

John F. Olson & Mohamed Dirir, Primary Authors

This report is based on research funded by an Enhanced Assessment Grant from the U.S. Department of Education, awarded to the Connecticut Department of Education, in partnership with the Council of Chief State School Officers and its ASES and TILSA SCASS consortia. Publication of this document shall not be construed as endorsement of the views expressed in it by the U.S. Department of Education, the Connecticut Department of Education, or the Council of Chief State School Officers.

COUNCIL OF CHIEF STATE SCHOOL OFFICERS

Steven L. Paine (West Virginia), President
Susan Gendron (Maine), Past-President

Gene Wilhoit, Executive Director

Council of Chief State School Officers
One Massachusetts Avenue, NW, Suite 700
Washington, DC 20001-1431
Phone (202) 336-7000
Fax (202) 408-8072
www.ccsso.org

Copyright © 2010 by the Council of Chief State School Officers, Washington, DC
All rights reserved.

Technical Report
for
Studies of the Validity of Test Results for Test Accommodations

Connecticut Enhanced Assessment Grant

Establishing the Validity of Test Accommodations and Score Interpretations for Students with Disabilities: A Collaboration of State-based Research

A Project of the Connecticut State Department of Education
and the
Council of Chief State School Officers
State Collaboratives on Assessment & Student Standards (SCASS)
Assessing Special Education Students (ASES)
Technical Issues in Large-Scale Assessment (TILSA)

Report prepared by:
John F. Olson, Principal Investigator and Project Coordinator,
Mohamed Dirir, Project Analyst,
and the
CTEAG Project Management Team

Consortium of States in CTEAG: Connecticut, Kentucky, Michigan, and Nevada

The project was funded by a grant from the U.S. Department of Education
(S368A070009)

June 2010

Council of Chief State School Officers
Washington, DC

Acknowledgments

This research project benefited from the support and advice of members of two of CCSSO's State Collaboratives on Assessment and Student Standards: the Technical Issues in Large-Scale Assessment (TILSA) collaborative, led by Doug Rindone; and the Assessing Special Education Students (ASES) collaborative, led by Sandra Warren. The authors want to especially thank the project management team, the Connecticut State Department of Education staff, and the state department staff members of the validity studies for their work and participation in the project.

CTEAG Project Management Team:
Barbara Beaudin, Connecticut DOE and Project Director
Mohamed Dirir, Connecticut DOE and Project Analyst
Joe Amenta, Connecticut DOE
Gil Andrada, Connecticut DOE
Janet Stuck, Connecticut DOE
John Olson, Principal Investigator and Project Coordinator
Doug Rindone, TILSA Advisor
Sandra Warren, ASES Advisor
Martha Thurlow, Advisor/Partner
Eileen Ahearn, Advisor/Partner
Phoebe Winter, Project Evaluator
Adam Petermann, CCSSO Liaison

State Validity Study Project Leaders:
Joe Amenta, Connecticut
Gil Andrada, Connecticut
Mohamed Dirir, Connecticut
Ken Draut, Kentucky
Roger Ervin, Kentucky
Joseph Martineau, Michigan
Steve Viger, Michigan
Adam Wyse, Michigan
Carol Crothers, Nevada
Cindy Sharp, Nevada
Andrew Parr, Nevada

In addition, the authors want to thank the state representatives and the technical reviewer, Laura Kramer, for their thorough and thoughtful reviews of the draft technical report.

Acknowledgments (continued)

States participating in the Assessing Special Education Students (ASES) and Technical Issues in Large-Scale Assessment (TILSA) collaboratives, 2007–2010:

ASES	TILSA
Alaska	Arizona
Arizona	Arkansas
Arkansas	California
California	Connecticut
Colorado	Delaware
Connecticut	Dept. of Defense Education Activity (DODEA)
Delaware	District of Columbia
District of Columbia	Florida
Florida	Georgia
Georgia	Illinois
Hawaii	Kansas
Illinois	Kentucky
Iowa	Louisiana
Kansas	Michigan
Kentucky	Minnesota
Louisiana	Mississippi
Maryland	Missouri
Michigan	North Carolina
Minnesota	Ohio
Mississippi	Pennsylvania
Nebraska	Rhode Island
New Mexico	South Carolina
North Carolina	Tennessee
Ohio	Texas
Oregon	Utah
Pennsylvania	Washington
Rhode Island	West Virginia
South Carolina	Wisconsin
South Dakota	Wyoming
Texas	
Utah	
Washington	
West Virginia	
Wisconsin	
Wyoming	

Table of Contents

Acknowledgments……1

I. Abstract…….4

II. Purpose and Overview of the CTEAG Project……5

III. Review of Recent Test Accommodations Literature……...11

IV. Methodology and Analysis Procedures……...27

V. Results and Findings from State Validity Studies……38
Connecticut Study #1…….39
Connecticut Study #2…….47
Kentucky Study…….54
Nevada Study…….61
Michigan Study……..68

VI. Overall Summary and Discussion of Accommodations Validity Studies ……..……75

VII. Additional Information Provided in the CTEAG Guidebook, Project Database, and Lessons Learned Report……..…89

References…….……90

Appendices……96

A. Glossary of Key Terms and Acronyms……..97

B. CTEAG Project Master Schedule……..…98

C. Summary of Project Evaluation……..…….103
.

D. Item and Test Statistics from State Studies…….132

E. Examples of Additional and Post Hoc Analyses…….138

F. Examples of Statistical Analysis Procedures Used…….…141

G. Accommodations Bibliography: Experimental Studies……..149

I. Abstract

The No Child Left Behind Act (NCLB) requires that states offer accommodations on grade-level assessments in order to provide tests that are accessible to as many students with disabilities as possible. However, little research has been conducted on the validity of accommodated score interpretations or the effectiveness of test accommodations. From 2007–2009, with Connecticut serving as the lead state and with the support of the Council of Chief State School Officers (CCSSO), a consortium of 38 states who are members of two of the larger State Collaboratives on Assessment and Student Standards (SCASS) groups—Assessing Special Education Students (ASES) and Technical Issues in Large-Scale Assessment (TILSA)—participated in a special research project called the Connecticut Enhanced Assessment Grant (CTEAG). The project was funded by a grant from the U.S. Department of Education (S368A070009). The research conducted by CTEAG provides empirical evidence to help establish the validity of inferences from accommodated tests based on specific student accessibility needs.

In this Technical Report, detailed information is provided on the research design that was used, the types of accommodations selected and other details of the validity studies that were conducted by states, and the guidelines that were shared with states to help them prepare and organize for the studies. In addition, details are provided on the samples, test instruments, data files, and analyses conducted by project staff, which included statistical analysis of items and test scores, exploratory and confirmatory factor analysis to identify any structural changes in test constructs due to accommodations, and repeated measures analysis of variance to examine main and interaction effects. Detailed results of the analyses are reported and the findings discussed.

Deliverables from the project include this technical report, a separate CTEAG guidebook, a lessons learned report with contextual information on the validity studies, and an associated database that provides student data, procedures, statistical data, and other information for evaluating the validity of test results from accommodated assessments.

II. Purpose and Overview of the CTEAG Project

The main goal of the CTEAG project was to conduct in-depth studies of the validity of score interpretations from selected test accommodations used for students with disabilities. The project's purpose was to empirically examine the effects of accommodations on test performance, comparing students with disabilities (SWD) to students without disabilities (SWOD) and accommodated with non-accommodated test administrations, and to look at the impact, if any, of accommodations on the psychometric characteristics of the items and tests. The products of the project include a guidebook summarizing the findings from the multiple validity studies that were conducted, along with this technical report and an extensive database of information from the research, so that all states can use and refer to the evidence when justifying the use of certain accommodations and the validity of their own accommodated assessments and interpretation of scores.

Because the final reports and database include information on research designs, procedures, and findings from the five validity studies, they are a valuable resource for every state that needs to reference existing data and evidence to inform, evaluate, and support its decisions on the use of various types of accommodations. All CTEAG project documents and materials can be found on the CCSSO website (http://www.ccsso.org). Also, because the results from the states cover a variety of different read aloud accommodations (for example, read alouds done by the teacher or trained test administrator and read alouds delivered by computer), the project provides a resource that few, if any, states would be able to develop by themselves. This compilation and sharing of information will be a great benefit for states in meeting NCLB peer review requirements for standards and assessments by providing additional validity information concerning inferences about student achievement and the performance of students with disabilities.

In addition, information from the collection of CTEAG reports can be used by state staff and others who have decided to conduct validity studies on test accommodations, and for those who are thinking about possibly conducting such a study in the future.

Overview of CTEAG Project

The lead state for the CTEAG project was Connecticut. With the support of CCSSO, a large consortium of states from the ASES and TILSA SCASS groups participated in this special research project, conducted between October 2007 and June 2010 (a list of all participating states in the two SCASS groups is shown in the Acknowledgments). The combined strength of this consortium brought together the expertise and practical experience of state personnel who are working with state

assessments, special education, accommodations decisions, technical issues, and validity evidence to support their programs.

During the course of the CTEAG project, the Connecticut State Department of Education (CSDE) coordinated closely with CCSSO to manage the project's work and the subcontracts with specific researchers. Work on the project was directed by a Project Management Team (PMT) and included experts in assessment, special education, and technical issues. State members from the TILSA and ASES SCASS groups led the validity studies in their states.

CTEAG Project Personnel

During the course of the CTEAG project, Connecticut coordinated closely with CCSSO to manage the project's work and subcontracts with specific researchers. Work on the project was directed by a Management Team led by the project director, Barbara Beaudin. She was assisted by CSDE staff: Mohamed Dirir, Joe Amenta, Gil Andrada, and Janet Struck. Mohamed Dirir also served as the analyst for the project.

The Management Team for this project included experts in assessment, special education, and technical issues. Many had extensive experience working successfully with state collaboratives. Others on the Management Team included John Olson, principal investigator and project coordinator; Doug Rindone, TILSA SCASS Advisor; Sandra Warren, ASES SCASS Advisor; Adam Petermann, CCSSO liaison; and Phoebe Winter, project evaluator. Martha Thurlow and Eileen Ahearn served on the Management Team as advisors and partners to the project.

Members of the TILSA and ASES SCASS groups led the validity studies in their states. They included: Mohamed Dirir and Joe Amenta from Connecticut, Ken Draut and Roger Ervin from Kentucky, Steve Viger and Adam Wyse from Michigan, and Carol Crothers and Andrew Parr from Nevada.

The following table shows the organizational structure for the project.

Organizational Structure and Key Personnel for the CTEAG Project

Connecticut State Department of Education
Barbara Beaudin, Project Director
Mohamed Dirir, Psychometrician and Project Analyst
Joe Amenta, CT State Project Lead
Gil Andrada, CT State Project Lead
Janet Struck, CT State Education Consultant

Participating States and Representatives
Connecticut: Mohamed Dirir, Joe Amenta, Gil Andrada
Kentucky: Ken Draut, Roger Ervin
Michigan: Steve Viger, Adam Wyse
Nevada: Carol Crothers, Andrew Parr

Council of Chief State School Officers
(Primary Contractor)
Doug Rindone, TILSA Advisor
Sandra Warren, ASES Advisor
Adam Petermann, CCSSO Liaison

State Collaborative on Assessment and Student Standards (SCASS)
Technical Issues in Large-Scale Assessment (TILSA) Collaborative
Assessing Special Education Students (ASES) Collaborative

Subcontractors and Partners
John Olson, Principal Investigator and Project Coordinator
Martha Thurlow, Project Advisor and Partner
Eileen Ahearn, Project Advisor and Partner
Phoebe Winter, Project Evaluator

CTEAG Project Timeline (abbreviated version)

A detailed master schedule was developed for the duration of the CTEAG project. The timeline ran from October 2007 through February 2010, with an extension granted through June 2010 to disseminate all products from the project to states. The work of the CTEAG was organized around the regularly scheduled meetings of the ASES and TILSA SCASS groups, when CTEAG project members would come together for meetings and interactions. (Note: A detailed version of the master schedule can be found in the appendix.)

Date	Key Project Activities and Tasks
Oct.- Dec. 2007	• Connecticut awarded Enhanced Assessment Grant • CTEAG project kickoff at combined TILSA and ASES SCASS Meeting in Salt Lake City • Initiate biweekly Project Management Team conference calls to coordinate plans, review questions from states, and solve problems • Survey states for participation in validity studies • Develop "Guidelines for Conducting Validity Studies" document and share with states
Jan.- Mar. 2008	• Hold 1st meeting of CTEAG (in Atlanta) and make presentations at combined TILSA and ASES SCASS Meeting • Refine plans for the development of split-half forms and identification of student samples in each state • States finalize split-half test forms and select samples for Spring '08 administrations

Date	Key Project Activities and Tasks
Apr.-June 2008	• Conduct interim formative evaluation of project • Begin collecting published and unpublished research on test accommodations validity research • At TILSA and ASES SCASS Meetings, review information and progress to date from spring administrations and suggest refinements to analysis plan • Make presentation on CTEAG project at CCSSO National Conference on Student Assessment (NCSA) in Orlando
July-Sep. 2008	• Develop draft outline for guidebook • States finalize test forms and select samples for Fall '08 administrations
Oct.-Dec. 2008	• Hold 2nd meeting of CTEAG (in San Diego) and make presentations to TILSA SCASS on results to date, analysis plans, draft outline for validity studies guidebook, and evaluation feedback • Submit proposal on CTEAG project for presentation at 2009 CCSSO NCSA
Jan.-Mar. 2009	• Hold 3rd meeting of CTEAG (in Orlando) and make presentations to ASES SCASS on results to date and get input on revised draft outline for validity studies guidebook • States finalize test forms and select samples for Spring '09 administrations • Draft summary of accommodations validity research
Apr.-June 2009	• Hold 4th meeting of CTEAG (in L.A.) and make presentations at combined TILSA and ASES SCASS Meeting to share findings from state studies, discuss challenges in conducting studies and lessons learned, review draft information in guidebook • Prepare draft summaries of analyses and findings • Make presentations on results from EAG projects at CCSSO NCSA Pre-session
July-Sep. 2009	• Begin writing guidebook with actual data and results from studies included • Begin drafting technical report and project database contents • Conduct 2nd interim formative evaluation of project

Date	Key Project Activities and Tasks
Oct.-Dec. 2009	• Complete draft of guidebook and share draft reports with state members and other interested parties for review and feedback • Hold 5th meeting of CTEAG (in St. Louis) and make presentations at combined TILSA and ASES SCASS Meeting on findings from state studies and get feedback on draft reports and outlines • Conduct formative evaluation of project • Make presentation on CTEAG project at 2009 Northeastern Educational Research Association (NERA) symposium
Jan.-Mar. 2010	• Disseminate Validity Studies Guidebook • Hold 6th meeting of CTEAG (in New Orleans) and make presentations at combined TILSA and ASES SCASS Meeting on project status and get feedback on plans for final reports and data • Complete drafts of technical report and database and share with state members and other interested parties for review and feedback • Conduct final evaluation of project
Apr.-Jun. 2010	• Finalize technical report and database for publication • Hold 7th and final meeting of CTEAG (in Detroit) with combined TILSA and ASES SCASS Meeting and special workshop to disseminate products, highlight findings, and instruct states on how to best use the information from this project • Make presentation on findings from validity studies at CCSSO NCSA Session in Detroit • Post all materials on CSDE and CCSSO websites • Prepare final project report for ED • Prepare final project evaluation

Summary of Approach Used for CTEAG Project

As described earlier, the purpose of this project was to conduct research on the validity of various types of commonly used test accommodations. Validity was evaluated by comparing the accommodations' effectiveness for special education students who require the accommodation to their effectiveness for general education students who do not typically use the accommodation. The specific impact of the use of accommodations on test scores was analyzed. (More details on the research design, methodologies, and analysis procedures are included in a subsequent section of this report.)

- The lead state for this grant was Connecticut.

- The other states that volunteered to conduct validity studies were Kentucky, Michigan, and Nevada.
- A project management team comprised of CSDE staff, researchers, experts in the field, and representatives from CCSSO and states oversaw and coordinated all project activities.
- CCSSO provided logistical support, key staff, and consultants to assist with activities for the project.
- The TILSA and ASES SCASS groups collaborated on the project. TILSA provided the psychometric and measurement "thought leadership" and ASES provided the special education content and process "thought leadership" for this effort.
- The individual states that volunteered to conduct validity studies each carried out a study in their state, assisted in the development and administration of the specific accommodated assessment, and provided data to the project's statistical analyst.
- CTEAG project staff conducted all analyses, summarized the results, and wrote the reports for the project.

III. Review of Recent Test Accommodations Literature

The *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999; "the joint standards" or "the standards") provides important recommendations for documenting the validity of assessment accommodations. These are very similar to NCLB requirements for technically sound assessments, as described in the legislation and peer review guidance. Chapter 10 of the joint standards provides background information on testing individuals with disabilities and addresses many issues related to accommodations, strategies of test modification, and issues in reporting scores. An underlying principle in the standards is ensuring the validity of the assessments. As stated in Standard 10.1: "In testing individuals with disabilities, test developers, test administrators, and test users should take steps to ensure that the test score inferences accurately reflect the intended construct rather than any disabilities and their associated characteristics extraneous to the intent of the measurement" (p. 101).

The standards define an accommodation as "an action taken in response to a determination that an individual's disability requires a departure from an established testing protocol" (p. 101). Over 70 different accommodations in eight categories (i.e., motivation, assistance prior to testing, scheduling, setting, directions, assistance while testing, use of equipment/adaptive technology, and changes in format) have been identified in the testing literature (Elliott, Kratochwill, & Schulte, 1998). Among the issues at the forefront of conversations in many testing contexts today are how states will report progress for students with Individualized Education Programs (IEPs) as part of NCLB, what accommodations are used, how they are implemented, and to what extent scores are comparable from tests administered under standard and non-standard conditions.

Federal and State Policy Issues Related to Assessment Accommodations Research

Accommodations have been, and continue to be a topic of heightened interest for a number of reasons. Participation in national, state, and district assessments is required by both the federal special education law, the Individuals with Disabilities Education Act (IDEA). Furthermore, the federal Elementary and Secondary Education Act (ESEA), reauthorized and amended as the No Child Left Behind Act (NCLB), requires universal participation in state assessments and the National Assessment of Educational Progress (NAEP). First and foremost, it is no longer a question of whether students with disabilities need to be participants in national, state, and district assessments. The issue before assessment experts and state policymakers now is *how* to ensure inclusion. Many students with disabilities require accommodations; data from the 2003–2004 Annual Performance Reports submitted by states to the federal government indicated that, on

average, over 60% of school-age students with disabilities used accommodations during assessments (Thurlow, Moen, & Altman, 2006).

According to the Common Core of Data from the National Center for Education Statistics, in the 2004-2005 school year nearly 6 million of 48.7 million students in the U.S. had IEPs (NCES, 2006). In many cases the disabilities that prompt these IEPs make it difficult for some students to perform to their full potential on tests under standard conditions. Thus, while not an exact barometer of test accommodation use, these statistics do indicate that on average across the states about 13% of elementary and secondary students have had teams of educators and specialists determine whether or not a student with a disability needs an accommodation. Clearly, one approach to assessment cannot always fit all students' needs because test-takers across many testing contexts often vary by more than just proficiency. The presence of one or more disabilities can impact how students interact with and complete tasks in a testing situation. The use of test accommodations is often a necessity, as is the need for research-based policy to guide practice.

As the field has moved forward, both in terms of policy development and in terms of research, it has been pushed at the same time by federal activities in which state assessment systems are undergoing reviews by the United States Department of Education's via the Standards and Assessment Peer Review process. The peer reviewers are assessment and policy experts from across the nation who review each state's submission and make recommendations to the USED. During peer review, each state supplies documentation that the state's assessment system is in compliance with required technical and policy elements from NCLB. One requirement of the peer reviews is that students with disabilities will participate in assessments with accommodations if needed, and thus, accommodations provided by states are also scrutinized as part of the assessment system. Some of the specific aspects of the peer review guidance (U.S. Department of Education, 2004) that address accommodations for students with disabilities are Critical Elements 4.3 and 4.6:

> ***4.3*** *Has the State ensured that its assessment system is fair and accessible to all students, including students with disabilities and students with limited English proficiency, with respect to (each of) the following issue(s):*
>
> *(d) Does the use of accommodations… yield meaningful scores? (p.37)*
>
> ***4.6*** *Has the State evaluated its use of accommodations?*
>
> *(b) How has the State determined that scores for students with disabilities that are based on accommodated administration conditions will allow for valid inferences about these students' knowledge/skills and can be combined meaningfully with scores from non-accommodated conditions? (p. 40)*

State policymakers, practitioners, and researchers define accommodations in a variety of ways. For example, Thurlow and Bolt (2001) defined testing accommodations as "changes in assessment materials or procedures that address aspects of students' disabilities that may interfere with the demonstration of their knowledge and skills on standardized tests. Accommodations attempt to eliminate barriers to meaningful testing, thereby allowing for the participation of students with disabilities in state and district assessments." (p. 3).

A major challenge surrounding accommodations is determining which changes to test materials, settings, and procedures will or will not result in alterations to the intended construct. Historically, this concern has not always been a top priority. For some time, the major focus was on the participation of students in the assessments—accommodations were seen as an avenue to participation, and the distinction between whether a change compromised the construct being assessed was not necessarily the major factor of interest (Thurlow, 2007). It has taken time for the field to adjust its focus to clearly distinguish between accommodations (those changes in materials, setting, or procedures that do not result in changes to the construct being measured) and modifications (those changes in materials, setting, or procedures that do result in changes to the construct being measured), and for research and policy to address these concepts.

Summary of Recent Accommodations Research

The body of research on test accommodations has continued to grow over the years, although it is still relatively small. A summary of recent research on the topic can be found in an online bibliography from the National Center on Educational Outcomes (http://www.cehd.umn.edu/NCEO/Things%20to%20delete/oldAccomStudies.htm). As Martha Thurlow stated in a recent American Educational Research Association paper summarizing research findings on accommodations (2007, p. 2): "There are many definitions of testing accommodations, yet all now converge on the same concepts. In general, any type of change to testing materials, setting, or procedures that does not alter what is being measured is considered to be an accommodation."

Some researchers refer to the elimination of construct irrelevant variance—in other words, the variance associated with extraneous features of test administration (Fuchs, Fuchs, Eaton, Hamlett, & Karns, 2000). Demonstrations of test changes that maintain the validity of the construct that an assessment is attempting to measure have been challenging for the field, and numerous approaches to research have been pursued (Thurlow, McGrew, Tindal, Thompson, Ysseldyke, & Elliott, 2000; Sireci, Scarpati, & Li, 2005; Tindal, 1998), including single subject designs, "boost" studies, and "differential boost" studies.

Two reports from the National Center on Educational Outcomes (NCEO) provide excellent summaries of research studies on the effects of test accommodations (NCEO,

2006, 2007). Note that a forthcoming NCEO report summarizing more recent research on accommodations was in the process of being written in 2010 but was not yet available for use in this report. In "*A summary of research on the effects of test accommodations: 2002 through 2004*" (Technical Report 45) by Johnstone, Altman, Thurlow, & Thompson (2006), the authors present the case that "as states and school districts strive to meet the goals for adequate yearly progress required by NCLB, the use of individual accommodations continues to be scrutinized for effectiveness, threats to test validity, and score comparability" (p. i). The report summarizes 49 empirical research studies completed on test accommodations between 2002 and 2004, and provides direction in the design of critically needed future research on accommodations.

The primary purpose of the summary of 2002-2004 accommodations research was to determine the effects of accommodations use on the large-scale test scores of students with disabilities. In their executive summary, Johnstone et al. described the studies as having certain defining characteristics. The majority of the studies tested students using norm-referenced or criterion-referenced tests, in the content areas of math or reading/language arts. Participants were varying percentages of students without disabilities and students with disabilities. Students with learning disabilities were studied most frequently among students who receive special education services. Equal numbers of research studies involved fewer than 100 participants, 100-1,000 participants, and more than 1,000 participants of multiple age categories. Individual accommodations showed either differential item functioning or no differential item functioning depending on the study. The findings shared no common theme, with various accommodations shown to have both a positive and non-positive effect on scores.

Following is an overall summary of the findings from individual studies on test accommodations done in 2002-2004, focusing on the effects on test scores and score comparability.

Studies of the Effects of Accommodations on Test Scores

As noted in the report, studies that used experimental and extant data analysis methods typically had two main goals: understanding the effect of accommodations on test scores and understanding the effects of accommodations on the psychometric qualities of items. Among the research that examined these types of effects, those that found a positive effect on the test scores of at least some of the SWD who were sampled included studies of the following accommodations: computerized administration (Pomplun, Frey, & Becker, 2002), read aloud accommodation (Helwig, Rozek-Tedesco, & Tindal, 2002; Meloy, Deville, & Frisbie, 2002), video administration (Burch, 2002; Tindal, 2002), extended time (Bridgeman, Cline, & Hessinger, 2004), and assistive technology (Landau, Russell, Gourgey, Erin, & Cowan, 2003; MacArthur & Cavalier, 2004). Two studies found that some students with learning disabilities did not benefit from computer or video accommodations (Burch, 2002; Barton, 2002) and another found that calculator usage did not have an effect on student scores (Scheuneman, Camara,

Cascallar, Wendler, & Lawrence, 2002). Two other studies found that extended time also did not yield improved scores for students with disabilities (Elliott & Marquart, 2003, 2004).

Studies of Item and Score Comparability

NCEO also reported that researchers found mixed results in terms of score comparability for items based on analyses of the psychometric properties of test accommodations. When various accommodations including read aloud, extended time, and computerized administration were studied, Barton (2002), Barton & Huynh (2003), Cahalan, Mandinach, & Camara (2002), Huynh, Meyer, & Gallant-Taylor (2002), and Kobrin & Young (2003) found no change in item comparability. However, Bolt & Bielinski (2002), Choi & Tinker (2002), and Thornton, Reese, Pashley, & Dalessandro (2002), all found that tests administered orally, with extended time, or via computer resulted in changed item difficulty or constructs.

Studies of Read Aloud Accommodations

Mixed results were found from the 11 studies on read aloud accommodations. For example, Helwig et al., (2002), Huynh, Meyer, & Gallant (2004), Janson (2002), Meloy, Deville, & Frisbie (2002), Tindal (2002), and Weston (2003) all found that read aloud accommodations had a positive effect on scores for SWD. However, Bolt & Bielinski (2002) and McKevitt & Elliott (2003) found read aloud accommodations had no significant impact on student scores.

In terms of item comparability, two studies found items to be comparable, whether presented under standard conditions or with read aloud accommodations (Barton, 2002; Barton & Huynh, 2003), whereas three studies found that read aloud accommodations did have an impact on item comparability (Bolt & Bielinski, 2002; Meloy et al., 2002; and Weston, 2003).

Overall, several general patterns emerged from NCEO's review of accommodations research in the years 2002-2004. The majority of research concentrated on the effect of accommodations use for SWD and the effects on score validity due to accommodations use. Although there were 36 studies in total that investigated scoring and validity, there was little consensus among researchers and findings were sometimes contradictory. Research indicated that accommodations were either beneficial or not beneficial for SWD. Likewise, researchers did not reach consensus on whether accommodations changed the construct of the item assessed. Therefore, because of the lack of consistent findings and because some results were relatively disparate, the authors suggested a need for further research.

From its analysis of 49 studies from 2002-2004, NCEO recommended that future research needed:

- a clear definition of the constructs tested in future research,
- a reduction in confounding factors,
- increased study of institutional factors affecting accommodations judgment, and
- exploration of students' perceptions on the usefulness of accommodations.

Finally, the authors also recommended that future research should focus on improvements in all these areas, but also on the positive effects of field-testing potential items in accommodated formats in addition to standard formats.

A more recent summary of research on the topic, "*A summary of the research on the effects of test accommodations: 2005-2006*" (Technical Report 47) by Zenisky & Sireci (2007), states that NCLB has had the effect of "an increasing convergence of longtime policy and psychometric discussions about the use of various test accommodations and score interpretations from accommodated and non-accommodated administrations. At the same time, much work remains" (p. i). The report provides an update on the state of the research on testing accommodations and identifies promising areas of research to further clarify and enhance understanding of emerging issues. The authors present practitioners and researchers with a number of insights into both the current state of research on test accommodations and the directions that future research might take. They found that most of the research published in 2005-2006 fell into one of two categories:

(1) empirical studies of student scores from assessments administered under accommodated and non-accommodated conditions, and
(2) research activities that were more descriptive in nature, aimed at identifying the accommodations used with different test populations or how accommodations use is perceived by different stakeholder groups.

The technical report summarized research on the effects of test accommodations in 2005 and 2006, and found 32 published research studies on the topic. Among the main points highlighted in the report were:

- **Purpose:** The majority of the research that was included sought to evaluate the comparability of test scores when assessments were administered with and without accommodations.
- **Types of assessments, content areas:** Mathematics and reading were the most common content areas involved in the 2005-06 research, and a wide variety of assessment types were used in these studies.
- **Disabilities and accommodations:** Generalized learning disabilities were the most common disabilities identified among participants in the research,

accounting for nearly half of the studies. Extended time (alone and bundled with other accommodations) was the single most studied accommodation, but oral accommodations (such as read aloud and audiocassette presentation) were also considered in multiple studies, as was computerized administration.

- **Research design:** Over 70% of the studies reported primary data collection on the part of the researchers, rather than drawing on existing archival data sets. Almost half of the studies involved experimental or quasi-experimental designs.
- **Findings:** Most of the oral presentation and timing accommodations empirically tested were found to have positive effects on scores, although some studies reported no effects.

Following is an overall summary of the findings from the individual studies on test accommodations that were published in 2005-2006, focusing on the characteristics of the studies that were done, research on the interaction hypothesis, and recommendations for additional accommodations research.

Types of Studies

Ten of the studies carried out in 2005-2006 evaluated the comparability of scores from standard and non-standard administrations and included both SWD and SWOD in their samples. Many implemented the full range of designs identified in Thurlow et al. (2000). Of the non-experimental studies that were conducted, most were surveys, but the research also included case studies and observations of assessment practices. Over half of the research studies used primary data in their investigations rather than drawing on extant data sets. Among the academic measures used in the research, some were state assessments used for NCLB purposes, but many studies involved norm-referenced assessments, such as the TerraNova (Gibson, Haaeberli, Glover, & Witter, 2005; Kettler, Niebling, Mroch, Feldman, Newell, Elliott, Kratochwill, & Bolt, 2005; Lang, Kumke, Ray, Cowell, Elliott, Kratochwill, & Bolt, 2005) or the Stanford Achievement Test. Similar to previous summaries of accommodations research (Johnstone, Altman, Thurlow, & Thompson, 2006; Sireci et al., 2005), the domains of mathematics and English language arts (reading, writing, and other related skills) were the most frequently studied content areas.

Types of Accommodations Studied

The 32 research studies identified in 2005-2006 reported a wide variety of accommodations used with different student populations. However, just seven specific types of accommodations were empirically studied and those were quite narrowly focused primarily in two categories (presentation and timing/scheduling). The authors

noted that this finding was in contrast to earlier summaries of accommodations research by Johnstone et al. (2006) and Thompson, Blount, & Thurlow (2002), where 11 different accommodations within four categories were reported as being studied empirically in each of those two reviews. Extended time was the most frequently studied accommodation in 2005-2006 and the results were generally consistent with those from previous research literature that had shown extended time to have a positive effect on the scores of SWD. However, it is important to note that most, if not all, state assessments have moved toward the use of untimed tests for all students which has minimized if not eliminated the need for further study of the benefits and effects of extended time test accommodations.

Research on the Interaction Hypothesis

Sireci, Scarpati, & Li (2005) explained the validity of accommodations through an "interaction hypothesis" or the theoretical assumption that test accommodations will lead to improved test scores for students who need accommodations, but little or no improvement for students who do not need accommodations (i.e., SWD receive a boost in scores as a result of accommodations whereas SWOD do not receive a boost or receive a less pronounced boost in scores). In other words, this hypothesis proposes that SWD will benefit to a greater extent from accommodations than SWOD, i.e., there will be an interaction effect.

The interaction hypothesis was the topic of the article by Sireci et al. (2005), in which evidence supporting an interaction with respect to extended time was compiled. A revised hypothesis was based on the finding that both SWD and SWOD benefited from extended time, but the SWD exhibited relatively greater score gains. This revision is consistent with differential boost theory (Fuchs & Fuchs, 2001; Thompson et al., 2002). Note that more discussion of the differential boost theory is presented later in this chapter.

The empirical results reported by Fletcher, Francis, Boudousquie, Copeland, Young, Kalinowski, & Vaughn (2006); Lesaux, Pearson, & Siegel (2006); and Kettler et al. (2005) provided support for the idea that SWD needed accommodations and benefited from their use, while SWOD did not benefit from them (at least not to the same extent). In Fletcher et al. (2006), only SWD benefited from the use of the orally administered test given in multiple sessions, while Lesaux et al. (2006) and Kettler et al. (2005) found similar results for the extended time and various IEP-assigned accommodations, respectively.

Additional research that continues to delineate the interaction hypothesis and that reduces construct irrelevant variance for SWD without introducing any new effects for SWOD still appears to be necessary. Because accommodations represent departures from the standard testing protocol and typically were considered to benefit only SWD for whom they are deemed appropriate, the authors suggest that future studies should

continue to implement research designs that explicitly address the interaction hypothesis and differential boost to inform practice.

Recommendations for Research that Connects to Policy and Practice

The reviews of test accommodations issues completed by Sireci et al. (2005), Sireci (2005), and Stretch & Osborne (2005), respectively, were focused mostly on the interaction hypothesis, score comparability and interpretation, and extended time accommodations, but together offered many important directions for future study. As described in the NCEO technical report, "how accommodations are operationalized is one area where greater definition or clarification may be warranted, as is improved guidance for users of scores from accommodated and non-accommodated administrations about appropriate test score inferences. In addition, while many of the studies reported that the use of accommodations had some positive effect on test scores, it was noted that variations across studies in the operational definitions of those accommodations does challenge the extent to which findings can be generalized across studies. Furthermore, while much work has been done, another challenge for research is to construct true experiments to assess the effects of accommodations' use on test scores and their consequences for students with and without disabilities alike." (p. ii)

In their summary, Zenisky & Sireci (2007) consider what has been learned from accommodations studies done in 2005-2006 and previous years to identify how the research findings to date can offer advances in knowledge about the effects of accommodations, as well as inform policy and practice. They note that in 2005-2006, as in previous years, the wide variations across operational definitions, tests, populations, settings, and contexts still limit policy implications to all but the most general conclusions and interpretations, and although decisions on the use of accommodations increasingly involve high-stakes consequences, interpreting scores from accommodated and non-accommodated administrations remains, in many cases, as much art as science.

In addition, because the assessment policies of NCLB require that all students are included in state assessments and that disaggregated results for SWD and other groups be reported, it also is important that valid measures of students' performance be obtained. In particular, for many SWD, valid measurement means providing accommodations that do not change the construct measured and make the test more accessible. The authors emphasized that the need to better understand what the research on test accommodations tells us is more important than ever before. Thus, it will be essential to continue to review and summarize the research conducted in this area and to question whether changes in assessment and accommodations policies need to be made.

Recent Research on Computer-based Accommodations

While computerized administration is increasingly being considered for use across many different testing contexts, the research on different aspects of computer

technology as a test accommodation is not yet conclusive, as is the case with much of the research on paper and pencil test accommodations. This is due in part to ongoing operational challenges of implementing computer-based tests in practice or for research purposes. Nevertheless, computers hold much promise for allowing students to use innovative formats and tailoring the presentation of the test to their individual needs (e.g., magnifying text, pacing in audio presentation, digital read alouds).

As reported in Johnstone et al. (2006), findings from investigations of the computer as an accommodation have not been definitive. However, they found that the presentation accommodation of scrolling or paging through passages did not seem to have any effect on student scores one way or another, but further study seemed warranted for comparing the effects for SWD and SWOD, rather than only SWD. Because of the many ways that computer-based tests can be designed and administered for different purposes and content areas, a focused program of research on operationally defining and evaluating computerized assessment accommodations is needed. The review by Meyen, Poggio, Seok, & Smith (2006) on the use of computerized-adaptive testing as a strategy for testing SWD likewise provided an important direction for future research, but the authors suggested that computer use should be implemented carefully with respect to universal test design and with the goal of minimizing construct-irrelevant variance.

In a recent paper entitled, "Computer-Based Read Aloud Accommodation: Comparative Effect on Student's Mathematics Test Performance" (Russell, Higgins, Miranda, & Hoffmann, 2009) an ongoing research program on read aloud accommodations is described. The authors acknowledge that use of this accommodation often presents several logistical challenges to large-scale standardized testing, including finding sufficient test proctors and classroom space to provide the accommodation individually for students, enabling students to control the pace of testing and frequency with which portions of an item are read aloud, and ensuring that reading is provided in a high-quality manner. Russell et al. propose that computer-based technologies hold much potential for overcoming many of these challenges, and in their paper, present findings from a pilot study of the effects for a computer-based read aloud accommodation on the mathematics test performance of high-school students. In their study, students were randomly assigned to one of three conditions to compare the effect of (1) a computer-based read aloud to (2) a human delivered read aloud and to (3) a test delivered without the read aloud accommodation. The findings indicate that test scores for students with special needs related to reading, as well as scores for ELL students, increased when a computer-based read aloud was provided as compared to the human delivered accommodation.

The authors state that the research provides preliminary evidence that a computer-delivered read aloud accommodation is both feasible and effective for the targeted population. Based on the findings from their pilot study, they recommended additional research be conducted that would focus specifically on the effects of computer-based delivery of the test versus computer-based delivery of the read aloud accommodation.

Secondly, while this research is being conducted, they suggested that state assessment programs also should begin exploring the feasibility of incorporating a computer-based read aloud accommodation into their testing programs. Thirdly, given that they found the read aloud accommodation had a positive effect on the performance of general education students, it was also recommended that research be conducted on whether validity would be increased by allowing all students to have test items read aloud (if the students desire) when the construct being measured is anything other than reading skills or reading comprehension.

Recent Resources and References from the ASES SCASS for Addressing Accommodations Issues

In 2006, the ASES SCASS Accommodations Monitoring workgroup discussed ways to provide more information to states about monitoring of accommodations as required by NCLB and as addressed in the guidance and requirements for peer review. Several reports were issued from the work of this group. Working in conjunction with NCEO, a quick reference document entitled "Hints and Tips for Addressing Accommodations Issues for Peer Review" (Christensen, Lail, & Thurlow, 2007) provided a summary of findings from peer review notes of acceptable evidence of accommodations monitoring as well as recommendations for best practices on the selection of accommodations. The information can be useful to states in responding to accommodation criteria in the Standards and Assessment Peer Review process and the need for continued improvement after approval. Examples of evidence that are acceptable as well as insufficient for the accommodations elements are highlighted in the document. States were encouraged to consider the recommendations of best practices for assessment accommodations for students with disabilities. Among the critical elements that are addressed in the document is the issue that accommodations use needs to provide meaningful scores and valid inferences about student knowledge and skills. This issue is pertinent to the research conducted by the CTEAG and the type of evidence that the project provides to states.

Another useful product that came recently from the work of the ASES SCASS Accommodations Monitoring workgroup and the collaborative effort between the ASES SCASS and NCEO was the report "Improving Accommodations Outcomes: Monitoring Instructional and Assessment Accommodations for Students with Disabilities" (Christensen, Thurlow, & Wang, 2008). This document presents a process for states, districts, and schools to use in monitoring accommodations for instruction and assessment and it provides guidance on serving SWD who participate in large-scale assessments. In addition, guidelines were developed for states to use and examples of best practice provided so that accommodations can be used appropriately by all states. Among the conclusions in the document are that states need to evaluate their policies, practices, and procedures for accommodations, and the evaluation may include a review

of the current research on accommodations and analysis of current state data. This information can then be used by the state to make improvements that reflect the needs of the state.

Research Designs and Approaches for Studying Test Accommodations

In the recent book *Large Scale Assessment and Accommodations: What Works?* (Laitusis & Cook, 2007) Cara Laitusis of the Educational Testing Service authored a chapter on the variety of approaches for studying accommodations on assessments (Chapter 6, ibid, 2007). In this chapter, she provides policymakers, researchers, and test developers with a summary of research-based approaches for determining the validity and reliability of test scores obtained under accommodated conditions. The chapter reviews the design of research studies that can be used to examine the validity of accommodated and standard test scores for SWD and SWOD as well as other studies that can be conducted to improve test design and accommodation policies for SWD. The topics and areas described in the Laitusis chapter include various types of special research studies, small sample designs, research designs to examine psychometric comparability, and differential boost designs.

According to Laitusis, researchers are using a variety of types of research studies that focus on improving accommodation policy and practice. These types of studies include field testing of new item types and test forms on samples of SWDs, use of cognitive labs or think aloud studies, and scaffolding test accommodations to examine their impact at the item level. For example, Laitusis proposes small field tests that include accommodations as a cost-effective way to identify problems with new item types prior to their operational administration. These small studies also provide a means to collect several types of survey data, such as information on teacher or student perceptions, accessibility, and data on SWD. These data can then be used to inform individuals who make decisions about testing accommodations at both the individual and policy levels. Additionally, these data and improved policy decisions can subsequently be used to inform test development, to create alternate test formats, or to determine best practice on the implementation of testing accommodations.

Cognitive labs, or think aloud studies, have been used to examine issues of test design for SWD as well as ELLs. This methodology provides a useful means for investigating how individual test takers approach a particular item type or use of a test accommodation. In addition, another type of research study that has been used recently is scaffolded accommodations, which obtains information on students' performance under standard conditions and then asks the student to attempt the item again with an accommodation. This type of design can provide additional information on the student's meta-cognitive skills in determining if they can complete the item without accommodation.

Psychometric Indicators for Research on Test Accommodations

As described by Laitusis (2007), a variety of psychometric indicators are useful to include in the research designs and can be examined to provide testing programs with additional evidence on the impact of testing accommodations. In 1982, the National Academy of Sciences recommended five psychometric indicators for examining effects for accommodations, which included:

1. Test reliability
2. Test factor structure
3. Item functioning
4. Predictive validity
5. Impact of accommodations on admission decisions

(Note that the last item is not as relevant today since many testing companies have decided to no longer flag test scores for "non-standard administrations.")

In 2006, the USED issued a paper that addressed the validity of test scores and provided a framework for examining validity using seven different types of empirical validity evidence. This evidence included the first four NAS recommendations above as well as evidence from test design and development, scoring, and scale comparability studies. Much of the data for use as psychometric evidence from these two sets of recommendations can be collected during the test administration. Descriptions of several common types of psychometric analyses that can be done are presented below.

Analysis of test reliability can be used to compare the effects of accommodations for subgroups of the test taking population, such as SWD and SWOD. A number of different methods can be used to measure test reliability which, in general, is a measure of how consistently examinee's scores are rank ordered across administrations of a test. The most common metrics are measures of internal consistency reliability, split-half reliability, and alternate forms reliability. However, regardless of which metric is used, it is important to ensure that the degree of error in an assessment is appropriate for the purpose of the test and that subgroups of students are not impacted disproportionately.

A study of a test's internal structure also can be part of the test score validation process. For valid test score interpretation, scores on an accommodated test form given to SWD should measure the same underlying construct(s) as scores for SWOD who have taken the test under standard conditions. Factor structure can be analyzed for both SWD and SWOD, and for both accommodated and non-accommodated conditions. Typically, a series of exploratory and confirmatory factor analyses are conducted using either item level data or parcel data. An example of "parceling" data is to combine items from a particular strand or competency, such as grouping "algebra" items into a parcel when analyzing data from a test of general mathematics. Several researchers have used factor analysis in recent studies to investigate the structure of an assessment for SWD (Cook, Eignor, Sawaki, Steinberg, & Cline, 2006; Huynh & Barton, 2006; Huynh, Meyer, &

Gallant, 2004). Factor analytic types of studies are important because they provide valuable evidence that a test does or does not measure the specified construct for a particular group of students.

Analyses of item functioning and statistical characteristics, such as item analysis and Differential Item Functioning (DIF) are additional approaches that are useful for examining what the test is measuring for different populations. Item statistics, such as classical item difficulty and discrimination indices, can be compared for tests administered to SWD and SWOD, under accommodated and non-accommodated conditions. Although there are different procedures for conducting DIF analyses, these methods all examine the difference in item performance between two comparable groups of examinees, usually matched on total test score. DIF is typically conducted to compare male-female or White-Black groups, but it is possible for research studies to use this methodology with disability subgroups. However, sample sizes should not be less than 100 in each group, and preferably at least 300, in order to defensibly analyze DIF.

Differential Boost Model

In earlier years, researchers studying the effects of accommodations proposed that for the accommodation to support valid inferences, it should have a positive effect on scores of SWD and no effect on the scores of SWOD (Phillips, 1994). If the accommodation increased scores for both groups of students, the general conclusion at that time was that the accommodation provided an unfair advantage to some of the students who used it. This reasoning was based on the idea that students without disabilities did not have access needs that were addressed by the accommodation and any increases in their scores were probably due to a changing of item difficulties or the construct targeted by the test.

In recent years, research has progressed and the paradigm has shifted to one of "differential boost" and in particular, research on student access to test content and opportunity to perform (Laitusis & Cook, 2007). Researchers have proposed that an accommodation may support valid inferences if students with disabilities have a relatively greater increase in scores with the accommodation than students without disabilities. Just because the scores of students without disabilities also increase does not automatically invalidate the accommodation (Sireci, Scarpati, & Li, 2005). Recent research has acknowledged the fact that students who are not identified as having a disability may still have access needs that are alleviated by the accommodation. For example, many poor readers are not identified as having a communication-related learning disability, but they may still benefit from an accommodation in which all or part of a test is read aloud to them.

Laitusis focused attention on determining if the outcome of an accommodation used on a particular assessment yields valid test results and if it removes construct irrelevant variance to improve the validity of the test score. She argues that the differential boost model, or the interaction model, may be a preferred approach to use in

interpreting the effects of accommodations. Differential boost indicates that SWD receive significantly larger gains in overall test score from an accommodation than SWOD and, although both groups can receive improved test scores by using the accommodation, the gains for SWD are differentially larger. The interaction model has a similar premise and may use a repeated measures analysis of variance methodology to examine for a significant interaction between disability classification and accommodation when all students have taken the test under both accommodated and non-accommodated conditions. It is especially important to take this perspective into account when reviewing the findings from the CTEAG validity studies.

Conclusion

The approach that the CTEAG used in its studies takes many ideas from prior research as well as from work on test accommodations that is being done currently. Based on a review of recent research literature on accommodations, an evaluation of various recommendations that have been made regarding the need for additional research on the impact of test accommodations for the validity of test scores, and conversation with various experts on accommodations and state assessment staff on the kind of data that would be most useful to them, the CTEAG developed plans for conducting collaborative research in this area. The CTEAG approach used an experimental research design that examined differential performance boost using an interaction model. States identified accommodations that were important for them to study in the context of their own assessment policy and practice. A collection of psychometric indicators was used to analyze the effects of the accommodations. Additionally, the intent of the CTEAG project was to build upon the existing body of research, expanding the body of knowledge that currently exists on test accommodations and transferring principles found to be successful in strengthening the validity of state assessments. Indeed, given the policy emphasis that has been placed on the performance of subgroups of students in state and national assessments, the need for contributions to policy and psychometric understanding of the issues surrounding the use of test accommodations from researchers who are empirically studying these issues is at a critical point.

In the next chapter, the research design, methodology, and analysis procedures used in the CTEAG studies are described in detail.

IV. Methodology and Analysis Procedures

In this chapter, the research methodology used in the CTEAG studies is described in detail along with the procedures used for the research design, test instruments, data collection, and data analysis. By design, the five studies had many features that were common across the states, including the research questions addressed and data analyses used. These common issues are presented as a unit for all studies. On the other hand, design issues that are different from study to study are presented for each specific study. Features that may be unique to an individual state's study include the grade studied, test forms, samples, and the accommodation studied.

Research Questions

The CTEAG research examined three basic questions addressing three important aspects of test validity: item and test characteristics, test content structure, and accommodation effects on student performance. The first question that the studies sought to answer was whether test accommodations affect item statistics and test properties. This important question focuses on the building blocks of the data gathered for the study. Understandably, an accommodation that alters item properties could lead to inaccurate item scores and invalid test results. If the accommodation makes the item much easier, for example, it could unfairly favor the examinees who took the item under the accommodated condition. Similarly, changes in item discrimination due to accommodation could reduce that item's ability to correctly rank examinees. Any effects of the accommodation on the items would be expected to similarly impact the characteristics of the test. Test-level effects could include changes in average performance, variability, and reliability. Consequently, these effects could lead to invalid comparisons of test scores from the different administration conditions.

The second research question explored in the studies was whether the accommodation affects the dimensionality of the test or the test construct. In practice, tests are usually unidimensional, meaning they are designed to measure a single concept or trait. Restricting tests to measure one area/construct stems from the need to sharpen the focus of the tests so that the inferences and decisions to which the test results lead are well defined. Moreover, most assessment models assume that the assessment data are unidimensional. Hence, an accommodation that alters the dimensionality or the construct of the test may not be defensible because the accommodated test and the non-accommodated test may no longer be strictly comparable.

The third question examined by the CTEAG studies was whether the test accommodation affected student performance—specifically, whether the accommodation increased performance of SWD more than it increased performance of SWOD, and if so by how much. One of the objectives in examining this question was to assess the extent to which accommodations serve the purposes for which they were developed, i.e.,

increasing accessibility to the test for students with certain types of disabilities. To approach this question, the project examined the impact of accommodations on student performance in each of the five studies. Another objective related to this question was to examine if the accommodations gave an unfair advantage to students who used them over students who do not use them.

Research Design

The following description of the research design will permit users and researchers to be better able to interpret results from the validity studies. Information in this section includes a brief mention of the requirements for the studies and the guidelines provided by the project to support states in conducting their studies.

The intent of the project was to have a variety of different accommodations, grades, and content areas selected for the studies so that validity evidence could be gathered under a variety of conditions for the project as a whole. Each of the participating states first decided what accommodation to study and then determined the content area and grade level of the test used in the study.

Each state was guided to use the same 2 × 2 counterbalanced research design that had students (SWD or SWOD) crossed with accommodations (with or without) on the test forms used in their state. In each state, one form was selected for the study. The form was then divided into two parallel half-forms. Both groups of students took both half-forms under both conditions of administration, but each student took each half-form under one administration only. Further, the order of presentation of the administration conditions (accommodated half-form vs non-accommodated half-form) was randomized for most of the studies. This was intended to cancel out any order effects that might affect the outcomes (see the following chart).

Accommodations		**Students**	
		Students with Disabilities	Students without Disabilities
	Without	Split half A Split half B	Split half B Split half A
	With	Split half B Split half A	Split half A Split half B

To assist the states in preparing for the validity studies, detailed guidelines were developed and distributed. The guidelines provided steps for states to follow in preparing for and conducting their respective studies. The main goal was to standardize the research

studies as much as possible. The following document was shared with participating states in November 2007.

Guidelines in Planning for the Accommodations Validity Studies: Steps for States to Follow in Preparing to Conduct a Study

A. Building the project test forms for the study.

1. If at all possible, the participating state should use an existing form of the state test (such as a released form, alternative form, breach form, etc.). Another option is to assemble the project test forms from a pool of existing items that the state already has. If a state does not have an existing form or pool of items that are available, staff should contact the project to discuss possible ways to acquire a test for use in the study. Mohamed Dirir and John Olson will provide advice as necessary to state staff.
2. Divide the form into two parallel half-forms. The two half-forms must be as parallel as possible in the following features:
   i) The content domains measured should be the same, or as similar as possible, for the two parts.
   ii) The number of items from each item format (i.e., multiple-choice, open-ended), number of score points, and test length must be equal.
   iii) The difficulties of the two half-forms must also match. This could be achieved by matching the IRT difficulty parameters and other item statistics.
3. State staff will be responsible for creating the two parallel forms. This will require the involvement of the content and psychometric staff. Project staff will provide advice upon request for building the forms.
4. The project researchers and the participating state will be responsible for the review and approval of the parallel forms.
5. The participating state will arrange production, printing, shipping/receiving, and scoring of the test with its test vendor.

B. Selecting the subject area, grade, and type of accommodations to study.

1. Each state will be testing only one grade and one subject (mathematics or reading).
2. Select the type of accommodation you want to study. Some examples of accommodations to study are: read aloud, online test readers, scribes, calculators, extended time, and variations in test setting (small group, individual). For this project, it is important to have a variety of different accommodations studied, as well as subject areas and grades, so plans will need to be discussed with project staff before the decision is final.
3. Make sure you have a large enough number of students for the type of accommodation you want to study. The study requires a minimum of 150–200 students who are eligible (i.e., required per their IEP or 504 plan) for using the accommodation.
4. Although some students receive multiple accommodations as a normal part of their assessment, a single accommodation could be more reasonable and more straightforward to study. If possible, examine your data on accommodation use during the assessment to determine which combinations are most frequently used, and whether some accommodations are used alone by sufficient numbers of students.
5. There could be situations where students use more than one accommodation out of necessity. As an example, a student who needs a human reader would also have extended time and special setting accommodations. Normally, these three accommodations are interdependent, the latter two being necessary conditions for the main accommodation, the human reader. In such situations, you could focus on the major accommodation and assume that the secondary accommodations are part of the major accommodation.

C. Sampling students.

1. As noted above, a minimum of 150–200 students is needed for each of the two groups: the students with disabilities (SWD) group and the students without disabilities (SWOD) group.
2. Examine the SWD group in your state that is eligible for the selected accommodation first. Check the composition of this population (gender, ethnicity, SES, type of community, etc.). If the count of this group is more than 700, you may need to sample from the group due to the cost of testing.
3. After you are satisfied with your selection of the SWD group, select a sample from the SWOD population. As much as possible, this sample should match the demographics of the SWD group. It may be simpler to select this sample from the same schools where you will be testing the SWD group. The procedures used to select the SWD and SWOD samples for the study need to be reviewed by project staff before they are finalized.

D. Test administration.

1. Before the test administration, the state will need to provide written test administration instructions that are unique to this study for test examiners.
2. We would like to be as flexible as possible in determining the best schedule for conducting the study. The project management will discuss your test administration schedule and the project timelines with you in light of the state's testing cycle.
3. It might be a good idea to administer the project test forms as a pilot or supplemental test administration during the normal test administration window.
4. Since students without disabilities have never used accommodations before, you should let them practice the accommodation in a session or two about two to four weeks prior to the test administration to get them more familiar with it. It is suggested that these students use the accommodation once prior to the test administration and then again when the directions for administration are given to the students on the day of the assessment.
5. In case you are planning to use an accommodation that you have not used before, you will need to give practice time for both groups. For example, if your state is using a computer-delivered screen reader accommodation for the first time, you will need to give students a chance to become familiar with the screen reader and the technology that comes with it. As suggested before, students should be allowed to use the accommodation at least a couple of weeks prior to the administration as well as when directions are given on the assessment day. The procedures to be used for administration and practice will need to be reviewed by project staff before being finalized.

E. Data.

1. Project staff will conduct all analyses. The main data to be analyzed for the study are the student scores from the test administration. After the student responses have been scored, we will need student-level data, including raw data, scoring keys, item scores, student ID, and other information, from your test contractor.
2. We also plan to cross-reference test performance with disability categories, needs, and services for the SWD group. For that to happen, we will need you to share with us your special education data. Each student record should contain:
    - Student ID
    - Standard demographic data, such as gender, race/ethnicity, and SES
    - Information about the student's disability, such as type of disability and services
    - If the student's disability is academic, whether it is in the area of reading, math, or both
    - Teacher's judgment as to whether student is showing grade-level classroom performance (or some other indicator of student performance)
    - School information, such as name of school, grade, class
    - Item response data

- Any other indicators you think would be useful in the study

These variables will be on the same student record as the student response variables. The project will provide you with a list of all specific data requirements and discuss with you the format and layout for the files.

3. To make it possible for us to help you create and check the development of parallel forms, we need to receive the test data that is used to build the half-forms. This includes IRT item parameters, classical item and test statistics, item objectives or domains, and other data on item characteristics you may have.
4. We will work with you and your contractor on the best way to collect and code the test data before the test administration.
5. In order to be as flexible as possible, we will accept data in almost any pre-specified format, but we will need to work out the easiest way to transfer the data.
6. We assure you that we will honor the confidentiality of your data sets and protect against any tampering or misuse.

Test Instruments

Each state was responsible for creating the test forms for use in their study. As was noted above, in order to reduce the overall testing time for students, two parallel split-half forms were developed from the full form. Each half-form was administered with and without accommodation, with each student receiving one half-form with accommodations and the other without. Note that the Michigan study followed a design that was slightly different from the rest of the studies, which will be discussed later in this section. For the other four studies, the half-forms were based on either an existing test or a released form, or in one state, entirely new parallel and equivalent forms were developed from an existing item bank. The choice of using current or old items/forms was left to the state. To reduce the burden on states, no new item development was required for participation in the project. Each half-form consisted of half the items from the full test, for both the multiple choice (MC) and constructed response (CR) items. These short versions of an existing test were made by splitting the full form into halves that were parallel in content and equivalent in difficulty, based on pre-existing item parameters (pre-equated). Working with their vendors, state content staff in the mathematics or reading subject areas, along with the state's psychometrician, helped create the study forms and/or conduct reviews to ensure that the half-forms were indeed parallel in content and statistically equivalent.

As mentioned earlier, the Michigan study followed a test construction process that was different from the other studies. A pilot form of a state language arts assessment used in 2008 was developed by revising items to make them more appropriate for SWD, which included the use of enhanced directions/scaffolding and one less option for students to choose from on multiple-choice questions. For the purposes of the study, the enhanced directions were viewed as an accommodation. It was designed for use as a future alternate assessment based on modified achievement standards (AA-MAS). A test parallel to the 2008 form was constructed from an old form by matching item

discrimination and reliability of the two forms. Then the newly constructed test was administered to a sample of SWD and SWOD students in February 2009. The comparison of student performance was based on the item scores for the sample (both SWD & SWOD) on the fall 2008 regular test and the students' item scores on the newly assembled form that was administered in February 2009. An important point to note is that every student in the study sample took both forms, but only in one administration condition. Consequently, none of the items in the two forms were administered in both conditions, as was the case in the other four studies.

The test forms and accommodations that were used for the five studies are summarized in Table 4.1.

Table 4.1: Test Form Descriptions

Study	Grade	Subject	Accommodation studied	# of MC items per subform	# of CR items per subform	Total points per subform
Connecticut # 1	4	Reading	Computer reader with digital voice	20	0	20
Connecticut # 2	7	Mathematics	Computer reader with human voice	26	0	26
Kentucky	4	Reading	Read aloud by a teacher	22	0	22
Michigan	6	Language Arts	Fewer answer choices and enhanced directions	21	1	25
Nevada	7	Mathematics	Read aloud by trained reader	20	0	20

Note: MC = multiple choice; CR = constructed response

Samples

Based on the accommodation selected for its study, each participating state identified the students who required that specific accommodation. The characteristics of the sample of students with disabilities depended upon the accommodations used. For example, in a state studying a read aloud accommodation, all students in the tested grade level whose IEP specified a read aloud accommodation were eligible for the study. The majority of these students might have a learning disability in reading, but other disabilities such as vision impairment could be represented in the eligible pool. Depending on the size of the eligible population, states included either a sample or the entire pool of eligible students with disabilities in the study. Generally, a minimum of 150–200 SWD was needed in each state's sample. In order to make administration easier,

states were advised to sample schools that could provide the largest number of SWD who needed or used the accommodation of interest. This might allow states to test geographically close and fewer schools instead of selecting many different areas around the state for testing, although the guidelines also emphasized the importance of a representative sample of schools in the state. Once the special education sample was identified, a similarly sized sample of general education students was selected to participate in the study, matched on their general demographics. General education students were randomly selected from the same schools or classrooms if appropriate.

Sample sizes for each of the five studies are provided in Table 4.2. As can be seen, state samples were adequate in all studies. The smallest number of examinees was in Kentucky, which had 150 SWD and 145 SWOD, very close to the minimum numbers recommended for the studies.

Table 4.2. Sample Sizes for the Studies

Study	Number of schools	N of SWD	N of SWOD	Total Count
Connecticut # 1	18	206	200	406
Connecticut # 2	27	282	366	648
Kentucky	22	150	145	295
Michigan	Not available	500	194	694
Nevada	12	212	225	437

Data Collection

After the half-forms were created, reviewed, and approved, they were printed by each state's vendor to be similar in appearance to the regular state assessment and then shipped to the schools. Each state determined the best time frame for the administration of its accommodation study. Study assessments were conducted in a separate, special administration and not in conjunction with the operational administration of the state program, since it was important that this research effort not adversely affect the operational administration of the state assessments used for accountability and reporting of Adequate Yearly Progress (AYP). In the Michigan study, the newly constructed form was administered to each student in the sample and item scores from the same students in the 2008 census administration were used as the parallel control form.

In the other four studies, each student was tested with both half-forms used in the study, one half-form at a time, receiving one under the standard administration condition and the other with the accommodation. In addition, the order of the administrations was counterbalanced such that each half-form was administered first, in both standard and accommodated conditions. This counterbalancing procedure was recommended to cancel any possible order effects.

After the administration, the state's contractor scanned and scored the student responses and then sent a data file to the CTEAG project analyst, who conducted the statistical analyses for each state. Prior to the transmittal of this data, the project analyst

worked with each state to agree on a format layout of the data files to be sent. The data elements in the studies included district code, school code, generic student ID, form taken first, condition tested first, gender, and item scores. Item metadata, such as the sub-domains or traits measured by each item, also was provided by the states.

Analysis Procedures

Since the purpose of the studies was to examine the validity of test scores from the accommodated administration, a primary focus of the analysis was to examine content and construct validity and the psychometric characteristics of the tests. The validity investigations addressed three important areas: 1) item and test characteristics, 2) test structure, and 3) effects on performance for SWD in comparison to SWOD. As described in detail elsewhere in this report, the research questions were as follows:

1. Does the accommodation affect item statistics?
2. Does the accommodation change the construct and structure of the test? For instance, is there an impact on the factor structure under an accommodated administration?
3. Does the accommodation increase performance of students? Does it increase performance more for SWD than it does for SWOD?

To answer these questions, a variety of investigations and statistical analyses were performed, as described in the following sections.

1. Impact of accommodation on item and test statistics.

Item-level analyses compared classical item statistics (difficulty & discrimination) between administration conditions. These indicators, p-value and point-biserial, are the most common ones used in comparing items in two testing situations (DePascale, 2009; Abedi, Courtney, & Leon, 2003). Basically, items which became notably easier or harder under the accommodated condition may have been impacted by the accommodation. That is, there was substantial enough of an accommodation effect that the accommodation could be construed as a modification (that it changed the construct being tested). Similarly, a marked change in the point-biserial due to an accommodation is not desirable as it might affect the quality of the item and its strength in ranking students according to their ability levels. In the studies done by Connecticut, Kentucky, and Nevada, classical item difficulty and point-biserial correlation statistics were compared for the same items presented with and without accommodation. In the Michigan study, where items were not tested in two conditions, investigation of accommodation effects on item properties was not possible.

In addition to item-level comparisons, investigations of the impact of accommodations on test properties using statistics such us reliability, average score, and

variability has been widely recommended (Laitusis, 2007). Barton and Winter (2009), for example, suggested that comparability of tests can be verified by examining properties such as the validity of inferences from the tests, whether the tests measure the same construct(s), and if the tests are equally reliable. In the CTEAG studies, all of these test properties were examined. To ensure content validity, for example, the half-forms were built to have the same numbers of items in each content domain. The numbers of score points in each sub-domain also were equal. A comparison of the testing conditions with respect to construct inferences is discussed in the next section of this chapter. With regard to reliability, the Cronbach's Alpha statistics for the two administration conditions were compared. In the event that the accommodation increased or decreased test reliability, it could be argued that consistency of the test in measuring student achievement had been compromised. In other words, a change in reliability might indicate an unfair situation where some test takers might have more measurement error in their scores than others.

2. Impact of accommodation on test dimensionality and construct.

The second set of analyses that was conducted examined whether the accommodation altered the dimensionality of the test or the construct it was purported to measure. The importance of verifying dimensionality has been amply addressed in the literature (Hambleton, Swaminathan, & Rogers, 1991), and it has been commonly agreed that educational tests should be unidimensional. The unidimensionality requirement follows from both practical and theoretical standpoints. First, a unidimensional test sharpens the focus of the test and simplifies the interpretation of the results, enabling test users to make accurate inferences from the examinee's score. Second, most of the currently used test models (such as Item Response Theory) require unidimensionality. The unidimensionality assumption is used in many models applied to analysis of test data and test users are often concerned with an overall composite of skills, such as overall mathematics ability when looking at a summative / end-of-instruction test such as those used for accountability under NCLB.

In the first part of the dimensionality investigation for the CTEAG studies, an exploratory factor analysis of the test data was done using SAS Software (SAS, Inc, 2009). As is well documented in the psychometric literature, factor analysis of binary items does not work properly (see, for example, Yuan, Bentler, & Kano, 1997; Kishton & Widaman, 1994), and researchers often opt to parcel items into meaningful clusters. In all five studies, the parceling approach was followed, combining items on the basis of the sub-domain indicated in the item metadata. The newly formed item parcels from each testing condition (accommodated and standard) were then analyzed using the factor analysis procedure in SAS. The main statistical indicators examined were the eigenvalues for the first few factors, the differences between successive eigenvalues, the factor loadings of the input variables, and the percent of variance in each variable accounted for by the first factor. These indicators were compared across testing conditions, where

equivalence or similarity was interpreted as an absence of the accommodation's effect on test dimensionality.

In the second part of the factor analyses, a confirmatory factor analysis was implemented using structural equation modeling. The goal of this analysis was to examine the invariance of the test construct, represented by the item parcels, between the accommodated and standard condition. The AMOS software (Arbuckle, 2007) was used for the confirmatory factor analysis, processing data from the two administration conditions. For each data set, the variables from the parceled items were modeled to measure a single construct or factor purported to be measured by the test. In the first analysis, data from each administration condition were processed, one at a time. The graphical output from AMOS as well as the fit indices and convergence reports were examined for each analysis. Among the multiple statistics examined, the first was the fit index (χ^2, P, df), which provided an indication of how well the model fit the data. In all the CTEAG studies, an adequate fit was obtained for the unidimensional model, which indicated that the data were indeed unidimensional and were structurally organized in an appropriate way during the parceling.

In the second analysis, invariance of the test construct between the accommodated and standard conditions was tested. This was accomplished by simultaneously processing the two data sets in AMOS. Note that it was not possible to do this analysis for every study. The software tests the goodness-of-fit when different estimates of the parameters are constrained to be equal under the two administration conditions (Arbuckle, 2007; Hu & Bentler, 1999). Among the parameters that were set equal for the two conditions were the factor loadings of the parceled variables by the common factor, their intercepts, means and variances of the common factors, and the measurement residuals. In AMOS, the models for these equalities are hierarchically nested in the sense that each model includes the constraints of the preceding model. For example, the model for the equal variances of the common factors includes the model that requires equality in the means, the intercepts, and the factor loadings. At the end of each analysis, the fit statistics for all models were examined. In such analyses, a good fit indicated that there was measurement invariance between the testing conditions, which in turn suggested that the accommodation had not altered the test construct.

3. Impact of accommodation on test scores.

The last analysis of the CTEAG data investigated whether the test accommodation affected examinee scores for the two groups, SWD and SWOD. The main goal of this analysis was to examine if the accommodations selected by the participating states were serving the purposes intended, which were to increase the accessibly of the test material for SWD. As a result, the premise of the studies was to evaluate whether the accommodations were increasing test scores for SWD more than they were increasing them for SWOD. This approach is in line with what has been recommended in the literature for such studies. It has been documented that the

differential boost method is a better approach in evaluating how accommodations affect students' scores (see Laitusis, 2007). Underlying this approach is the premise that an accommodation is effective to the extent it raises scores for SWD more than it does for SWOD.

In order to investigate this issue, a repeated measures analysis of variance (ANOVA) was used. This statistical procedure examines whether accommodations significantly increased test scores for one or both of the groups, and if they increased test scores more for SWD than for SWOD. As such, a 2 × 2 counterbalanced design described earlier in the chapter was used with the data from the two groups of students (SWD and SWOD) and the two administration conditions (with and without accommodations). To meet the requirements for conducting a repeated measures ANOVA, data were rearranged so that each student's score under the standard condition served as a pre-test and his/her score under the accommodated condition served as a posttest. Main and interaction effects were examined by comparing differences among the scores for the two disability groups and the two testing conditions. In addition, the significance of the differences in scores between the two main effects (disability status and test condition), the significance of the interaction between them, and the effect sizes of all statistically significant effects were examined and reported.

In Chapter V, detailed results from the analyses conducted for the five CTEAG validity studies are presented.

V. Results and Findings from State Validity Studies

As described in the previous chapter, a number of analyses were conducted on the data from each of the state studies in order to examine the effects of test accommodations on item and test characteristics, student performance, and score validity. In this chapter, detailed results are reported, by section, for the five validity studies:

5.1 Validity of Digital Voice Text Reader Accommodation in a Reading Test for Connecticut
5.2 Validity of Human Voice Text Reader Accommodation in a Mathematics Test for Connecticut
5.3 Validity of Read Aloud Accommodation in a Reading Test for Kentucky
5.4 Validity of Read Aloud Accommodation in a Mathematics Test for Nevada
5.5 Validity of Enhanced Directions and Item Revision Accommodation in an English Language Arts Test for Michigan

Each of these detailed state summaries provides the following information:

- Overview/Purpose of Study
- Sample
- Test Instruments
- Design
- Data Analysis
    - Test Score Summary
    - Item Statistics
- Examination for Construct Changes
    - Exploratory Factor Analysis
    - Confirmatory Factor Analysis
- Repeated Measures Analysis of Variance (ANOVA)
- Conclusion

The summary of each state's study that is included in this chapter presents data from the results of the analyses and describes the findings. An overall summary across all five studies and a discussion of the findings is presented in Chapter VI.

5.1 Validity of Digital Voice Text Reader Accommodation in a Reading Test for Connecticut

Purpose

This study was the first of two studies conducted by the state of Connecticut as part of the CTEAG project. The purpose of this study was to investigate the validity of scores from use of a digital voice text reader accommodation in a computer-delivered reading comprehension test. Specifically, the study sought answers for the following three questions:

1) Does the text reader accommodation affect item properties?
2) Does the text reader accommodation affect the tested construct of reading comprehension?
3) Does the text reader accommodation affect the performance of SWD differently than SWOD?

Sample

The study sample consisted of 406 students who were in 7th grade in the spring of 2008. These students attended 18 public schools that are representative of Connecticut public schools. About half of the students (206) were SWD, and the rest were SWOD. In the sampling process, the representative schools were selected first. Next, all SWD in the identified schools who were eligible for a text reader accommodation were selected for the study. Eligibility was determined by the planning and placement team (PPT)[1] using information about the need for the accommodation and achievement records of the student contained in the individual education plan (IEP). Practically, the PPT recommends a reader accommodation when the committee believes that a student would have difficulty in decoding grade-level text due to his or her disability. Finally, a sample of SWOD who matched the SWD group in background characteristics was selected from the same schools.

Test

The instrument was a grade 7 computer-delivered reading comprehension exam with 40 multiple-choice items. In order to reduce the total testing time, the test was divided into two half-forms, Form A and Form B, which were parallel in content and equivalent in difficulty. Each half-form consisted of two passages from an old test administered in spring of 2007. All items in Form A were as they were administered in the old form, but most of the Form B items were replacement items written to match the old items from the same passage. The reading test measured two strands or domains:

[1] Terminology used by Connecticut for an IEP Team.

forming an initial understanding, and developing interpretation. In each form, eight items measured understanding and the remaining 12 items measured interpretation.

Design

A 2 × 2 design was used in which disability status was crossed with testing condition. All students took both forms; one in standard condition and the other in accommodated condition. Note that in this study the standard condition was administered first to all students. The chart below represents the design used to administer the test forms.

	Standard	Accommodated
Group 1 (SWD & SWOD)	A	B
Group 2 (SWD & SWOD)	B	A

Data Analysis

Test Score Summary

Table 5.1.1 presents the score summaries for all students in the study. The N count represents the number of students with valid scores in the corresponding half-form. For example, 210 students took Form A in the standard condition, while 172 took Form B in the standard condition. Similarly, 173 students took Form A in the accommodated condition, while 190 took Form B in the accommodated condition. For both conditions and groups, Form B seems to be more difficult than Form A. Hence, the two half-forms were equated in order to compare the total scores accurately. The equating process is described later.

Table 5.1.1: Test Descriptive Statistics

Form	Condition	Disability Status	N	Mean	SD
A	Standard	SWOD	96	14.7	3.5
		SWD	114	9.4	4.2
		All	210	11.8	4.7
	Accommodated	SWOD	82	15.1	2.7
		SWD	91	10.0	3.8
		All	173	12.4	4.2
B	Standard	SWOD	80	13.0	3.2
		SWD	92	8.3	3.3
		All	172	10.5	4.0
	Accommodated	SWOD	89	12.9	3.0
		SWD	101	8.7	3.1
		All	190	10.7	3.7

Item Statistics

The first question that the study addressed was whether the text reader accommodation affected the items. To answer this question, classical item difficulties and point biserial correlations between accommodated and standard conditions were compared. Table 5.1.2 presents a summary of the item analysis for Form A and Form B. As can be seen in the table, the item statistics did not change much. The average change in difficulties was -0.019 for Form A, and -0.005 for Form B. Similarly, the average change in point biserial correlation was 0.023 for Form A, and 0.033 for Form B. These results indicate that the text reader accommodation did not change the performance of the items as compared to their standard values.

Table 5.1.2: Summary of Item Statistics

Note: The Difference is based on a summary of the Standard statistic minus the Accomm statistic at the item level.

Examination for Construct Changes

The second question explored by the study was whether the text reader accommodation affected the structure of the reading test. To address this question, an exploratory factor analysis was conducted, followed by a confirmatory factor analysis. When conducting a general factor analysis, binary scores such as those from multiple-choice items do not work properly. To circumvent this problem, test items are normally parceled together to form multi-category variables.

For this study, five multi-category variables were built from each 20-item half-form. The items were parceled according to the domains they measured – forming an initial understanding and developing interpretation. For each half-form, the eight items for understanding were clustered into two variables and the 12 items for interpretation were clustered into three variables. Scores of four non-adjacent items were added to create each variable. As a result, two variables for understanding and three variables for interpretation, each taking a value between 0 and 4, were formed. These five variables were submitted to factor analysis.

Exploratory Factor Analysis

The results of the exploratory factor analysis are shown in Table 5.1.3. The factor loadings and the percents of variance in each variable accounted for by the first (primary) factor are similar and have the same patterns across testing conditions. Hence, the results imply that the text reader accommodation did not alter the dimensionality of the test.

Table 5.1.3: Factor Loadings and Variance by First Factor

Form	Variable	Factor Loadings		% variance by factor	
		Standard	Accomm.	Standard	Accomm.
A	Understand 1	0.762	0.710	58%	50%
	Understand 2	0.773	0.724	60%	52%
	Interpret 1	0.793	0.770	63%	59%
	Interpret 2	0.756	0.703	57%	49%
	Interpret 3	0.820	0.819	67%	67%
B	Understand 1	0.749	0.754	56%	57%
	Understand 2	0.812	0.782	66%	61%
	Interpret 1	0.334	0.575	11%	33%
	Interpret 2	0.677	0.529	46%	28%
	Interpret 3	0.755	0.764	57%	58%

The first five eigenvalues shown in table 5.1.4 indicate that both forms were unidimensional whether they were administered in standard condition or in accommodated condition.

Table 5.1.4: First Five Eigenvalues and their Differences

Form A				Form B			
Eigenvalues		Differ from next one		Eigenvalues		Differ from next one	
Standard	Accomm.	Standard	Accomm.	Standard	Accomm.	Standard	Accomm.
3.050	2.784	2.465	2.126	2.361	2.375	1.391	1.445
0.585	0.658	0.068	0.058	0.970	0.930	0.303	0.272
0.517	0.600	0.073	0.033	0.667	0.658	0.136	0.115
0.444	0.567	0.041	0.173	0.531	0.543	0.060	0.049
0.403	0.394	-	-	0.471	0.494	-	-

Confirmatory Factor Analysis

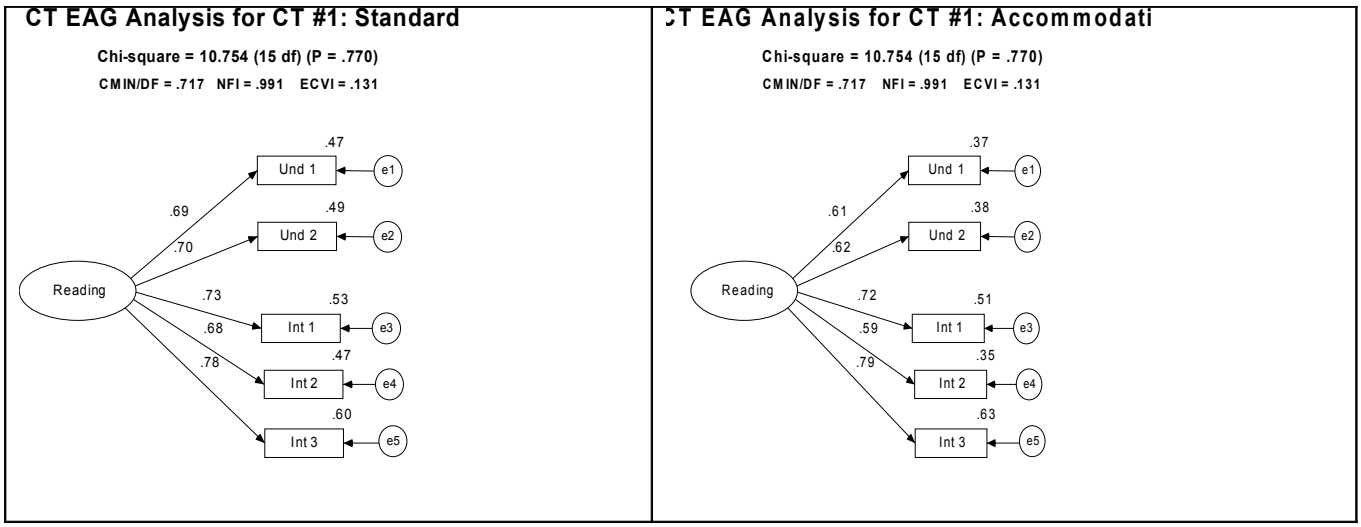
The goal in conducting the confirmatory factor analysis was to verify that the two administration conditions (standard and accommodation) had the same construct structure. To this end, a single factor was fit to the standard condition results, and then the same factor was fit to the accommodated condition results. Figure 5.1.1 shows the results from fitting the same factor model on the two testing conditions for Form A. The results for Form B were also similar.

The numbers on the arrows that go from the unobserved variable (reading factor) to the observed variables represent the factor loadings. From a practical point of view, the loadings are quite high, which indicates that the observed variables measure the reading factor very well. The numbers on the top of the observed variables represent the squared

multiple correlations. These indicate the percent of variance in each variable accounted for by the common factor. For both testing conditions, the squared multiple correlations were high. These indicators support the adequacy of the model fit.

Besides the chi-square (CMIN) and the probability statistics, there are three other fit indices presented in the charts. The first index, CMIN/DF, is the ratio of the Chi-square to degrees of freedom (DF). Values less than two indicate good model-data fit. Normed Fit Index (NFI) values higher than 0.9 indicate good model-data fit. Finally, the Expected Cross Validation Index (ECVI) represents the change in the minimization function that would be expected if the analysis were repeated with a sample similar to the current sample. The smaller the ECVI, the better the model fit.

Figure 5.1.1: Results of Confirmatory Factor Analysis



Furthermore, a multi-group analysis was used to test whether the model is invariant across the two data sets from the two testing conditions. That analysis tested the hypothesis that the model parameters for the two data sets are the same. Table 5.1.5 summarizes the results from testing the hypothesis that model parameters are equal for the two groups. Again, the data in the table were for Form A, but the results for Form B were similar. The models in the table are hierarchically nested, and each one adds more constraints to the preceding model. For example, the structural variance/covariance model implies that for the two testing conditions, the factor loadings are equal, measurement intercepts are equal, the factor means are equal, and the factor variances are equal. The last model (measurement residuals) adds another constraint to the above, requiring that the variances/covariances of the residuals are constant across the two testing conditions.

Table 5.1.5: Model Fit Statistics for Testing Equality of Parameters across Conditions

Equal Parameters for Models in Hierarchy	CMIN	DF	P	CMIN/DF	AIC
Factor loadings and intercepts	13.243	18	0.777	0.736	57.243
Above + factor means	15.198	19	0.710	0.800	57.198
Above + factor variances	17.196	20	0.640	0.860	57.196
Above + error variances (all parameters)	20.446	25	0.723	0.818	50.446

As shown in the above table, each of the models fits the data well, and none of them can be rejected. We also found that the most restrictive model, which constrains all parameters to be equal across groups, resulted in the best fit among the models (as demonstrated by the smallest AIC among all four models tested). All of the confirmatory factor analysis results indicate that the text reader accommodation did not change the underlying construct measured by the reading test.

Analysis of Variance

To examine whether the text reader accommodation affected reading scores for SWD differently than reading scores for SWOD, a repeated measures ANOVA with two independent variables (disability status [between subjects] and testing condition [within subjects]) and two dependent variables (total scores from standard and accommodated conditions) was conducted. To treat the testing condition as an effect, test scores in the standard and the accommodated conditions were treated as repeated measures – pretest and posttest, respectively. As mentioned earlier in the report, Form B was slightly more difficult than Form A. Therefore, it was important to equate the two forms so that proper comparisons could be made between the scores. For that, the total scores from the standard conditions of the two forms were equated using 20 multiple-choice items that were common for both groups. The two groups used for the equating were those students who took Form A in the standard condition and those who took Form B in the standard condition. The common items used for the equating were part of a live reading test administered in the spring of 2008.

The first analysis in the equating process was to examine whether the two equating groups were similar in reading ability estimated from the 20 common items. The average scores from the common items were 11.5 and 12.0, respectively, for group 1 (Form A) and group 2 (Form B). The small difference of 0.5 did not seem to matter in the equating process. Hence, a randomly equivalent groups linear equating of the two forms was conducted using data from the standard administration of the forms. The equated scores were used in the repeated measures ANOVA.

Table 5.1.6 shows the outcome of the Repeated Measures ANOVA. In part (a), it is shown that disability status significantly affects reading scores, testing condition does not significantly affect reading scores, and there were no interaction effects of disability status and testing condition on the reading scores. In part (b), the means are shown for the two main effects. These means were the input for Figure 5.1.2, which follows.

Table 5.1.6: Results for Repeated Measures ANOVA

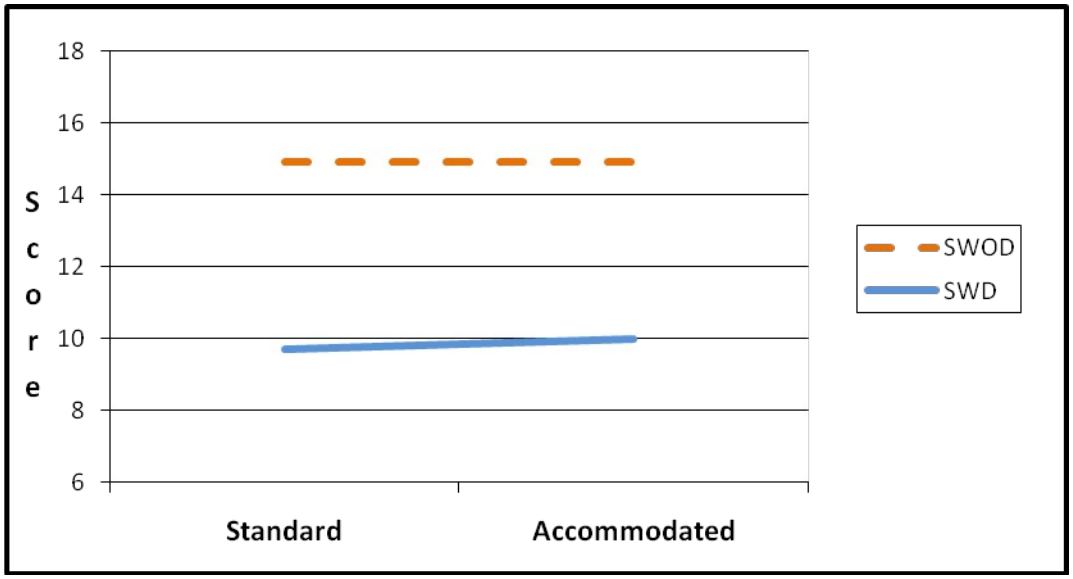
(a) Repeated Measures

Source	DF	SS	Mean Sq.	F value	Pr > F
Disability Status (between subjects)	1	4646.3	4646.3	222.77	<0.0001
Testing Condition (within subject)	1	2.1	2.1	0.39	0.5343
Disability × Condition (within subject)	1	4.0	4.0	0.74	0.3903

(b) Means

Disability status	Standard	Accommodated
SWD	9.7	10.0
SWOD	14.9	14.9

Figure 5.1.2: Plot of Means for Disability Status by Accommodation Condition (Connecticut Study #1)



Conclusion

This study was first in a series of collaborative state-based studies of the validity of test accommodations and score interpretations for SWD. Connecticut chose to study the text reader accommodation on a reading test for grade 7 students. The study investigated the text reader accommodation focusing on whether it changes item properties, whether it changes test construct, and whether it differently affects scores for SWD and scores for SWOD. For the first two concerns, it was found that the text reader accommodation did not affect item properties or test construct. For the third question, it was found that the text reader accommodation did not affect test scores differently for the two groups SWD and SWOD.

5.2 Validity of Human Voice Text Reader Accommodation in a Mathematics Test for Connecticut

Purpose

This study was the second of two studies conducted by the State of Connecticut as part of the CTEAG project. Its purpose was to examine the validity of inferences from mathematics scores obtained under accommodation condition. The accommodation selected for this study was a computerized text reader with a human voice on a mathematics test. Building on the outcome and experience gained from a previous study conducted by Connecticut, the text reader software and the testing process were revised and improved. The research questions addressed in the current study were similar to the first Connecticut study.

Sample

The sample for this study consisted of 648 students from 27 schools in the state. The students were in the 5th grade in the spring of 2009, and 282 of them were SWD while 366 were SWOD. The SWD were students identified as needing a reader accommodation on the mathematics test of the Connecticut Mastery Test. The SWOD sample was selected from the same schools as the SWD and was matched with SWD on background variables.

Test

The test was a grade 5 computer-delivered mathematics test built from a previously used form. It consisted of 52 multiple choice items that were split into parallel half-forms (26 items each). Each test measured four broad content standards: numbers and operations (12 items), geometry and measurement (6 items), data and statistics (5 items), and algebra (3 items).

Design

The test administration followed the 2 × 2 design developed for all CTEAG studies. Each student in the sample took a half-form under standard condition, and took the other half-form under accommodated condition. Furthermore, in order to control for possible order of administration effect, half of students in any school (both SWD & SWOD) were assigned to take the test under the standard conditions first while the other half was assigned to take the test under the accommodated condition first.

Data Analysis

Test Score Summary

Table 5.2.1 summarizes test score results for both groups (disability category) and testing condition (accommodated and standard). The N count represents the number of students with valid scores in the corresponding half-form. In both half-forms, there was a slight increase in the average scores for SWD from the without accommodation to the with accommodation condition.

The reverse was true for SWOD: a slight decrease in average score from the without accommodation to the with accommodation condition.

Table 5.2.1: Test Descriptive Statistics

Form	Condition	Disability Status	N	Mean	SD
A	Standard	SWOD	177	18.9	4.4
		SWD	148	13.4	4.8
		All	325	16.4	5.3
	Accommodated	SWOD	189	18.3	5.4
		SWD	134	13.6	5.4
		All	323	16.3	5.9
B	Standard	SWOD	189	19.9	4.5
		SWD	132	13.7	5.1
		All	321	17.3	5.7
	Accommodated	SWOD	177	19.2	5.2
		SWD	150	14.0	5.6
		All	327	16.8	5.9

Item Statistics

One of the issues examined in the study was whether the text reader accommodation affects item statistics. To address this question, items statistics were computed for both standard and accommodated conditions. Table 5.2.2 shows the means of the item difficulties and point biserial correlations for the two forms under both testing conditions. As shown, the point biserial correlations seem to have slightly increased with the accommodation, and in both forms, whereas the item difficulties and the reliabilities for the two conditions were close.

Table 5.2.2: Summary of Item Statistics

Form		Item Difficulty			Point Biserial Correlation		
	Stat	Standard	Accomm.	Difference	Standard	Accomm.	Difference
A	Mean	0.631	0.629	0.002	0.387	0.444	-0.056
	Min	0.212	0.168	-0.096	0.086	0.093	-0.235
	Max	0.898	0.870	0.096	0.578	0.623	0.061
	Reliability				0.847	0.880	
B	Mean	0.664	0.643	0.020	0.423	0.446	-0.022
	Min	0.374	0.311	-0.053	0.182	0.090	-0.175
	Max	0.972	0.915	0.086	0.573	0.589	0.126
	Reliability				0.869	0.880	

Note: The Difference is based on a summary of the Standard statistic minus the Accomm statistic at the item level.

Examination for Construct Change

Multi-level variables were created from the dichotomous item scores by adding scores from three items that measure the same content standard. In the process, two variables were created for the number and operations standard, two for geometry and measurement, one for working with data, and one for algebra and patterns. These six variables, which were each on a 0-3 scale, could be assumed a good representation of the structure of the 26-item test. To avoid over representation of the Numbers & Operations, and to obtain similarity in score ranges, six random items from this domain were not used in the parceling process. The six-variable data were submitted for factor analysis.

Exploratory Factor Analysis

The results of the exploratory factor analysis are provided in the tables below. Table 5.2.3 presents the factor loading and percent variance in each variable that is accounted for by a single factor. Both factor loadings and variances indicate that the test is unidimensional at both administration conditions. The magnitudes of the first eigenvalues compared to the remaining eigenvalues (see Table 5.2.4) also suggest that a single factor fits the test data.

Table 5.2.3: Factor Loadings and Variance by First Factor

		Factor Loadings		% variance by factor	
Form	Variable	Standard	Accomm.	Standard	Accomm.
A	Numb. & Operations	0.807	0.811	65%	66%
	Numb. & Operations	0.710	0.736	50%	54%
	Geom. & Measure.	0.816	0.777	67%	60%
	Geom. & Measure.	0.535	0.687	29%	47%
	Working with Data	0.790	0.805	62%	65%
	Algebra, Patterns	0.667	0.756	45%	57%
B	Numb. & Operations	0.668	0.759	45%	58%
	Numb. & Operations	0.700	0.721	49%	52%
	Geom. & Measure.	0.801	0.799	64%	64%
	Geom. & Measure.	0.730	0.709	53%	50%
	Working with Data	0.755	0.745	57%	55%
	Algebra, Patterns	0.687	0.715	47%	51%

Table 5.2.4: First Six Eigenvalues and their Differences

Form A		Form B	
Eigenvalues	Differ from next one	Eigenvalues	Differ from next one

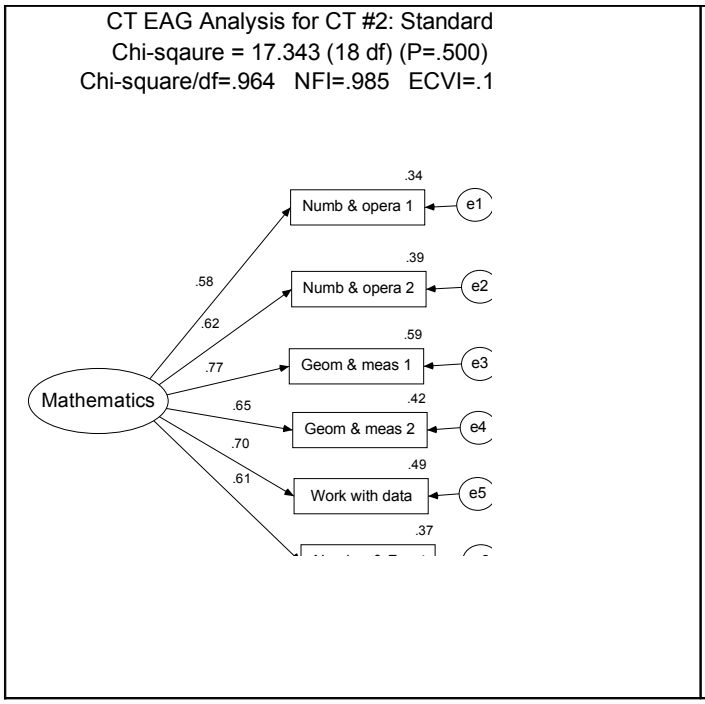
Standard	Accomm.	Standard	Accomm.	Standard	Accomm.	Standard	Accomm.
3.177	3.495	2.379	2.824	3.153	3.303	2.475	2.685
0.798	0.671	0.172	0.110	0.678	0.618	0.034	0.021
0.626	0.560	0.042	0.028	0.643	0.598	0.077	0.038
0.584	0.532	0.144	0.133	0.566	0.560	0.024	0.037
0.439	0.400	0.063	0.058	0.542	0.523	0.123	0.124
0.376	0.342	-	-	0.419	0.398	-	-

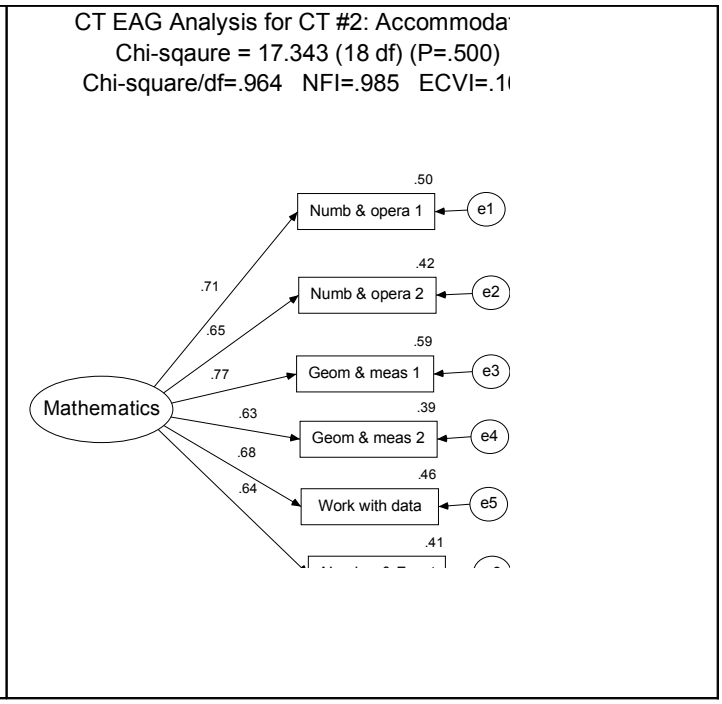
Confirmatory Factor Analysis

The goal for the confirmatory factor analysis was to express the test data (represented by the six variables) in a factor structure, and then examine whether the hypothesized structure holds. In other words, we tested whether the data has the factor structure it was purported to have. More importantly, we assessed whether the factor structure of the data changed with the text reader accommodation. For this analysis, structural equations modeling was used as implemented in AMOS software. Figure 5.2.1 represents the proposed model and how it fits both data sets: standard administration and accommodated administration. These results are for Form A (Form B results were similar).

The charts show that the structural model equally fits the two data sets (standard and accommodated) and that model-data fit was adequate. For instance, the factor loadings (represented by arrows from the mathematics factor to each of the six variables) were high in both testing conditions. Similarly, the squared-multiple correlations (on top of each observed variable) were moderate to high for both conditions. These indicate the amount of variation in each observed variable accounted by the common factor. Furthermore, the widely used chi-square indicates a good model fit. In addition, there are three more fit indices presented in the chart that support good model-data fit. These indices are the ratio of the chi-square to the degrees of freedom (less than 2.0 indicates good fit), Normed Fit Index or NFI (greater than 0.9 is recommended), and Expected Cross Validation Index (small value indicates good fit).

Figure 5.2.1: Results of Confirmatory Factor Analysis





A further analysis was conducted to support the structural equality of the data between the two testing conditions. Four models were fit that require equality or invariance of parameters across testing condition. The four models were hierarchically nested such that each one adds another constraint to the preceding model. Table 5.2.5 summarizes the results from testing the hypothesis that model parameters are equal for the two conditions for Form A. The simplest model (first) implies that the factor loadings are equal and the measurement intercepts (SMC) are equal. This model fits the data well. The last model (measurement residuals), which is the most constrained, requires all preceding models constraints, and adds that the variances/covariances of the residuals are constant across the two testing conditions. This model also fits the data well, and the models in between the first and last ones fit the data well. As shown in the table, the most restrictive model, which constrains all parameters to be equal across testing conditions, resulted in the best fit (smallest AIC) among the models.

Table 5.2.5: Model Fit Statistics for Testing Equality of Parameters across Conditions

Equal Parameters for Models in Hierarchy	CMIN	DF	P	CMIN/DF	AIC
Factor loadings and intercepts	31.798	28	0.283	1.136	83.798
Above + factor means	32.239	29	0.309	1.112	82.239
Above + factor variances	32.313	30	0.353	1.077	80.313
Above + error variances (all parameters)	36.025	36	0.467	1.001	72.025

Analysis of Variance

The last question that was examined in this study was whether the text reader accommodation affects math scores for SWD more than it does for SWOD. To investigate this issue, the data were modeled as a repeated measures problem. To that end, it was presumed that the standard condition scores would stand for the first measure and the accommodated condition scores would stand for the second measure. As a result, each student would have two scores; a pretest (standard) and a posttest (accommodated). Table 5.2.6 highlights the results of the repeated measures ANOVA. Part (a) presents the results for the Repeated Measures, and indicates that disability has a significant effect on scores (between subjects), that the text reader accommodation did not affect test performance as a whole, but that there was significant interaction effect between disability and testing condition. The next part (b) shows the means for the two main effects: disability status and test administration condition. These means are also shown in Figure 5.2.2.

Table 5.2.6: Results of Repeated Measures ANOVA

(a) Repeated Measures

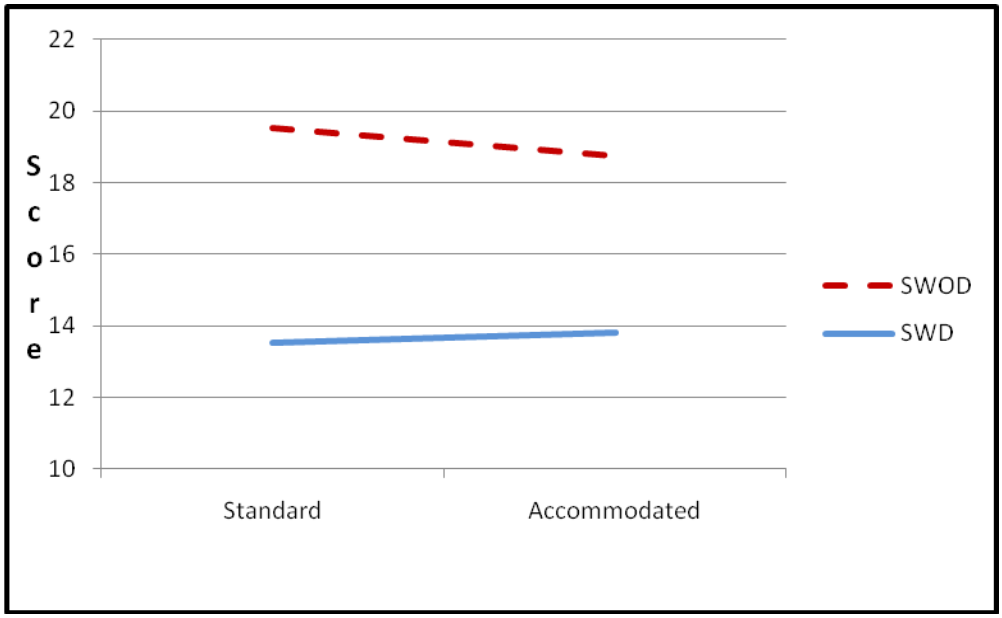
Source	DF	SS	Mean Sq.	F value	Pr > F
Disability Status (between subjects)	1	9388	9388	222.94	<0.0001
Testing Condition (within subject)	1	19	19	2.47	0.1162
Disability × Condition (within subject)	1	89	89	11.40	0.0008

(b) Means

Disability Status	Standard	Accommodated
SWD	13.52	13.80
SWOD	19.53	18.75

The chart below highlights the interaction between disability status and testing condition, which was found to be significant in the repeated measures ANOVA. Comparing the least squares means for the two groups, it was found that, due to the accommodation, average scores for SWOD slightly decreased while average scores for SWD slightly increased.

Figure 5.2.2: Plot of Means for Disability Status by Accommodation Condition (Connecticut Study #2)



Conclusion

The second study in Connecticut examined the effects of a text reader accommodation on mathematics test scores for grade 5 students. The results were encouraging regarding the use of the text reader accommodation for SWD. It was found that the accommodation did not change item difficulty, item point biserials, or test reliability. It also was found that the text reader accommodation does not change the dimensionality of the test or the test construct. A one-factor model fitted both standard and accommodated data sets. Finally, while there was no main effect for the text reader accommodation, there was a significant interaction effect between disability status and testing condition. Scores increased slightly for SWD with the accommodation, while scores decreased slightly for SWOD with the accommodation (the effect size for the interaction [0.01] was small).

5.3 Validity of Read Aloud Accommodation in a Reading Test for Kentucky

Purpose

The purpose of the study done in Kentucky was to examine the validity of test scores from a teacher read aloud accommodation used with a reading test for fourth-grade students. The research questions addressed in this study were similar to those for the other studies.

Sample

The study was part of collaborative effort by several states participating in the CTEAG project to examine the validity of the interpretations of test results from accommodated tests. The sample in the Kentucky study consisted of 295 students who were in 4th grade in the spring of 2009. About half of the students (150) were SWD, and the rest were SWOD. In the sampling process, 22 schools in 18 school districts were selected first. Next, all SWD in the identified schools who were eligible for a read aloud accommodation during the spring testing were selected for the study. Finally, a sample of SWOD from the same schools who matched the SWD group in background characteristics was selected.

Test

The test was based on a grade 4 state reading assessment consisting of 44 multiple-choice items that were selected from an item bank. In order to reduce the total testing time, two half-forms, Form A and Form B, parallel in content and equivalent in difficulty, were created. Each of the half-forms consisted of three passages that were accompanied by either 6 or 10 items. The reading test used in the study measured two strands: literary and informational. The 10-item passage measured the informational strand and the two 6-item passages measured the literary strand.

Design

A 2 × 2 design was used with disability status crossed with testing condition. All students took both half-forms, one in standard condition and the other in accommodated condition. To control for a possible order effect, half of the students in each disability category took one half-form first and the other half of the students took the other half-form first. The chart below represents the design used to administer the test forms.

	First Session	Second Session
Group 1 (SWD & SWOD)	A - Standard	B - Accommodated
Group 2 (SWD & SWOD)	B - Accommodated	A - Standard
Group 3 (SWD & SWOD)	B - Standard	A - Accommodated
Group 4 (SWD & SWOD)	A - Accommodated	B - Standard

Data Analysis

Test Score Summary

Table 5.3.1 presents the score summaries for all students in the study. The N count represents the number of students with valid scores in the corresponding half-form. The two half-forms seem to be equally difficult under both conditions and both groups. The only exception is that form A seems a bit harder than form B for SWD under standard administration. In general, the read aloud accommodation increased student performance for all students.

Table 5.3.1: Test Descriptive Statistics

Form	Condition	Disability Status	N	Mean	SD
A	Standard	SWOD	67	16.8	4.4
		SWD	80	10.5	5.2
		All	147	13.4	5.8
	Accommodated	SWOD	78	18.8	2.6
		SWD	70	14.5	4.7
		All	148	16.7	4.3
B	Standard	SWOD	78	17.1	3.6
		SWD	70	12.2	5.1
		All	148	14.3	5.3
	Accommodated	SWOD	67	18.7	2.4
		SWD	80	14.7	4.2
		All	147	16.5	4.1

Item Statistics

For both half-forms, all items became easier under the accommodation conditions. The point biserial correlations decreased under the accommodation condition, with some items' point biserials decreasing dramatically. For example, the point biserial correlation of an item in form B changed from 0.489 under standard administration to 0.037 under accommodated administration. Also, as a result of the decrease in score variability due to accommodation (see table 5.3.1), the reliabilities of the forms also decreased slightly under accommodation conditions. For example, the reliability for form A changed from 0.891 under standard administration to 0.833 under accommodated administration. For form B, the reliability changed from 0.866 under standard administration to 0.803 under accommodated administration. Table 5.3.2 summarizes the classical item statistics for the two forms and administration conditions.

Table 5.3.2: Summary of Item Statistics

	Stat	Difficulty	Point Biserial

Form		Standard	Accomm.	Difference	Standard	Accomm.	Difference
A	Mean	0.609	0.760	-0.152	0.490	0.396	0.094
	Min	0.401	0.493	-0.280	0.304	0.279	-0.112
	Max	0.830	0.932	-0.024	0.592	0.531	0.288
		Reliability			0.891	0.833	
B	Mean	0.651	0.751	-0.100	0.448	0.366	0.082
	Min	0.372	0.476	-0.310	0.111	0.037	-0.087
	Max	0.824	0.905	-0.005	0.580	0.581	0.452
		Reliability			0.866	0.803	

Note: The Difference is based on a summary of the Standard statistic minus the Accomm statistic at the item level.

Examination for Construct Changes

The next analysis addressed the extent to which the read aloud accommodation may have impacted the structure of the reading test. To this end, an exploratory factor analysis was conducted followed by a confirmatory factor analysis. Since linear factor models do not work well with binary item scores, test item scores were clustered together to form multi-level variables. For the Kentucky study, seven continuous variables were built from each 22-item form. The items were parceled together according to the content domains they measured – literary and informational. For each form, the 12 items for literary were combined into four variables and the 10 items for informational were combined into three variables. In order to make the range of new variables the same (0-3), a random item in the informational domain was not used in the parceling. Scores of three non-adjacent items were added to create each variable. As a result, four variables for literary and three variables for informational were formed. These seven variables were the input for the factor analysis, both exploratory and confirmatory.

Exploratory Factor Analysis

The results of the exploratory factor analysis are shown in Table 5.3.3. The factor loadings and the percents of variance in each variable accounted for by the single factor are presented for both testing conditions. The indices were similar for the two testing conditions, but, with few exceptions, mostly a bit higher in the standard administration condition. This could mean that in general, the single factor solution fits better in the standard administration data than in the accommodated administration data. In both conditions, however, the results indicate the test is unidimensional. Hence, the results imply that the read aloud accommodation did not notably alter the dimensionality of the test.

Table 5.3.3: Factor Loadings and Variance by First Factor

Form	Variable	Factor Loadings		% variance by factor	
		Standard	Accomm.	Standard	Accomm.
	Literary 1	0.829	0.735	69%	54%
	Literary 2	0.733	0.663	54%	44%

A	Literary 3	0.790	0.606	62%	37%
	Literary 4	0.723	0.623	52%	39%
	Information 1	0.711	0.760	51%	58%
	Information 2	0.766	0.758	59%	57%
	Information 3	0.768	0.724	59%	52%
B	Literary 1	0.763	0.454	58%	21%
	Literary 2	0.703	0.728	49%	53%
	Literary 3	0.768	0.722	59%	52%
	Literary 4	0.781	0.785	61%	62%
	Information 1	0.818	0.786	67%	62%
	Information 2	0.706	0.673	50%	45%
	Information 3	0.546	0.607	30%	37%

In Table 5.3.4, the first seven eigenvalues are shown. These data indicate that both forms are unidimensional, whether they were administered in standard condition or in accommodated condition.

Table 5.3.4: First Seven Eigenvalues and their Differences

Form A				Form B			
Eigenvalues		Differ from next one		Eigenvalues		Differ from next one	
Standard	Accomm.	Standard	Accomm.	Standard	Accomm.	Standard	Accomm.
4.052	3.412	3.421	2.627	3.739	3.311	2.936	2.370
0.632	0.785	0.050	0.094	0.803	0.942	0.175	0.208
0.582	0.691	0.024	0.079	0.628	0.734	0.026	0.113
0.558	0.613	0.061	0.005	0.603	0.621	0.107	0.052
0.498	0.608	0.090	0.090	0.495	0.569	0.097	0.125
0.407	0.518	0.136	0.144	0.399	0.445	0.066	0.067
0.271	0.373	-	-	0.333	0.378	-	-

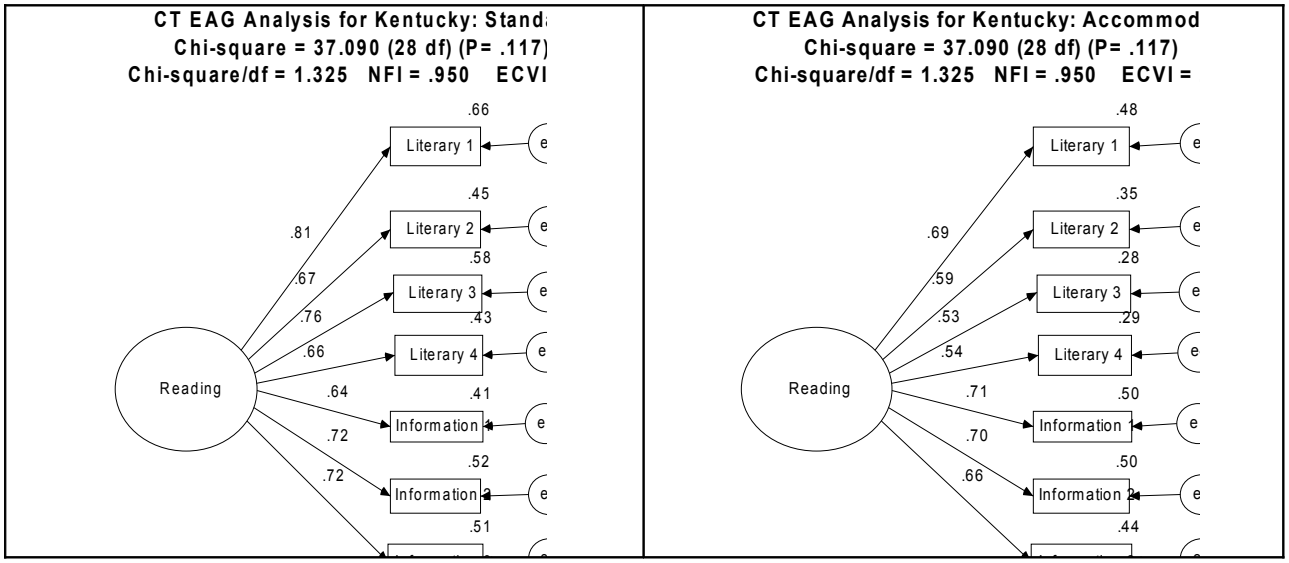
Confirmatory Factor Analysis

The goal in conducting the confirmatory factor analysis was to examine if the two administration conditions (standard and accommodation) have the same construct structure. To this end, a single factor was fit to the standard condition results, and then the same factor was fit to the accommodated condition results. The data for this analysis comprised the seven variables mentioned above, and all variables were hypothesized to measure a single reading factor. In Figure 5.3.1 the results from fitting the same factor model on the two testing conditions for Form A are shown. The results for Form B were similar.

The numbers on the arrows that go from the unobserved variable (reading factor) to the observed variables represent the factor loadings. From a practical point of view, the loadings are high, an indication that the observed variables measure the reading factor very well. The numbers on the top of the observed variables represent the squared-multiple correlations. These represent the percent of variance in each variable accounted for by the common factor. For both testing conditions, the squared-multiple correlations were high, another support for the adequacy of the model fit.

Besides the chi-square and the probability statistics, three other fit indices are presented in the charts. The first index is the ratio of the Chi-square and DF. It is recommended to be less than 2.0 to support model-data fit. The Normed Fit Index (NFI) is recommended to be higher than 0.9 to indicate good model-data fit. Finally, the Expected Cross Validation Index (ECVI) represents the change in the minimization process that would be expected if the analysis were repeated with a sample similar to the current sample. The smaller the ECVI, the better the model fit. In general, these three fit indices also show a good model fit.

Figure 5.3.1: Results of Confirmatory Factor Analysis



Analysis of Variance

As indicated in Table 5.3.1, SWD gained more in scores from the read aloud accommodation than SWOD. To test the significance of the gain, an ANOVA with repeated measures that had two independent variables (disability status and testing condition) and two dependent variables (total scores in standard and accommodated conditions) was conducted. To treat the testing condition as an effect, test scores in the standard and the accommodated conditions were treated as repeated measures; i.e. as pretest and posttest, respectively.

Table 5.3.5 presents the outcome of the ANOVA. In part (a), it is shown that disability status significantly affected reading scores, that the testing condition significantly affected reading scores, and that there was a significant interaction effect between disability status and testing condition on reading scores. In part (b), the means are shown for the two main effects of disability status and testing condition. Figure 5.3.2 also shows the means with respect to the disability status and testing condition.

Table 5.3.5: Results for Repeated Measures ANOVA

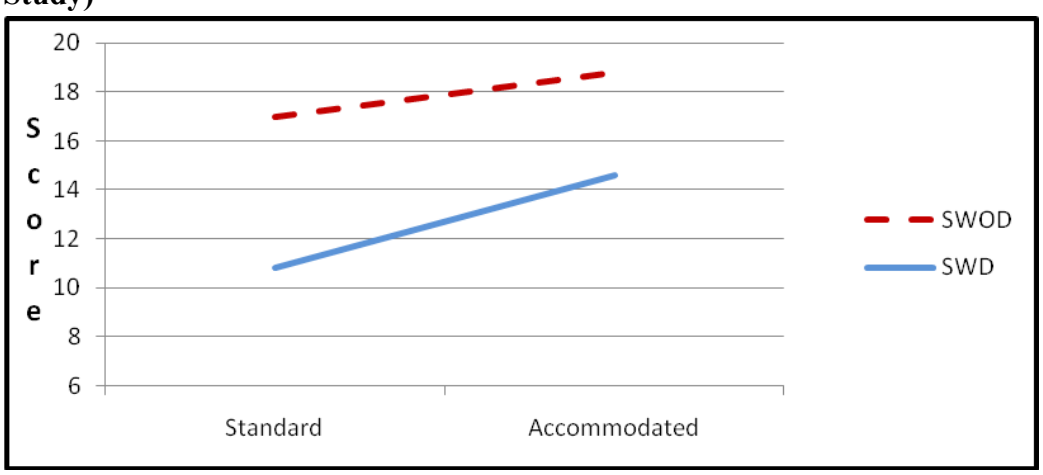
(a) Repeated Measures

Source	DF	SS	Mean Sq.	F value	Pr > F
Disability Status (between subjects)	1	3939.8	3939.8	155.62	<0.0001
Testing Condition (within subject)	1	1114.7	1114.7	124.82	<0.0001
Disability × Condition (within subject)	1	142.8	142.8	15.98	<0.0001

(b) Means

Disability Status	Standard	Accommodated
SWD	10.8	14.6
SWOD	17.0	18.8

Figure 5.3.2. Plot of Means for Disability Status by Accommodation Condition (Kentucky Study)



Conclusion

As a participating member of the consortium of states in the CTEAG project, Kentucky chose to study the effect of a read aloud accommodation on test scores for a reading test in grade 4. It was found that, at the item level, the read aloud accommodation made items a bit easier and the test reliability lower than in the standard administration. It was also found that the test structure was not altered by the read aloud accommodation. A one-factor model adequately fit both standard and accommodated data sets. Finally, the read aloud accommodation increased test scores for SWD more than for SWOD. The difference in score gain due to the read aloud accommodation between the two groups was significant. This was indicated by the significant interaction effect between accommodation and disability, as is visually displayed in Figure 5.3.2.

5.4 Validity of Read Aloud Accommodation in a Mathematics Test for Nevada

Purpose

The study was part of collaborative effort by several states participating in the CTEAG project to examine the validity of the interpretations of test results from accommodated tests. The Nevada study explored the impact of a read aloud accommodation used with a mathematics test, and the validity of the interpretations of its results. The research questions addressed in this study were similar to those for the other studies.

Sample

The sample comprised 437 students who were in the 7th grade in the 2008-2009 school year. About half of the sampled students (212) were SWD and the other half were SWOD. In the sampling process, schools in Nevada attended by at least several SWD were selected first. Next, all SWD in the identified schools who were eligible for a read aloud accommodation in mathematics during the spring testing were selected for the study. Finally, a sample of SWOD from the same schools who matched the SWD group in background characteristics was selected.

Test

As mentioned above, the test was based on the state mathematics assessment for grade 7. It consisted of 40 multiple-choice items, and was specifically built for the study, using Nevada's test blueprint and test development procedures for operational forms of the tests. The items were selected from a pool of released items, selecting items with p-values between 0.3 and 0.6. The 40 items were split in half to form two parallel half-forms, each consisting of 20 items. In addition to equality in average difficulty, the two half-forms were built to be similar in the content domains and mathematics skills they measure.

Design

The test and its administration were designed to follow the 2 × 2 plan that was used for the CT EAG studies. About half of the students, both SWD and SWOD, took items 1-20 (Form A) under the standard condition, then items 21-40 (Form B) under the accommodation condition. The other half took items 1-20 under the accommodation condition, then items 21-40 under the standard condition. The resulting counterbalanced design had disability status crossed with administration condition.

Data Analysis

Test Score Summary

A summary of the test score statistics is presented in Table 5.4.1. The N count represents the number of students with valid scores in the corresponding form. Differences in scores between standard and accommodated administration were very small, but consistent. In all score

changes, SWD seem to have benefited slightly more from the accommodation than SWOD. Furthermore, Form B (items 21-40) was a bit harder than Form A (items 1-20).

Table 5.4.1: Test Descriptive Statistics

Test Form	Condition	Disability Status	N	Mean	SD
A Items 1-20	Standard	SWOD	106	13.0	4.3
		SWD	106	8.7	3.5
		All	212	10.9	4.4
	Accommodation	SWOD	109	13.0	3.9
		SWD	106	9.0	3.6
		All	215	11.0	4.2
B Items 21-40	Standard	SWOD	108	12.3	3.9
		SWD	105	7.8	3.6
		All	213	10.1	4.4
	Accommodation	SWOD	106	12.6	4.0
		SWD	106	8.3	3.4
		All	212	10.5	4.3

Item Statistics

A summary of the classical item statistics, i.e. difficulty and point biserial correlation, is shown in Table 5.4.2. As shown, average difficulty did not change much when comparing the standard administration condition to the accommodated condition for Form A, but there was a slight change in average difficulty in Form B. Also, Form B seems to be slightly more difficult than Form A. The point biserial correlations were similar across forms, and stayed the same between the two administration conditions. As a result, the reliability indices were equivalent across forms and administration conditions.

Table 5.4.2: Summary of Item Statistics

Form		Item Difficulty			Point Biserial Correlation		
	Stat	Standard	Accomm.	Difference	Standard	Accomm.	Difference
A	Mean	0.551	0.559	-0.008	0.349	0.374	-0.025
	Min	0.373	0.362	-0.018	0.200	0.206	-0.140
	Max	0.824	0.826	0.086	0.478	0.560	0.076
	Reliability				0.810	0.790	
B	Mean	0.508	0.526	-0.018	0.355	0.350	0.004
	Min	0.270	0.212	-0.098	0.148	0.180	-0.210
	Max	0.670	0.704	0.128	0.483	0.483	0.185
	Reliability				0.790	0.790	

Note: The Difference is based on a summary of the Standard statistic minus the Accomm statistic at the item level.

Examination for Construct Change

In order to adequately process the data in factor analysis using a standard statistical package, the items were clustered into the skills they measure, which were identified as conceptual, procedural, and problem solving. The item scores were summed according to the skills they tap. For both forms, the result from the parceling was as follows: Two variables for conceptual, two variables for procedural, and two variables for problem solving. Each variable consisted of three item scores (0-3 scale), and two random items were not used in the parceling so that the scale of the resulting variables are the same. The six variables were then submitted for factor analysis.

Results of Exploratory Factor Analysis

The results of the exploratory factor analysis are in the tables 5.4.3 and 5.4.4. Table 5.4.3 presents the factor loadings and percent of variance in each variable that is accounted for by the single factor. The statistics for both the factor loadings and percent variances indicate that the test is unidimensional under both administration conditions. The magnitudes of the first eigenvalues compared to the remaining eigenvalues (see Table 5.4.4) also suggest that a single factor fits the test data.

Table 5.4.3: Factor Loadings and Variance by First Factor

Form	Variable	Factor Loadings		% variance by factor	
		Standard	Accomm.	Standard	Accomm.
A	Conceptual 1	0.657	0.622	43%	39%
	Conceptual 2	0.686	0.726	47%	53%
	Procedural 1	0.676	0.671	46%	45%
	Procedural 2	0.761	0.795	58%	63%
	Problem solving 1	0.740	0.702	55%	49%
	Problem solving 2	0.650	0.601	42%	36%
B	Conceptual 1	0.754	0.675	57%	46%
	Conceptual 2	0.600	0.604	36%	37%
	Procedural 1	0.691	0.680	48%	46%
	Procedural 2	0.721	0.782	52%	61%
	Problem solving 1	0.723	0.683	52%	47%
	Problem solving 2	0.668	0.630	45%	40%

Table 5.4.4: First Six Eigenvalues and their Differences

Form A				Form B			
Eigenvalues		Differ from next one		Eigenvalues		Differ from next one	
Standard	Accomm.	Standard	Accomm.	Standard	Accomm.	Standard	Accomm.

2.908	2.851	2.120	1.964	2.896	2.757	2.115	1.952
0.788	0.886	0.091	0.172	0.781	0.805	0.136	0.059
0.697	0.714	0.076	0.072	0.645	0.746	0.028	0.105
0.622	0.641	0.048	0.104	0.616	0.641	0.083	0.046
0.573	0.537	0.161	0.167	0.533	0.595	0.004	0.140
0.412	0.371	-	-	0.529	0.455	-	-

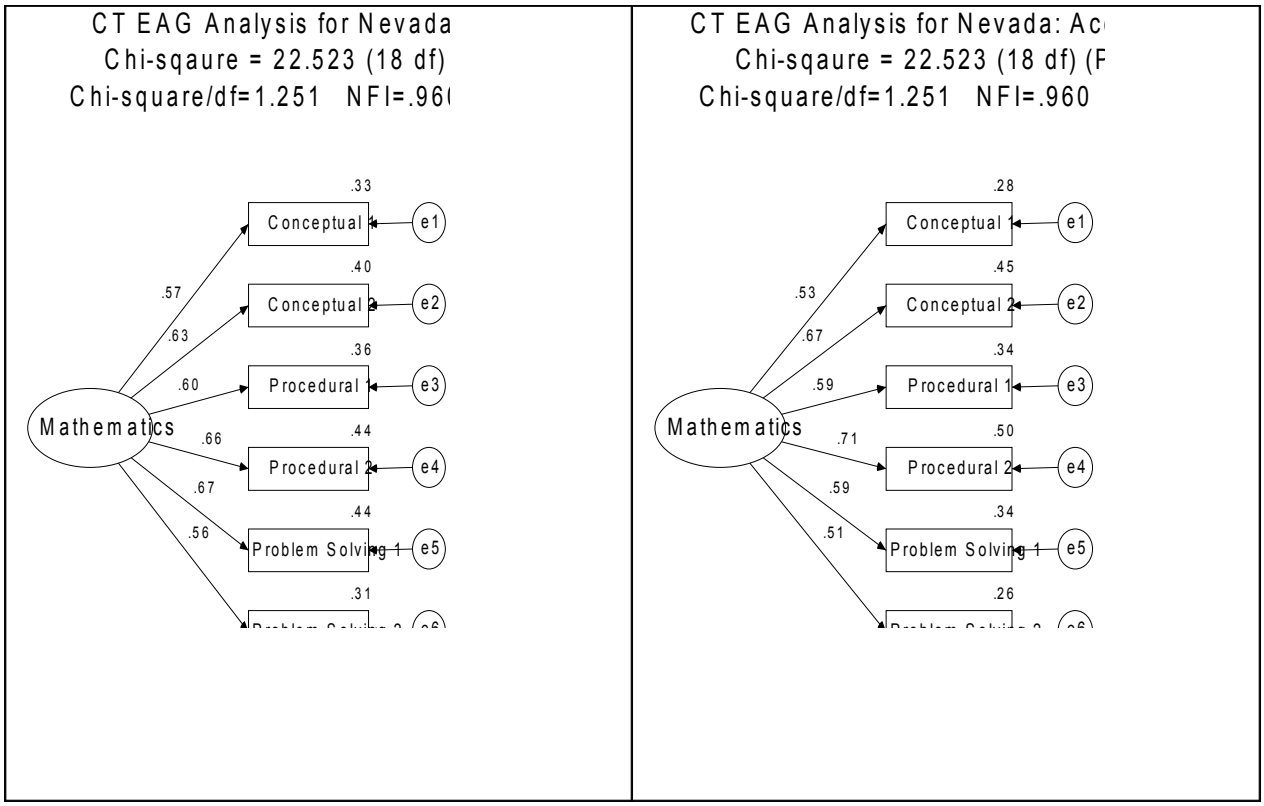
Confirmatory Factor Analysis

The goal in conducting the confirmatory factor analysis was to examine if the two administration conditions (standard and accommodation) have the same factor structure. For this purpose, a single factor was fit to the six variables from the standard condition results. The hypothesis was that the variables measured a single mathematics construct. The same factor was then fit to the six variables from the accommodated condition results. Figure 5.4.1 shows the results from fitting the same factor model on the two testing conditions for Form A. The results for Form B were also similar.

The numbers on the arrows that go from the unobserved variable (math factor) to the observed variables represent the factor loadings. For practical purposes, the loadings were moderate to high, an indication that the observed variables measure the math factor adequately. The numbers on the top of the observed variables represent the squared-multiple correlations. These represent the percent of variance in each variable accounted for by the common factor. For both testing conditions, the squared-multiple correlations were moderate; another support for the adequacy of the model fit.

Besides the chi-square and the probability statistics, three other fit indices are presented in the charts. The first index is the ratio of the Chi-square and DF. It is recommended to be less than 2.0 to support model-data fit. The Normed Fit Index (NFI) is recommended to be higher than 0.9 to indicate good model-data fit. Finally, the Expected Cross Validation Index (ECVI) represents the change in minimization process that would be expected if the analysis were repeated with a sample similar to the current sample. The smaller the ECVI, the better the model fit. All indices point to a good model-data fit.

Figure 5.4.1: Results of Confirmatory Factor Analysis



To further support the structural equality of the two data sets from the two testing conditions, goodness of fit was compared for different models of the data. Four models that require equality or invariance of parameters across testing condition were fit. The four models were hierarchically nested such that each one adds another constraint to the preceding model. Table 5.4.5 summarizes the results from testing the hypothesis that model parameters are equal for the two conditions for Form A. The simplest model (first) implies that the factor loadings are equal and the measurement intercepts (SMC) are equal. This model fits the data well. The last model (measurement residuals), which is the most constrained, requires all preceding models constraints, and adds that the variances/covariances of the residuals are constant across the two testing conditions. This model also fits the data well, and the models in between the first and last fit the data well as well. As shown in the table, the most restrictive model, which constrains all parameters to be equal across testing conditions, resulted in the best fit (smallest AIC) among the models.

Table 5.4.5: Model Fit Statistics for Testing Equality of Parameters across Conditions

Equal Parameters for Models in Hierarchy	CMIN	DF	P	CMIN/DF	AIC
Factor loadings and intercepts	29.886	29	0.369	1.067	81.886
Above + factor means	29.947	29	0.417	1.033	79.947
Above + factor variances	30.110	30	0.460	1.004	78.110
Above + error variances (all parameters)	33.185	36	0.603	0.922	69.185

Analysis of Variance

In this analysis, data were processed such that the standard condition would stand for the first measure, and the accommodated test would stand for the second measure. The tables below highlight the results of the repeated measures ANOVA. Part (a) presents the results for the Repeated Measures. The table indicates that disability had a significant effect on scores (between subjects), that testing condition did not have a statistically significant effect on test performance (p=0.0523), and that the interaction between disability and testing condition was not significant. The next part (b) shows the means under the conditions, which were the input for the following chart.

Table 5.4.6: Results for Repeated Measures ANOVA

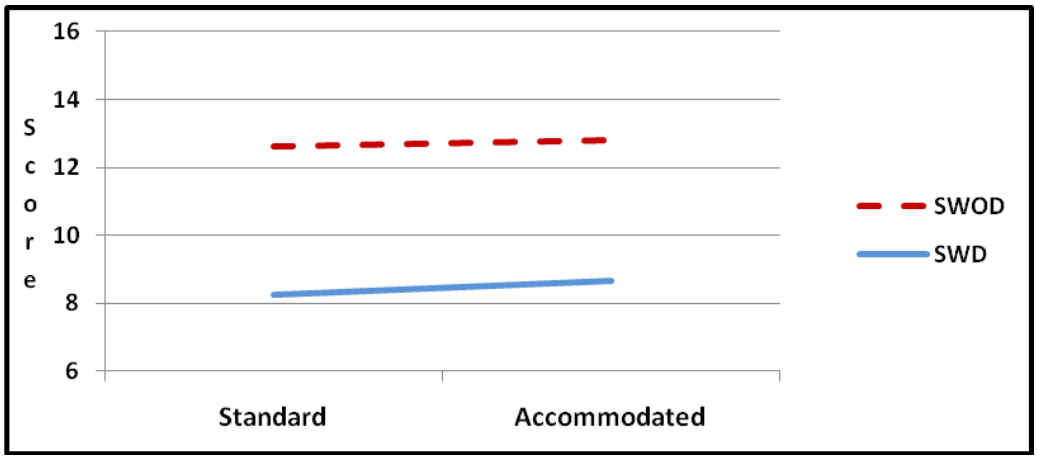
(a) Repeated Measures

Source	DF	SS	Mean Sq.	F value	Pr > F
Disability Status (between subjects)	1	3830	3830	158.80	<0.0001
Testing Condition (within subject)	1	18	18	3.79	0.0523
Disability × Condition (within subject)	1	2.4	2.4	0.49	0.4827

(b) Means

	Standard	Accommodated
SWD	8.27	8.67
SWOD	12.62	12.81

Figure 5.4.2: Plot of Means for Disability Status by Accommodation Condition (Nevada Study)



Conclusion

The Nevada study examined the effect of a read aloud accommodation on mathematics test scores for grade seven students. At the item level, the read aloud accommodation did not change item difficulty, item point biserials, or test reliability. Furthermore, the read aloud accommodation did not change the test dimensionality and the two administration conditions had similar factor structures, with a one factor model fitting both standard and accommodated data sets. Finally, the read aloud accommodation did not have a significant effect overall or a differential effect on the two disability groups.

5.5 Validity of Enhanced Directions and Item Revision Accommodation in an English Language Arts Test for Michigan

Purpose

The study was part of collaborative effort by several states participating in the CTEAG project to examine the validity of the interpretations of test results from accommodated tests. The main purpose of the Michigan study was to investigate the effectiveness of enhanced directions as an accommodation for SWD. This entails assessing whether the accommodation increases test scores for SWD more than SWOD. Another purpose was to investigate the extent to which the studied accommodation preserves the construct and the factor structure of the test.

Sample

The participants in the study consisted of 500 SWD and 194 SWOD. Schools that represented the state of Michigan in demographic composition, location, and socioeconomic status were chosen to participate in this study and all SWD in those schools were included in the study. Finally, groups of SWOD who matched SWD in important demographic indicators within the same schools were selected and included in the study.

Test

Two forms of a grade 6 English Language Arts assessment were built for the study. First, a form which would be used with enhanced directions/scaffolding and reduced options for MC items was developed using items released prior to 2008. This test was a pilot form for the MEAP-Access, a new alternate assessment based on modified achievement standards (AA-MAS). The MEAP-Access measures the same grade-level content expectations as the Michigan Educational Assessment Program (MEAP), the state's general education assessment under their accountability program. Item properties and statistics from prior administrations were used in developing the new MEAP-Access form.

Then, a form parallel with the MEAP-Access in terms of reliability and content composition was developed from released MEAP items. The parallel form consisted of released items from the October 2008 administration of the MEAP. The item scores in the regular parallel MEAP form were used as the standard test. The item scores for the MEAP-Access form, which was administered in February 2009, were used as the accommodated test. There were no common items across the two forms. Each of the forms had two types of items; reading items and writing items. The reading section consisted of 21 multiple-choice (MC) items. The writing section consisted of three MC items and one open ended (OE) item, which was scored using a 0-5 rubric. A total of 694 students took both forms.

Data Analysis

Test Score Summary

The test data were analyzed by form (MEAP & MEAP-Access), by subtest (reading & writing), and by disability status (SWD & SWOD). Table 5.5.1 shows the means and the standard deviations by form, disability status, and subtest. The N count represents the number of students with valid scores in the corresponding half-form. The combined score of the reading and the writing parts (TOTAL) also are presented.

A quick look at the table reveals that SWOD performed better on both the MEAP and MEAP-Access forms. This result is not surprising. However, the differences between the two groups were smaller in MEAP-Access than in MEAP. For both groups, average scores were notably higher in MEAP-Access, but the gains between MEAP and MEAP-Access were greater for the SWD than for SWOD. For example the average total scores increased from MEAP to MEAP-Access by 7 points (7.5 to 14.5) for SWD compared to 4.7 points (17.2 to 21.9) for SWOD. In addition, the differences in the standard deviations were much smaller in MEAP-Access than in MEAP. For example, the standard deviations for reading MEAP were 1.5 for SWD and 4.6 for SWOD, while the standard deviations for reading MEAP-Access were 3.7 and 3.8, respectively for SWD and SWOD. These results suggest that the accommodation used in the study increased achievement and spread scores for SWD more than for SWOD.

Table 5.5.1: Test Descriptive Statistics

Test	Subtest	Statistic	Disability status	
			SWD (500)	SWOD (194)
MEAP (Standard)	Reading	Mean	4.8	12.9
		SD	1.5	4.6
	Writing	Mean	2.7	4.3
		SD	1.2	1.4
	Total	Mean	7.5	17.2
		SD	2.0	5.5
MEAP-Access Accommodated	Reading	Mean	11.0	16.9
		SD	3.7	3.8
	Writing	Mean	3.5	5.0
		SD	1.4	1.5
	Total	Mean	14.5	21.9
		SD	4.4	4.8

Item Statistics

Table 5.5.2 presents item statistics for MEAP and MEAP-Access. Although the items were not the same for the two forms, the MEAP form was built to match the MEAP-Access form

in terms of reliability and content. Hence, it is fair to assume that the two forms had similar average difficulties based on the general population and under regular administration.

The results show that after the two forms were taken by the same sample of students, the MEAP-Access form turned out to be less difficult than MEAP. In reading, for example, the mean classical item difficulty was 0.337 for MEAP and 0.603 for MEAP-Access. The mean point biserial correlations for reading were similar, 0.384 and 0.371, respectively for MEAP and MEAP-Access. Although the writing part consisted of very few items, the same trend as in reading resulted in writing. That is, the MEAP-Access test was easier than the MEAP test.

Table 5.5.2: Summary of Item Statistics

	MEAP (Standard)			MEAP-Access	
Subtest	Statistic	Difficulty	Biserial	Difficulty	Biserial
	Mean	0.337	0.384	0.603	0.371
	Min	0.203	0.247	0.406	0.251
Reading	Max	0.546	0.462	0.712	0.466
Writing	Mean	0.401	0.319	0.621	0.358
MC	Min	0.290	0.232	0.579	0.296
	Max	0.483	0.385	0.695	0.409
Reliability	0.833			0.832	

Examination for Construct Changes

The extent to which the two forms, MEAP and MEAP-Access, have the same factor structure or construct was explored. To address this inquiry, an exploratory factor analysis was first conducted, followed by a confirmatory factor analysis. When conducting general linear factor analysis, binary scores such as those from multiple-choice items do not work properly. To circumvent this problem, test items are normally parceled together to form multi-level variables.

For the reading section of each form, four variables were built from the 21 items. Scores on five non-adjacent items were added to create each variable. As an example, scores on items 1, 5, 9, 13, and 17 were added to form the first variable. As a result, each variable was on a 0-5 scale. One random item was not used so that the new variables would all have the same range of values. For the writing portion, the open-ended item (0-5 scale) was assigned as one variable, and the remaining three items were combined as a second writing variable. These six variables, four reading and two writing, were submitted to factor analysis.

Exploratory Factor Analysis

The results of the exploratory factor analysis are shown in Table 5.5.3. The factor loadings and the percents of variance in each variable accounted for by the single factor were similar and had the same patterns across testing conditions. Assuming that the two forms were parallel in construct, the results imply that the enhanced directions accommodation did not appear to impact the underlying construct of the test.

Table 5.5.3: Factor Loadings and Variance by First Factor

Variable	Factor Loadings		% variance by factor	
	MEAP	MEAP-Access	MEAP	MEAP-Access
Reading 1	0.792	0.790	63%	62%
Reading 2	0.792	0.752	63%	57%
Reading 3	0.768	0.784	59%	61%
Reading 4	0.752	0.747	56%	58%
Writing MC	0.663	0.630	44%	40%
Writing OE	0.561	0.623	31%	39%

Shown in table 5.5.4 are the first five eigenvalues from the factor analysis, which suggest that both forms appear to be measuring a single prominent dimension.

Table 5.5.4: First Five Eigenvalues and their Differences

MEAP		MEAP-Access	
Eigenvalues	Differ from next one	Eigenvalues	Differ from next one
3.164	2.396	3.147	2.410
0.768	0.112	0.737	0.065
0.656	0.147	0.672	0.128
0.509	0.018	0.543	0.081
0.491	0.078	0.462	0.024

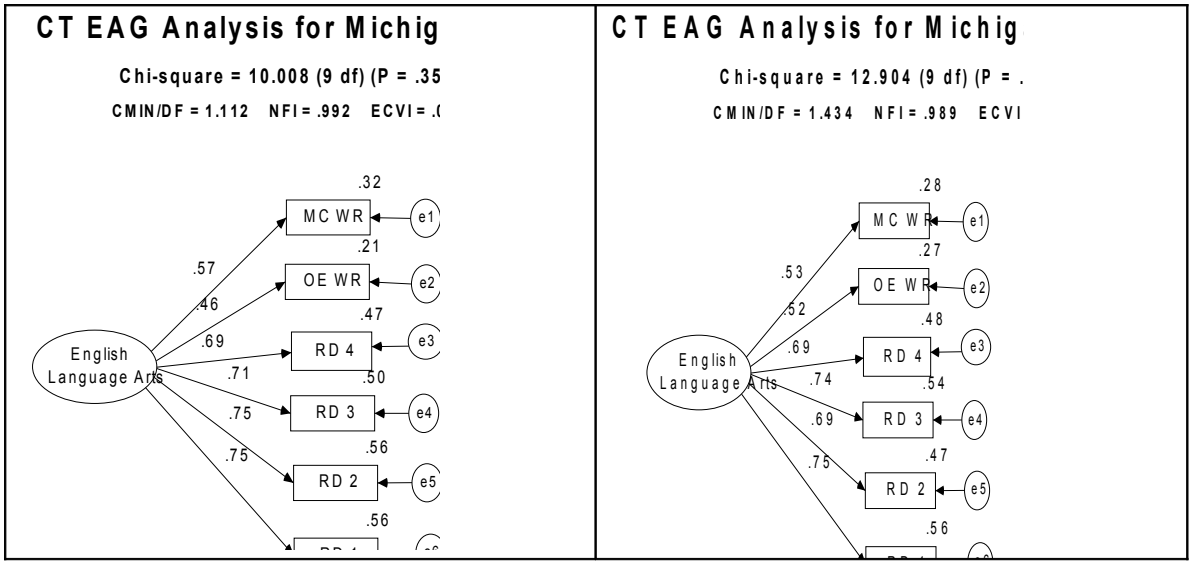
Confirmatory Factor Analysis

The goal in conducting the confirmatory factor analysis was to investigate whether the two forms (MEAP and MEAP-Access) had the same factor structure. To this end, a single factor model was fit to the MEAP results, and then the same factor model was fit to the MEAP-Access results. Figure 5.5.1 shows the results from fitting the same factor model on the two forms.

The numbers on the arrows that go from the unobserved variable (English Language Arts factor) to the observed variables represent the factor loadings. For practical purposes, the loadings are quite high, which indicates that the observed variables measure the common factor very well. The numbers on the top of the observed variables represent the squared-multiple correlations. These indicate the percent of variance in each variable accounted for by the common factor. For both test forms, the squared-multiple correlations were high, an indication that the model fit was good.

Besides the chi-square and the probability statistics, there are three other fit indices presented in the charts. The first index, CMIN/DF, is the ratio of the Chi-square and DF. It is recommended to be less than 2.0 to support model-data fit. The Normed Fit Index (NFI) is recommended to be higher than 0.9 to indicate good model-data fit. Finally, the Expected Cross Validation Index (ECVI) represents the change that would be expected if the analysis were repeated with a sample similar to the current sample. The smaller the ECVI, the better the model fit. All three fit indices indicated good fit.

Figure 5.5.1: Results of Confirmatory Factor Analysis



Analysis of Variance

For examining whether the enhanced directions accommodation affects scores for SWD differently than scores for SWOD, a Repeated Measures ANOVA with two independent variables (disability status & test form) and two dependent variables (total scores in MEAP & MEAP-Access) was conducted. To treat the testing condition as an effect, test scores in MEAP and MEAP-Access were treated as repeated measures – pretest and posttest, respectively. It was assumed that MEAP and MEAP-Access could be treated as repeated measures because both MEAP and MEAP-Access were built to be similar from a pool of released MEAP items and the only difference between MEAP and MEAP-Access test forms was the inclusion of enhanced directions on MEAP-Access.

Table 5.5.5 shows the outcome of the Repeated Measures ANOVA. In part (a), it is shown that disability status significantly affected reading scores, test form also affected test scores, and there was a significant interaction effect between disability status and test form. In part (b), the means are presented for the two main effects and shown in Figure 5.5.2, which follows.

Table 5.5.5: Results for Repeated Measures ANOVA

(a) Repeated Measures

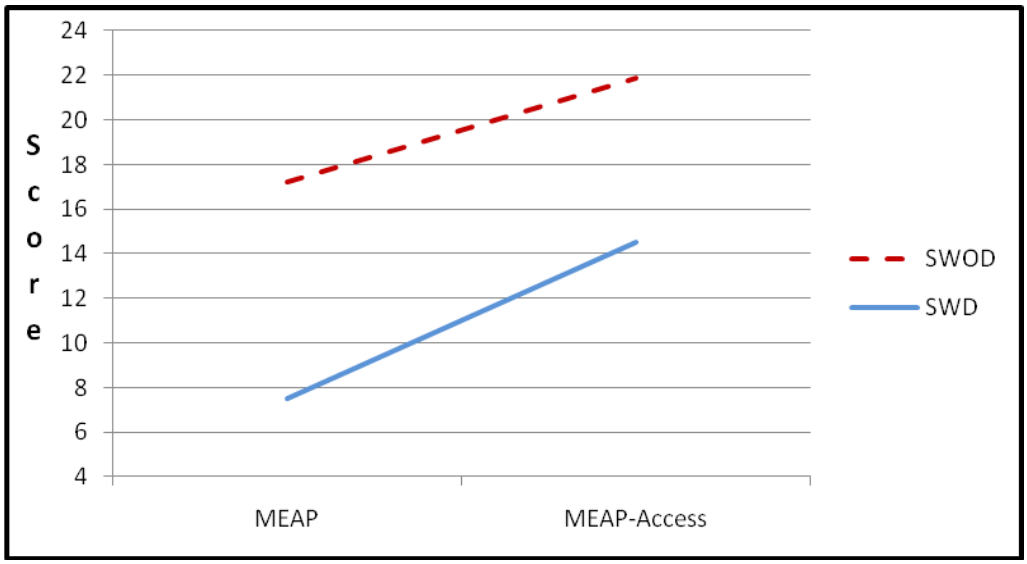
Source	DF	SS	Mean Sq.	F value	Pr > F
Disability Status (between subjects)	1	20635.6	20635.6	926.32	<0.0001

Test Form (within subject)	1	9531.76	9531.76	1014.03	<0.0001
Disability × Form (within subject)	1	368.82	368.82	39.24	<0.0001

(b) Means

Disability status	MEAP	MEAP Access
SWD	7.5	14.5
SWOD	17.2	21.9

Figure 5.5.2: Plot of Means for Disability Status by Accommodation Condition (Michigan Study)



Conclusion

In terms of the overall scale (i.e. total score), the MEAP-Access form was found to be easier than the MEAP form. With the assumption that the two forms were similar in difficulty when they were built, it is fair to say that the accommodation (enhanced directions) made the MEAP-Access test easier than MEAP. The results also indicate that the SWD group benefited more from the accommodation than the SWOD group did.

For interpretation purposes, the factor structure of the two forms was found to be similar. Both forms were found to have a single factor, which could be interpreted as an English Language Arts factor. The examination of the accommodation effects on test scores revealed that (1) disability status has significant effect on test scores, (2) the accommodation has significant effect on test scores, and (3) the accommodation effect was different between SWD and SWOD (interaction effect). The last finding suggests that SWD gained more from the accommodation than SWOD did.

VI. Overall Summary and Discussion of Accommodations Validity Studies

As described in the previous chapters, in order to conduct a validity study, each state had to organize its study according to the research design, obtain a sample of students, conduct the test administration, and provide data files for analysis. Then, a collection of analyses was done for each state's study to look at the impact of the test accommodation on validity. The detailed results from all the analyses were presented in Chapter V.

In the current chapter, an overall summary of the validity studies, their general characteristics, information on how well the studies met the project's criteria for design and samples, and a summary of the findings across the five studies are presented. A discussion of the results in the context of each research question and implications for the education community follows. Additional follow-up analyses that were conducted to help clarify some of the results also are described.

The table on the next page provides an overview of the studies and includes information such as the accommodations used, content areas and grades involved, and when the test administration occurred. As shown in this table, in four of the five studies, states decided to study a read aloud accommodation. However, states used several different approaches for this. Connecticut used a computerized text reader to deliver the accommodation, with a digital voice for a reading comprehension test for its first study, and a recorded human voice for a mathematics test for its second study. Kentucky and Nevada used trained human readers for the read aloud accommodations, but in different grades and subject areas—Kentucky in reading at grade 4 and Nevada in mathematics at grade 7. Michigan compared the results from a pilot version for its alternate assessment based on modified achievement standards (AA-MAS) to the general test. The state viewed the enhanced directions/scaffolding as an accommodation for its AA-MAS. Four states administered their studies at a variety of grade levels in the areas of mathematics and reading, while Michigan targeted reading and writing in their English language arts assessment.

Because of timing issues and state schedules for their regular ongoing state assessment programs, the administration dates for the studies varied somewhat. Validity studies were conducted in a time frame that ranged from May 2008 to May 2009. Two states administered the studies in the fall-winter time frame (late 2008 to early 2009). All states administered their studies at times that did not conflict with their regular state assessments.

Summary of General Characteristics of the Five State Validity Studies

Topics/Issues	Connecticut # 1	Connecticut #2	Kentucky	Michigan	Nevada
Accommodation	Computerized text reader with a digital voice	Computer-delivered with a recorded human voice	Read aloud by a trained human reader	Enhanced directions/ scaffolding and fewer response options	Read aloud by a trained individual (e.g., teacher or aide)
Grade	7	5	4	6	7
Content Area	Reading Comprehension	Mathematics	Reading	English Language Arts	Mathematics
Administration: Test Date	May 2008.	Spring 2009.	Mar. 2009.	Oct. 2008 and Jan. 2009.	Nov.-Dec. 2008.
Study Design, Items, and Forms	Computer-based test forms with 20 MC items per form.	Computer-based test forms. Used old secure math test with 26 MC items per form.	Developed forms from a pool of non-state items that were similar in content to KY forms, with 22 MC items per form.	Used released items (24 MC and 1 CR). Data from MEAP-Access for SWD and MEAP admin done earlier.	Used 20 MC items per form.
Vendor	Measurement Inc.	Measurement Inc.	Measured Progress	Measurement Inc.	Measured Progress and WestEd
State contact(s)	B. Beaudin M. Dirir G. Andrada J. Amenta J. Struck	B. Beaudin M. Dirir J. Amenta G. Andrada J. Stuck	K. Draut R. Ervin	S. Viger J. Martineau J. Griffiths A. Wyse	C. Crothers C. Sharp A. Parr

Summaries of Findings from State Validity Studies

In Chart 1, general information is provided on whether states met two important criteria for conducting the studies: use of the 2 × 2 research design and adequate sample size. All states except Michigan followed the design that was specified in the guidelines. Michigan used an alternate design that was approved for the study after a series of discussions with the CTEAG PMT. Michigan's results were included because, despite the change in the design, their study

still provides important information about the use of accommodations in tests used for accountability under current federal regulations. The design utilized a pilot version of an assessment based on the general state assessment, the Michigan Education Assessment Program (MEAP), in preparation for a new state test, the MEAP-Access. At the time of the study, the MEAP-Access was being designed as an AA-MAS developed by the state under NCLB regulations. For this validity study, a pilot form was developed to be parallel to the MEAP in content using released items from the test. The pilot form had one answer choice removed and was accompanied by enhanced directions/scaffolding. Students in the study took the items from the two forms at separate times and the forms did not share any items (i.e., there were no common items between MEAP and the pilot form). Because of this, some of the analyses could not be performed (as described further below).

All states met the minimum sample size requirements of at least 150–200 students in each subgroup.

Chart 1. General Information on Requirements Criteria for Studies

General Information	**Connecticut # 1**	**Connecticut # 2**	**Kentucky**	**Michigan**	**Nevada**
2 × 2 Design Followed?	✓	✓	✓	Alternate Design Used	✓
Adequate Sample?	✓	✓	✓	✓	✓

In Chart 2, the findings from the analyses for the three research questions are presented. For the first research question, the item difficulties and point-biserial correlations (a measure of discrimination) were examined. In general, the item statistics were unchanged across the two testing conditions, except in the case of Kentucky, where items had higher p-values (percent correct) under the accommodated condition (i.e., students performed better), and values for some point-biserial correlations decreased. Because no items were common across testing conditions in Michigan, the effect of the accommodation on individual item statistics could not be analyzed.

For the second research question, the exploratory and confirmatory factor analyses of test structure found no changes in dimensionality or factor structure across the five studies. For the third question, the analyses of main and interaction effects found that the accommodations significantly increased test scores for students in two of the five studies (Kentucky and Michigan). Analysis of the interaction effects found that scores for SWD increased more than scores for SWOD in three of the studies (Connecticut #2, Kentucky, and Michigan). In Connecticut study #2, scores for SWD increased slightly, whereas scores for SWOD decreased

under the accommodated condition. In both Kentucky and Michigan, scores for both groups increased, with those for SWD increasing much more than those for SWOD.

Plots of the interaction effects for these three studies were shown in the detailed results presented in Chapter V.

Chart 2. Summary of Findings from CTEAG Validity Studies

Accommodation Effects	**Connecticut # 1**	**Connecticut # 2**	**Kentucky**	**Michigan**	**Nevada**
Item Statistics Not Changed	✓	✓	Changes in item stats	Not applicable	✓
Test Dimensionality Not Changed	✓	✓	✓	✓	✓
Test Structure Not Changed	✓	✓	✓	✓	✓
Significantly Increased Test Scores	NO	NO	YES	YES	NO
Significantly Increased Test Scores for SWD More Than SWOD (Disability × Accommodation)	NO	YES	YES	YES	NO

Overall Discussion

In this section, a synthesis of the findings across the five validity studies is presented along with a discussion and interpretation of the results in light of their practical use by states and others. In order to make fair comparisons across the studies, the overall findings for Connecticut studies #1 and #2, Kentucky, and Nevada will be discussed first. The findings for Michigan will be discussed separately since that study was conducted under different conditions and focused on the use of a pilot form of an AA-MAS.

<u>Research Question 1: Does the accommodation affect item statistics?</u>

Overall, as was shown in Chart 2, the various test accommodations had little effect on item statistics. In the two studies conducted by Connecticut and the one by Nevada, the

difficulties and discrimination indices did not change, nor did the reliabilities of the test forms. Thus the use of read aloud test accommodations did not seem to have any impact on or change the item and test characteristics, which supports the validity of interpretations drawn from accommodated test administrations in those states. However, for the Kentucky study, several of the items became easier and the point-biserial correlations decreased under the accommodation condition, with some items' point-biserials decreasing dramatically. As a result of the decrease in test score variability, the reliabilities of the forms used in Kentucky also decreased slightly under the accommodated condition. A possible interpretation of this effect is that the read aloud accommodation used with the Kentucky reading assessment made the items in the test much more accessible and thus easier for students, as well as somewhat less discriminating because of less variability in the scores.

Research Question 2: Does the accommodation change the content structure of the test? For example, is there an impact on the factor structure under accommodated and standard administrations?

In general, the use of test accommodations did not seem to have any effect on the test constructs for the various assessments that were part of the validity studies. Based on the results from the exploratory and confirmatory factor analyses, none of the read aloud accommodations changed the dimensionality of the tests or had an impact on the content structure (i.e., test constructs). Similar factor structures were found for both the accommodated and non-accommodated test forms. Thus, based on the findings from the validity studies conducted for this research project, it is reasonable to say that when comparing accommodated to non-accommodated (standard) test administrations, the use of read aloud accommodations had no impact on the constructs being measured by the assessments.

Research Question 3: Does the accommodation increase performance of students? Does it increase performance more for students with disabilities than it does for students without disabilities?

In answer to this question, the results from the validity studies are mixed. The findings for the analysis of main effects from the repeated measures ANOVA found that the read aloud accommodations significantly increased test scores for all students (SWD and SWOD combined) in one of the studies (Kentucky). The accommodations did not significantly increase scores for students overall in the two Connecticut studies or in Nevada (although scores did increase in Nevada, but not quite enough to be statistically significant). For the former two studies, this may have had something to do with the mode of test administration used in Connecticut (i.e., use of a computerized read aloud and students' overall lack of familiarity with the computer delivery). The significant main effects found for Michigan are discussed below.

However, besides examining main effects, the possibly more interesting research question is whether there was a differential boost effect from the test accommodations. Again, the results were mixed. Analysis of the interaction effects found that scores for SWD increased more than scores for SWOD in two of the four studies (Connecticut #2 and Kentucky). This

finding provides support in the potential benefits of using the read aloud accommodations to make the test more accessible for SWD and thereby improve their performance. In addition, some SWOD in these two states also benefitted from use of the accommodations, which is a reasonable outcome when one considers that some students who do not have a disability may have access needs that also are helped by the read aloud accommodation. Note that no significant interaction effects were found for the Connecticut #1 or Nevada studies. The significant interaction effects found for Michigan are discussed below.

Discussion of Results from Michigan's Validity Study

As described earlier, the validity study conducted by Michigan used a slightly different approach from those of the other states in the CTEAG project. While the other studies in this project focused on accommodations to their general assessment, the Michigan study looked at a pilot form for a future alternate assessment (MEAP-Access) based on modified achievement standards (AA-MAS), which was compared to performance on a parallel form of the general assessment (MEAP). Both MEAP and MEAP-Access will measure the same grade-level content expectations, but the achievement standards will be different.[2]

In late 2008, as Michigan was beginning the process to design the new MEAP-Access assessment for use as an AA-MAS, the state had a unique opportunity to evaluate the validity of enhanced directions on the English Language Arts portion of their assessment within the framework of the CTEAG study. For the pilot form, MEAP items were revised to make them more appropriate for students with disabilities, which included the use of enhanced directions/scaffolding and fewer options for students to choose from on multiple-choice questions. For the purposes of the study, the enhanced directions were viewed as an accommodation. The use of enhanced directions with the pilot MEAP-Access form allowed the test administrator to read and clarify parts of the test to help students better access the reading and writing portions of the assessments.

The objectives of the Michigan validity study were to examine whether the use of enhanced directions alters item properties and test structure, and whether the revisions affect test scores for SWD differently than test scores for SWOD. However, due to lack of common items between the two forms that were studied (MEAP and pilot MEAP-Access) it was not possible to assess the effects of the enhanced directions/scaffolding on individual test items since none of the items were used under the two conditions—once in the regular administration and once in the AA-MAS condition. Looking at students' performance on the test forms in terms of average difficulty, it was found that average performance on the MEAP-Access form was higher than on the MEAP form. Thus, in keeping with the intention of an AA-MAS, the changes made the set of test items on the pilot form less difficult, on average, than their original iteration. The results also

[2] According to the U.S. Department of Education (2007), "A modified academic achievement standard is an expectation of performance that is challenging for eligible students, but may be less difficult than a grade-level academic achievement standard" (p. 14).

indicated that the SWD group benefited more from the enhanced directions/scaffolding condition than the SWOD group did.

For interpretation purposes, the factor structure of the two forms was found to be similar (i.e., the changes did not affect the test construct being measured). Both forms were found to have a single underlying factor, which could be interpreted as an English Language Arts factor. The examination of the main and interaction effects on test scores revealed that the use of enhanced directions had a significant and substantial effect on increasing test scores, and the interaction effect was significant, with SWD differentially gaining more from the accommodation than SWOD did. Although both SWD and SWOD showed increases in their test scores, there was a differential boost and the effect was much larger for SWD. Therefore, based on the results from these analyses, it appears that the pilot MEAP-Access with enhanced directions/scaffolding used in Michigan had a major impact on increasing the accessibility of the test and on increasing the scores of students, especially benefiting the SWD group.

Follow-up Exploratory Analyses of State Data

During the course of reviewing the data and preliminary results from the state studies, questions were sometimes raised concerning data issues, student performance, and how to interpret the results. CTEAG project staff conducted a number of follow-up investigations with individual states to gather more information or verify issues with the data, discuss possible interpretations of the findings, and/or propose additional analyses that could be done to dig deeper into the data. Often, these exploratory analyses helped generate more hypotheses regarding the effects of the accommodations on students. This information was used to guide additional research and discussion.

For example, while looking at the impact of accommodations on the statistical properties of the items and test forms, changes in some of the item statistics were found for Kentucky's study, which was the only study that had this type of effect. In conjunction with Kentucky, individual items were reviewed to determine if there was something different about the ones that had changes in difficulty or discrimination. Based on these reviews, no plausible explanations were found. Following further discussions, additional analyses were conducted to look specifically at the distribution of scores for students and whether there were any differences for low versus high performing students on their performance with and without the read aloud accommodation. Again, nothing that would explain the changes in total test form statistics was found.

Further discussions with Kentucky staff and their vendor generated more hypotheses for research. One of the discussions addressed the decrease that was noted in test reliability associated with the accommodation condition. It was thought that the decrease was probably the result of an overall decrease in score variability from use of the accommodation, and although it was not a large decrease, project staff agreed to look more closely into it. Thus, another post hoc analysis was done, this time to look more closely at the impact of the test accommodation on

reliability by subgroup. Both Cronbach's Alpha and standard errors of measurement (SEM) were calculated for SWD and SWOD so that differences in reliabilities could be interpreted in light of the score variability of each subgroup and test form. From this analysis, it was found that the magnitude of SEMs indicated that any observed differences in score precision among the test forms, and between SWD and the whole sample, were trivial.

In addition, one more analysis was proposed to investigate the hypothesis that passage length may be related to changes in item statistics, i.e., longer passages would have more changes in item statistics, so project staff analyzed the data from Kentucky further. However, for this analysis, it was difficult to tie item or score changes to passage length since five of the six passages used in Kentucky's reading test had very similar total word counts. However, one passage that had more items and less word count showed smaller changes in item statistics than the others. It is important to note that, because of the small sample sizes, it was difficult to make any definitive conclusions from some of these post hoc analyses that looked at subgroup performance.

One of the things discovered in doing this research project was the importance of examining the results carefully and trying to make interpretations in relation to possibly unexpected findings. Sometimes, additional hypotheses were generated and follow-up exploratory analyses conducted to look at the data in different ways. For example, another issue that received attention from the project team and state participants pertained to the differential boost hypothesis. In examining the findings from the ANOVA, it was noted that significant interaction effects were found for the studies done in Connecticut (#2), Kentucky, and Michigan. The study done in Connecticut, although it showed scores to increase slightly for SWD, actually showed scores to decrease for SWOD, which was not in the direction of a differential boost. The significant effects found for Michigan were likely due to the different items and tests that were used, in combination with the enhanced directions/scaffolding, which had a strong effect on boosting performance for students. The significant effects found for Kentucky led to much additional discussion with Kentucky staff, some of which was described above. One of the most important explanations concerning the impact of the read aloud accommodation was that it increased accessibility to the test for both SWD and SWOD, but much more so for SWD. As noted elsewhere in this report, it was hypothesized that some SWOD who also may have access needs benefited from use of the accommodation. The differential boost hypothesis seems to be supported based on the data from Kentucky, as well as from Michigan.

In addition, in order to investigate the hypothesis that the accommodations may be affecting students in different ways based on their overall performance on the test, follow-up analyses were conducted on the data sets from all five studies to look at the data in different ways. Another series of repeated measures ANOVAs were run, after manipulating the samples in different ways to make different groupings based on performance. Groupings were made for the top ¾ performers (students whose scores were in the top three quartiles of the distribution), bottom ¾ performers (students whose scores were in the lower three quartiles of the distribution), two halves of SWD, and two halves of SWOD. The means for the subgroups were examined and compared to the results obtained from the original analyses conducted for each

study. One observation noted based on this analysis was that, on average, students in the lower performing group sometimes showed more gains in scores than those in the higher performing group. This could mean that the accommodations had more of a positive effect, such as improved accessibility to the test to a greater degree for the lower performing students. This effect was more evident for SWD, although SWOD also exhibited some gains from the accommodation. In general, no definite patterns from this follow up analysis were noted across all the studies. However, this type of research may be of interest for further study. It is recommended that other researchers who may want to do this type of analysis should take care to obtain much larger sample sizes in order to perform subgroup analyses in various ways. The results from this collection of analyses are provided in Appendix F.

One other analysis came about following a discussion with the Technical Advisory Committee for Connecticut, who was being briefed on the CTEAG studies. When asked about the effect sizes for the five studies, additional analyses were run to calculate them for the significant accommodation and interaction effects for each study, and the data summarized. Large effect sizes were found for the testing condition effect in Kentucky (0.30) and for the test form effect in Michigan (0.59). Small effect sizes (0.05 or less) were seen for the other studies. Although a significant interaction effect was found for Connecticut study #2, it had a small effect size. The results from this analysis also are provided in Appendix F.

Another analysis that was conducted on an exploratory basis included an ANCOVA of the Nevada data to see if use of a covariate might help to better interpret the results. After reviewing the data, it was decided to forego this approach and maintain the repeated measures ANOVA approach as it provided cleaner and more straightforward results for interpretation.

Because of the different design used for the Michigan study, numerous discussions were held with Michigan's staff to review their data, and project staff worked closely with the state to review the results and interpret the findings. Although it was not possible to conduct some analyses, such as for item statistics, other analyses were done.

In addition to the variety of analyses that were conducted to attempt to discover more about the effects of the accommodations, additional hypotheses were generated during the course of the study, which unfortunately, could not always be answered because of limitations with the data or design, or both. Many of these ideas are summarized in the next section of this report as possible areas for further research.

In conclusion, it is important to note that the research conducted by the CTEAG does not answer all questions concerning accommodations and the types of students that may need certain ones. Many of these additional analyses were exploratory and done ad hoc or post hoc to figure out more about the data and student performance. This research approach should be viewed as a method of inquiry to generate hypotheses for more detailed follow-up analysis or further research.

Practical Applications and Implications for Use of Results from CTEAG Validity Studies

This technical report has provided a summary of the research design, procedural guidelines, and methodologies used for the collection of validity evidence on test accommodations, as well as details of the findings from the five state studies and discussion of the results. In this section, information for state assessment staff and others on conducting validity studies for test accommodations and interpreting the results is presented, emphasizing the practical use of the data and possible future applications. Because the research can provide information at a variety of levels, this section also focuses on the implications for the use of the results for the following groups: states, teachers, policymakers, and researchers.

States

One of the most important uses of the information from the CTEAG project is to help states refine their decision-making procedures with respect to the appropriate accommodation of students with disabilities on the general assessment and the appropriate inclusion of students on the alternate assessment. Accommodation decisions need to be supported by data, and the collection of studies from this project provide data on various types of read aloud accommodations that states can use for increasing the accessibility of their assessments while maintaining the validity of interpretations drawn from those assessment results.

In discussion with state members who participated in the project, a number of ways that the results can be used to benefit states were identified. These included use with peer review, justifying states' use of specific accommodations for accountability, validating the use of read aloud accommodations with various state tests, providing access to additional data from validity studies (such as the project database), and possibly providing assistance toward standardizing the use of accommodations and obtaining a more consistent and cost-effective product from vendors.

In addition, the CTEAG project provides evidence that collaborative research can be conducted by a group of states and that all can benefit from the work. Although each state may be somewhat different in its capacity to do research, with appropriate support and guidance a state can conduct its own validity study on test accommodations. The guidelines provided in this project can guide any state that wants to do such a study.

States also can learn from one another based on the findings that are shared from this type of collaborative project. It is feasible that a state that uses a similar type of accommodation in its assessment program, such as a read aloud delivered by a trained individual or a read aloud delivered by computer, can assume that the findings from one state's study will generalize to its own as long as the state has carefully considered the comparability of the constructs being measured. In other words, an accommodation that increases the accessibility of the test for students with disabilities in one state would do likewise in another state, given that other aspects of the state content standards and assessment programs are similar.

As noted earlier, states continue to require additional evidence in support of their decisions on the use of various accommodations. The results from the CTEAG project can help provide this information. Because the final reports and database include information on research designs, procedures, and findings from the five validity studies, these products can be a valuable resource for any state that needs to reference existing data and evidence to inform, evaluate, and support its decisions on the use of various types of accommodations. Also, because the results from the states cover a range of different accommodations, the project provides a resource that few, if any, states would be able to develop alone. This compilation and sharing of information can be a great benefit for states in meeting NCLB peer review requirements for standards and assessments by providing additional validity information about student achievement and the performance of students with disabilities.

Teachers

Many educators can benefit from more guidance based on findings from scientifically based research projects to help them make informed decisions concerning the best ways to assess their students. Classroom teachers, in particular, will benefit by using the results from the CTEAG studies to inform their decisions about which accommodations should be used with individual students. It is important that teachers understand how certain accommodations impact the validity of the tests they give and the interpretations of scores from the tests, either for the state tests or on teacher-developed classroom assessments. It is hoped that teachers also will become more sophisticated with their use of data and findings from research studies as they work with or serve on IEP teams to select which accommodations will be most appropriate for increasing access for individual students.

Policymakers

Policymakers should use empirical evidence when making policies related to which accommodations should be allowed on specific tests and how test scores will be interpreted if a specific accommodation is used. Policy decisions should be based on research findings like the results from the CTEAG validity studies. This empirical data can inform and support the validity of the assessment program.

It is important that policymakers recognize their influence on and support of additional research on test validity. Policymakers can encourage more research on accommodations data within their state assessment programs through such actions as making operational test data available to researchers or collecting more information on the use of accommodations in their jurisdiction. More data on the specific types of disabilities, such as ADHD, dyslexia, or other learning disabilities, and what accommodations were provided to these students could also be useful to collect.

Researchers

As noted earlier, the research design used in the CTEAG studies can easily be reproduced by additional states in future years to study the validity of other accommodations. Other researchers external to the state agencies also can utilize the design if they wish to conduct studies of this type. The findings can then be added to the growing body of knowledge that is being built from this project and others. It is hoped that a growing compilation of validity studies on test accommodations is created, with each new study being added to the existing collection. In this way, a shared body of knowledge will exist for all interested in accessing more information on the topic.

In addition, the project is providing data from the five studies in a database that can be used by researchers to conduct different types of analyses to answer their particular questions. More information on the database can be found in the following section.

Although a number of issues described in the previous section are challenging to the field (small samples, heterogeneity of samples, bundled accommodations, etc.), researchers should be able to address many of these issues in the future and continuously improve the designs and approaches used to study test accommodations.

Need for Additional Research

Despite the numerous vexing issues and challenges in conducting these types of studies, as described above, much progress has been made in recent years in carrying out research on the validity of test accommodations. Increasing amounts of high-quality research are being done in this area and the findings are being shared more widely so they can have a positive impact on policy and practice. In the Appendix, a list of experimental studies and recent research on the validity of accommodations is provided (see Accommodations Bibliography).

It is critical that researchers continue doing these types of studies in spite of the many challenges and issues, as the results will only help to increase the accessibility of assessments for students with disabilities and improve the overall quality and validity of assessment results. The CTEAG project and its collaborative approach to conducting and compiling state-based research studies is just a small, but significant, step forward in helping to answer the many questions pertaining to the validity of accommodations.

Study Limitations and Areas for Further Research

Throughout the course of the CTEAG project, the researchers noted various issues where methodologies and approaches could be addressed further and possibly be improved. In this section, it is important to discuss possible problem areas that may affect interpretations of the results, with the expectation that, if possible, these areas be addressed in future research.

First, the results from this collection of validity studies are specific to the accommodations that were used and the contexts in which they were applied. The majority of studies examined the effects of read aloud types of accommodations, so it is not necessarily possible to generalize the results that were found to all other accommodations.

Carrying out research on students with disabilities is often challenging for a number of reasons. For example, obtaining adequate sample sizes for SWD in order to conduct statistical analyses can be difficult, if not sometimes impossible. Many of the techniques typically used to analyze test data are methodologies that require large samples—for example, factor analysis or structural equation modeling usually requires sample sizes of at least 300 observations (students). Other statistical approaches usually require at least 100 students per group and preferably more to have enough power to detect small effect sizes. However, in some states it may be impossible to obtain sample sizes of even 100 for low-incidence disabilities within small populations of students. For example, the number of students who use a Braille accommodation at a given grade level may be less than 10 in an entire state. Although the CTEAG project required states to select accommodations for their validity study only if there were at least 150 students in each group, the results from the studies may have been stronger if the sample sizes had been larger. In addition, other types of analyses that are useful for comparing the performance of subgroups of students could have been done with larger sample sizes, for example, conducting a Differential Item Functioning (DIF) analysis to evaluate for item-level differential performance.

Another area that needs to be considered when doing these types of validity studies is that the SWD samples are often quite heterogeneous. There is much variation in any type of disability and the classification of students into disability categories may cover a broad definition of that disability. The natural variability in the student samples needs to be taken into account when planning for a study and in interpretations of the findings.

In order to interpret the results more fully, CTEAG attempted to collect data on the specific types of disabilities that students in the state samples had. However, many states did not have this information in their data files and some do not collect it as part of their assessment programs. So, this was not possible. This is an area in which more data on SWD would be useful for researchers.

A related issue is that many students, such as students who have visual or auditory disabilities, use a bundle of accommodations during the test administration. This makes the study of a single type of accommodation problematic if the student is used to getting a package of accommodations, for example a read aloud is often given in an individual administration with extended time. Isolating the impact of just one accommodation and disentangling the effects on the validity of test scores can be difficult. This issue was discussed in depth by the CTEAG PMT, and unfortunately there was no simple answer. In discussions with the states, it was decided to study a single accommodation in each of the studies, given the realities mentioned above. In the future, it would be beneficial for researchers to study the effects of combinations or bundles of accommodations on test validity, as the results would be more comparable to what actually is being done in the classroom and on state assessments.

Finally, researchers have learned that an important consideration to be taken into account is determining more clearly which students benefit from an accommodation that is intended to increase access and provide students with the opportunity to demonstrate their knowledge, skills, and abilities. Follow-up analyses of large-group data and studies that include measures of student

access needs will help us better provide accommodations that support valid inferences. In addition, more study of the provision of accommodations to students is needed—for example, there is little research about the degree to which the accommodations selected by IEP teams are the accommodations most appropriate for increasing access.

VII. Additional Information Provided in the CTEAG Guidebook, Project Database, and Lessons Learned Report

The other deliverables from the CTEAG project are a guidebook, a project database, and a report describing the lessons learned by states when conducting validity studies of test accommodations.

The guidebook serves as a practical guide for state assessment staff and others on conducting validity studies for test accommodations and interpreting the results. It provides a synthesis of the findings from the five state studies conducted for this project, with an emphasis on the use of the data and possible future applications.

The project database includes all data from the studies, at the student and item level, as well as results from the statistical analyses that were conducted for each of the state studies. Additional information on the studies is provided in the database so other researchers can replicate the analyses or do further analyses using the data.

The lessons learned report provides information on the context for the studies, challenges encountered by states in implementation, and suggestions for other states that may consider doing these types of studies.

The collection of all four products from this project can be a valuable resource to states and others interested in the topic. They are available as PDFs on the CCSSO (www.ccsso.org) and CSDE (www.sde.ct.gov/sde) websites.

References

Abedi, J., Courtney, M., & Leon, S. (2003). *Effectiveness and validity of accommodations for English language learners in large scale assessment* (CSE Technical Report No 608). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for Educational and Psychological Testing*. Washington, D.C.: American Educational Research Association.

Arbuckle, J.L. (2007). *Amos 16.0 User's Guide*. Chicago: SPSS, Inc.

Barton, K.E. (2002). Stability of constructs across groups of students with different disabilities on a reading assessment under standard and accommodated administrations (Doctoral dissertation, University of South Carolina, 2001). *Dissertation Abstracts International*, 62/12, 4136.

Barton, K.E. & Huynh, H. (2003). Patterns of errors made by students with disabilities on a reading test with oral reading administration. *Educational and Psychological Measurement*, *63*(4), 602-614.

Barton, K., & Winter, P.C. (2009). Alternative formats: A review of the literature. In Phoebe C. Winter (Ed.), *Evaluating the Comparability of Scores from Achievement Test Variation*, (pp. 228-236). Washington, DC: Council of Chief State School Officers.

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

Bridgeman, B., Cline, F., & Hessinger, J. (2004). Effect of extra time on verbal and quantitative GRE scores. *Applied Measurement in Education*, *17*(1), 25-37.

Burch, M. (2002). Effects of computer-based test accommodations on the math problem-solving performance of students with and without disabilities (Doctoral dissertation, Vanderbilt University, 2002). *Dissertation Abstracts International*, 63/03, 902.

Cahalan, C., Mandinach E., & Camara, W.J. (2002). *Predictive validity of SAT I: Reasoning test for test-takers with learning disabilities and extended time accommodations.* New York, NY: The College Reporting Board.

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.

Christensen, L.L., Lail, K.E., & Thurlow, M. L. (2007). *Hints and tips for addressing accommodations issues for peer review.* Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Christensen, L.L., Thurlow, M.L., & Wang, T. (2008). *Improving Accommodations Outcomes: Monitoring Instructional and Assessment Accommodations for Students with Disabilities.* Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Cook, L., Eignor, D., Sawaki, Y., Steinberg, J., & Cline, F. (2006). *Using factor analysis to investigate the impact of accommodations on the scores of students with disabilities on English language arts assessments*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

DePascale, C. (2009). Evaluating linguistic modifications: An examination of the comparability of plain English mathematics assessment. In Phoebe C. Winter (Ed), *Evaluating the Comparability of Scores from Achievement Test Variation*, (pp. 71-96). Washington, DC: Council of Chief State School Officers,

Elliott, S.N., Kratochwill, T.R., & Schulte, A.G. (1998). The assessment accommodation checklist: Who, what, where, when, why, and how*? Teaching Exceptional Children, 31* (2), 10–14.

Elliott, S.N., & Marquart, A.M. (2003). *Extended time as an accommodation on a standardized mathematics test: An investigation of its effects on scores and perceived consequences for students with varying mathematical skills.* Madison, WI: University of Wisconsin-Madison, Wisconsin Center for Education Research.

Elliott, S.N., & Marquart, A.M. (2004). Extended time as a testing accommodation: Its effects and perceived consequences. *Exceptional Children*, *70*(3), 349-367.

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

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

Fuchs, L.S., Fuchs, D., Eaton, S.B., Hamlett, C.L., & Karns, K.M. (2000). Supplementing teacher judgments of mathematics test accommodations with objective data sources. *School Psychology Review, 29*(1), 65–85.

Gibson, D., Haaeberli, F.B., Glover, T.A., & Witter, E.A. (2005). Use of recommended and provided testing accommodations. *Assessment for Effective Intervention, 31* (1) [Special issue: Testing Accommodations: Research to Guide Practice], 19-36.

Hambleton, R.K., Swaminathan, H., & Rogers, J. (1991). *Fundamentals of Item Response Theory*. Newbury Park, CA: Sage publications.

Helwig, R., Rozek-Tedesco, M.A., & Tindal, G. (2002). An oral versus a standard administration of a large-scale mathematics test. *The Journal of Special Education*, *36*(1), 39-47.

Hu, L. & Bentler, P.M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*. 6, 1-55.

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

Huynh, H., Meyer, J.P., & Gallant-Taylor, D. (2002). *Comparability of scores of accommodated and non-accommodated testings for a high school exit examination of mathematics.* Paper presented at the annual meeting of the National Council on Measurement in Education. New Orleans, LA.

Huynh, H., Meyer, J.P., & Gallant, D. J. (2004). Comparability of student performance between regular and oral administrations for a high-stakes mathematics test. *Applied Measurement in Education*, *17*(1), 39-57.

Janson, I.B. (2002). The effects of testing accommodations on students' standardized test scores in a northeast Tennessee school system (Doctoral dissertation, East Tennessee State University, 2002). *Dissertation Abstracts International*, 63/02, 557.

Johnstone, C.J., Altman, J., Thurlow, M.L., & Thompson, S.J. (2006). *A summary of research on the effects of test accommodations: 2002 through 2004* (Technical Report 45). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Kettler, R.J., Niebling, B.C., Mroch, A.A., Feldman, E.S., Newell, M.L., Elliott, S.N., Kratochwill, T.R., & Bolt, D.M. (2005). Effects of testing accommodations on math and reading scores: An experimental analysis of the performance of students with and without disabilities. *Assessment for Effective Intervention, 31*(1) [Special issue: Testing Accommodations: Research to Guide Practice], 37-48.

Kishton, J.M., & Widaman, K. (1994). Unidimensional versus domain representative parceling of questionnaire items: An empirical example. *Educational and Psychological Measurement*, 54, 757-765.

Kobrin, J.L., & Young, J.W. (2003). The cognitive equivalence of reading comprehension test items via computerized and paper-and-pencil administration. *Applied Measurement in Education*, *16*(2), 115-140.

Laitusis, C.C. & Cook, L.L. (Eds.) (2007). *Large Scale Assessment and Accommodations: What Works?* Washington, DC: Council for Exceptional Children.

Landau, S., Russell, M., Gourgey, K., Erin, J.N., & Cowan, J. (2003). Use of talking tactile tablet in mathematics testing. *Journal of Visual Impairment and Blindness*, *97*(2), 85-96.

Lang, S.C., Kumke, P.J., Ray, C.E., Cowell, E.L., Elliott, S.N., Kratochwill, T.R., & Bolt, D.M. (2005). Consequences of using testing accommodations: Student, teacher, and parent perceptions of and reactions to testing accommodations. *Assessment for Effective Intervention, 31*(1) [Special issue: Testing Accommodations: Research to Guide Practice], 49-62.

Lesaux, N.K., Pearson, M.R., & Siegel, L.S. (2006). The effects of timed and untimed testing conditions on the reading comprehension performance of adults with reading disabilities. *Reading and Writing, 19*, 21-48.

MacArthur, C.A., & Cavalier, A.R. (2004). Dictation and speech recognition technology as test accommodations. *Exceptional Children*, *71*(1), 43-58.

McKevitt, B.C., & Elliot, S.N. (2003). Effects and perceived consequences of using read aloud and teacher-recommended testing accommodations on a reading achievement test. *The School Psychology Review*, *32*(4), 583-600.

Meloy, L.L., Deville, C., & Frisbie, D. (2002). The effect of a read aloud accommodation on test scores of students with and without a learning disability in reading. *Remedial and Special Education*, *23*(4), 248-255.

Meyen, E., Poggio, J., Seok, S., & Smith, S. (2006). Equity for students with high-incidence disabilities in statewide assessments: A technology-based solution. *Focus on Exceptional Children, 38*(7), 1-8.

National Center for Education Statistics. (2006). *Common Core of Data (CCD): School Years 2004 through 2005.* Washington, DC: U.S. Department of Education, Institute of Education Sciences.

National Research Council (1999). *Uncommon Measures: Equivalence and Linkage among Educational Tests*. Washington, DC: National Academy Press.

O'Neil, T., Sireci, S.G., & Huff, K.L. (2003-2004). Evaluating the consistency of test content across two successive administrations of a state-mandated assessment. *Educational Assessment, 9*, 129–151.

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

Pomplun, M., Frey, S., & Becker, D. (2002). The score equivalence of paper-and-pencil and computerized versions of a speeded test of reading comprehension. *Educational and Psychological Measurement, 62*(2), 337-354.

Russell, M., Higgins, J., Miranda, H., & Hoffmann, T. (2009). *Computer-Based Read Aloud Accommodation: Comparative Effect on Student's Mathematics Test Performance.* Boston College.

SAS Institute Inc. (2009). *SAS User's Guide*. Cary, NC: Author.

Scheuneman, J.D., Camara, W.J., Cascallar, A.S., Wendler, C., & Lawrence, I. (2002). Calculator access, use, and type in relation to performance in the SAT I: Reasoning test in mathematics. *Applied Measurement in Education*, *15*(1), 95-112.

Sireci, S. G. (1998). Gathering and analyzing content validity data. *Educational Assessment, 5*, 299–321.

Sireci, S. G. (2005). Unlabeling the disabled: A perspective on flagging scores from accommodated test administrations. *Educational Researcher, 34*(1), 3-12.

Sireci, S., Li, S., & Scarpati, S., &. (2003). *The effects of test accommodations on test performance: A review of the literature.* (Center for Educational Assessment Research, Rep. No. 485). Amherst, MA: School of Education, University of Massachusetts Amherst.

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

Stretch, L.S. & Osborne, J.W. (2005). Extended test time accommodations: Directions for future research and practice. *Practical Assessment, Research, and Evaluation, 10*(8). Retrieved August 5, 2006 from http://pareonline.net/pdf/v10n8.pdf.

Thompson, S., Blount, A., & Thurlow, M. (2002). *A summary of research on the effects of test accommodations: 1999 through 2001* (Technical Report 34). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Available at http://education.umn.edu/NCEO/OnlinePubs/Technical34.htm.

Thornton, A.E., Reese L.M., Pashley P.J., & Dalessandro S.P. (2002). *Predictive validity of accommodated LSAT scores.* Pennsylvania: Law School Admission Council.

Thurlow, M.L. (2007, April). *Research Impact on State Accommodation Policies for Students with Disabilities.* NCEO, University of Minnesota.

Thurlow, M.L. (2007). State policies and accommodations: Issues and implications. In C. Cahalan-Laitusis & L. Cook (Eds.), *Accommodating students with disabilities on state assessments: What works?* Arlington, VA: Council for Exceptional Children.

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. Available from www.nceo.info/OnlinePubs/Technical26.htm.

Thurlow, M.L., Moen, R., & Altman, J. (2006). *Annual performance reports: 2003-2004 state assessment data* (available on the NCEO Web site at http://www.nceo.info/OnlinePubs/APRsummary2006.pdf). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Tindal, G. (1998). *Models for understanding task comparability in accommodated testing*. A publication for the Council of Chief State School Officers, Washington, DC. Retrieved May 19, 2006, from the World Wide Web: http://education.umn.edu/nceo/OnlinePubs/Accomm/TaskComparability.htm.

Tindal, G. (2002). *Accommodating mathematics testing using a videotaped, read-aloud administration.* Washington, DC: Council of Chief State School Officers.

U.S. Department of Education. (2004, April). *Standards and assessment peer review guidance: Information and examples for meeting the requirements of the No Child Left Behind Act of 2001.* Washington, DC: Office of Elementary and Secondary Education.

U.S. Department of Education. (2006). *Validating assessments for students with disabilities. In Models for large-scale assessment for students with disabilities.* Washington, DC: Office of Elementary and Secondary Education.

Weston, T.J. (2003). *NAEP Validity Studies: The validity of oral accommodation testing.* Washington, DC: National Center for Education Statistics.

Yuan, K.H., Bentler, P.M., & Kano, Y. (1997). On average variables in a confirmatory factor analysis model. *Behaviormetrika*, 24(1), 71-83.

Zenisky, A.L., & Sireci, S.G. (2007). *A summary of the research on the effects of test accommodations: 2005-2006* (Technical Report 47). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

APPENDICES

A. Glossary of Key Terms and Acronyms

B. CTEAG Project Master Schedule

C. Summary of Project Evaluation

D. Item and Test Statistics from State Studies

E. Examples of Additional and Post Hoc Analyses

F. Examples of Statistical Analysis Procedures Used

G. Accommodations Bibliography: Experimental Studies

APPENDIX A
Glossary of Key Terms and Acronyms

AA-MAS – Alternate Assessment based on Modified Achievement Standards. An AA-MAS is designed to enable some students with disabilities to meaningfully participate in the statewide assessment system. A modified academic achievement standard is an expectation of performance that is challenging for eligible students, but may be less difficult than a grade-level academic achievement standard.

ANOVA – analysis of variance. A statistical procedure that determines whether the means of several groups are all equal.

Construct irrelevant variance – sources of variance associated with extraneous features of the test that are not in the intended construct that is being measured.

Differential boost – indicates that SWD receive significantly larger gains in overall test score from an accommodation than SWOD, and although both groups can receive improved test scores by using the accommodation, the gains for the SWD are differentially larger.

Factor analysis – a statistical method used to reduce a large number of variables to a smaller set of highly intercorrelated variables, or "factors."

IEP – Individualized Education Program.

NCLB – No Child Left Behind. The federal legislation for the Elementary and Secondary Education Act, signed into law in 2002.

SWD – students with disabilities.

SWOD – students without disabilities.

Test construct – the concept that a test intends to measure; the knowledge, skills, and abilities related to a specific area (e.g., mathematics, reading comprehension, etc.) that is being measured by a test.

Validity – the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests. Although classical models divided the concept into various "validities," such as content validity, criterion validity, and construct validity, the modern view is that validity is a single unitary construct.

APPENDIX B
CTEAG Project
Master Schedule

Establishing the Validity of Test Accommodations and Score Interpretations for Students with Disabilities: A Collaboration of State-based Research

A Project of the Connecticut State Department of Education and the Council of Chief State School Officers
State Collaboratives on Assessment & Student Standards (SCASS):
> Technical Issues in Large-Scale Assessment (TILSA)
> Assessing Special Education Students (ASES)

Date	Key Project Activities and Tasks	Evaluation[3]
June-Sep. 2007 (before award)	• At TILSA and ASES SCASS meetings, discuss potential research and analysis designs and solicit state participants for accommodations studies • Prepare for CTEAG project kickoff	
Oct.- Dec. 2007	• Connecticut awarded Enhanced Assessment Grant • At combined TILSA and ASES SCASS Meeting in Salt Lake City — provide an overview of CTEAG plans to all state members. Discuss types of studies, details of test administration plans, plans for analyses, and address any questions from states. • Initiate biweekly Project Management Team conference calls to coordinate plans, review questions from states, and solve problems • Survey states for participation in validity studies • Develop "Guidelines for Conducting Validity Studies" document and share with states • Follow-ups with states to confirm participants for validity studies • Discuss plans for the	

[3] Project evaluation activities are continuous throughout life of the project. The evaluator provides feedback at each management team meeting in addition to conducting the activities in the timeline.

Date	Key Project Activities and Tasks	Evaluation
	development of split-half forms for accommodation studies in each state • Determine plans for identification of student samples in each state • Submit proposal on CTEAG project for presentation at 2009 CCSSO National Conference on Student Assessment (NCSA)	
Jan.- Mar. 2008	• Refine timeline and design • Hold 1st meeting of CTEAG (in Atlanta) • Make presentations at combined TILSA and ASES SCASS Meeting — update states on plans, status of project, and progress to date • Refine plans for the development of split-half forms for accommodation studies and identification of student samples in each state • States finalize split-half test forms and work with vendors in preparation for Spring ’08 administrations • States select samples for studies (Spring ’08 administrations)	Attend combined meeting of TILSA and ASES SCASS. Review proposal in light of implementation plans.
Apr.-June 2008	• Conduct interim formative evaluation of project • Begin collecting published and unpublished research on test accommodations validity research • CT sends forms to schools and conducts test administrations (Spring ’08 administration) • At TILSA and ASES SCASS Meetings, review information and progress to date from spring administrations and suggest refinements to analysis plan • Make presentation on CTEAG project at CCSSO NCSA in Orlando	Survey state team leaders. Prepare report of survey with recommendations and discuss with management team. Present findings to TILSA and ASES (Doug, Sandra).

Date	Key Project Activities and Tasks	Evaluation
July-Sep. 2008	• CT's vendor scores test booklets and sends data files to project analyst (Spring '08 administration) • Begin initial statistical analyses (item analysis, DIF, score distributions, etc.) • Develop draft outline for guidebook • States finalize test forms and give to vendors in preparation for Fall '08 administrations • States select samples for studies (Fall '08 administrations) • Submit proposals for symposia at 2009 American Educational Research Association (AERA) meeting • Prepare annual project progress report for ED	
Oct.-Dec. 2008	• Hold 2nd meeting of CTEAG (in San Diego) • Make presentations to TILSA SCASS on results to date, analysis plans, draft outline for validity studies guidebook, and evaluation feedback • Collect information for research summary (literature review) and preliminary guidebook • NV ships forms to schools and conducts test administrations (Fall '08 administration) • Submit proposal on CTEAG project for presentation at 2009 CCSSO NCSA	Attend consortium meeting and make suggestions as needed.

Date	Key Project Activities and Tasks	Evaluation
Jan.-Mar. 2009	• NV's vendor scores test booklets and sends data files to project analyst (Fall '08 administration) • Conduct statistical analyses (item analysis, DIF, score distributions, etc.) of Fall '08 administrations in NV • Hold 3rd meeting of CTEAG (in Orlando) • Make presentations to ASES SCASS on results to date and get input on revised draft outline for validity studies guidebook • States finalize test forms and give to vendors in preparation for Spring '09 administrations • States select samples for studies (Spring '09 administrations) • MI conducts test administration for validity study • Draft summary of accommodations validity research	Attend consortium meeting and make suggestions as needed.
Apr.-June 2009	• KY and CT conduct test administrations for validity studies (Spring '09 administration) • MI scores data and sends data file to project analyst • Hold 4th meeting of CTEAG (in L.A.) • Make presentations at combined TILSA and ASES SCASS Meeting — provide update on status of project, share findings from state studies, discuss challenges in conducting studies and lessons learned, review draft info in guidebook • Make presentation on preliminary results from state's validity studies at CCSSO NCSA Pre-Session in L.A.	

Date	Key Project Activities and Tasks	Evaluation
July-Sept. 2009	• States' (KY and CT) vendors score test booklets and send data files to project analyst (for Spring '09 administration) • Conduct statistical analyses (item analysis, DIF, score distributions, etc.) • Prepare draft summaries of analyses and findings • Begin writing guidebook with actual data and results from studies included • Begin drafting technical report and project database contents • Submit proposals for symposia at 2010 AERA or National Council on Measurement in Education (NCME) meetings • Prepare annual project progress report for ED	Conduct second interim formative evaluation of project, focusing on states that decided not to participate in the study after initially agreeing considering participation: What conditions led to changing their participation decision? What can be done to make it more feasible for them to participate in such studies? What challenges need to be overcome? The goal is to collect information and advice that will help states to conduct needed studies.
Oct.-Dec. 2009	• Complete draft of guidebook and share draft reports with state members and other interested parties for review and feedback • Hold 5th meeting of CTEAG (in St. Louis) • Make presentations at combined TILSA and ASES SCASS Meeting on findings from state studies and get feedback on draft reports and outlines • Revise information in Guidebook and make final revisions based on feedback from states • Submit proposal on CTEAG project for presentation at 2010 CCSSO National Conference on Student Assessment • Conduct formative evaluation of project • Make presentation on CTEAG project at 2009 NERA symposium	Attend consortium meeting and present second interim evaluation findings. Revise as needed based on feedback. Final formative evaluation – project states: What worked and what didn't in terms of implementing the study (e.g., start-up, communication, interdepartmental support)? The goal is to add to the 2nd evaluation results in terms of what works and what needs to be overcome. Discuss plans for final project evaluation with PMT – materials needed, outline, etc.

Date	Key Project Activities and Tasks	Evaluation
Jan.-Mar. 2010	• Disseminate Validity Studies Guidebook • Hold 6th meeting of CTEAG (in New Orleans) • Make presentations at combined TILSA and ASES SCASS Meeting on project and get feedback on plans for next reports and data • Complete drafts of technical report and database and share with state members and other interested parties for review and feedback • Revise reports and make final revisions to products based on feedback from states • Prepare for dissemination workshop • Conduct final evaluation of project	Final project evaluation – success of studies (design, interpretability of results), state goals, generalizability of results, dissemination, management structure.
Apr.-Jun. 2010	• Finalize technical report and database for publication • Hold 7th and final meeting of CTEAG (in Detroit) • Combined TILSA and ASES SCASS Meeting and special workshop to disseminate products, highlight findings, and instruct states on how to best use the information from this project • Make presentation on findings from validity studies at CCSSO NCSA Session in Detroit • Post all materials on CSDE and CCSSO websites • Prepare final project report for ED	

Notes:

Revised timeline includes 12-month no-cost extension granted by ED that extends the project through March 2010 and 2nd extension that took project to June 2010.

Project Management Team meets by teleconference biweekly throughout life of project

CTEAG state members meet face-to-face semi-annually

CTEAG meetings held in conjunction with combined TILSA and ASES SCASS meetings

APPENDIX C
Summary of Project Evaluation

Evaluation of the Project, Establishing the Validity of Test Accommodations and Score Interpretations for Students with Disabilities

Phoebe C. Winter, Project Evaluator

Pacific Metrics

March 23, 2010

The project, *Establishing the Validity of Test Accommodations and Score Interpretations for Students with Disabilities,* was funded by an Enhanced Assessment Grant awarded to a consortium of states led by the Connecticut State Department of Education (CSDE) and facilitated by the Council of Chief State School Officers (CCSSO).[4] The goal of the project was to work collaboratively with a group of states to conduct validation studies of a number of assessment accommodations commonly offered to students with disabilities. Each state selected the accommodation, subject area(s), and grade level(s) to study, with guidance from the project leaders. Although accommodations and assessments varied, states were asked to use a common design in their studies. The planned design was based on administering two mini-tests constructed to be parallel to each other in content and difficulty to students with disabilities (SWD) who required the accommodation studied and a demographically similar group of students without disabilities (SWOD). Each student would receive one short (approximately half-length) test with accommodations and one short test without accommodation. The design is characterized as follows in the project *Technical Report* (Olson, Dirir and CTEAG [Connecticut Enhanced Assessment Grant] Project Management Team, 2010):

Accommodations		**Students**	
		Students with Disabilities	Students without Disabilities
	Without	Split half A Split half B	Split half B Split half A
	With	Split half B Split half A	Split half A Split half B

Note: Split half refers to the short test

[4] The project was funded by grant S368A070009 from the U.S. Department of Education.

Project Structure

Four state departments of education, Connecticut, Kentucky, Michigan and Nevada, participated as research sites, and state members of two of CCSSO's State Collaborative on Assessment and Student Standards (SCASS) groups – Technical Issues in Large-Scale Assessment (TILSA) and Assessing Special Education Students (ASES) – served as advisors in the design stage of the project and reacted to in-progress research reports. CCSSO served as the facilitator for the states involved in the project and as the primary contractor for the project, channeling funds for carrying out the research to participating states and convening participating states, the project management team, and the TILSA and ASES SCASS, with Doug Rindone, the TILSA SCASS advisor, serving as CCSSO's primary contact. A project management team consisting of both members internal to the study and external members with expertise in educational measurement and in assessing students with disabilities provided ongoing advice to the project. John Olson, a consultant to CCSSO, served as principal investigator and project coordinator. In addition to working closely with the states in coordinating their data collection and analysis plans, reporting to various audiences, and authoring or co-authoring project reports, Olson led the project management team in bi-monthly conference calls and periodic in-person meetings. Data analysis was carried out by Mohamed Dirir, a psychometrician on the CSDE staff. Appendix A contains a list of the project management team members and state project leaders.

An important factor in project planning and implementation across four states, multiple advisors, and the primary contractor was a defined communication and information flow that was flexible enough to allow for immediate reaction to project needs. CCSSO coordinated all fiscal activities, working directly with state contacts. Olson and Dirir coordinated research and analysis activities, working with state researchers in designing studies and data collection and they, along with Rindone, were available to provide in-progress advice to states about the research and any attendant policy issues. Olson coordinated the project management team and conveyed its advice to participating states.

Overview of Evaluation Activities

The project evaluation was designed to be primarily formative, with the evaluator making recommendations to enhance ongoing project work based on participation in bi-monthly project management team meetings, reviews of project planning documents, attendance at project meetings, discussions with project state staff and SCASS TILSA and ASES members, and results of surveys of project state staff and other SCASS state staff members. While much of the evaluation process was, by its nature, not formally documented, several interim evaluation reports are contained in Appendix B:

1. **Evaluator Notes and Comments/Recommendations, February 14, 2008**
   Initial notes from review of the overall project design

2. **Connecticut EAG State Team Leader Survey, Spring 2008**
   Results of a survey designed to gather information that would help in the next phases of the grant

3. **Results of ASES/TILSA Member Survey on Collaborative Research, January, 2010**
   Results of a survey about conditions that would facilitate and encourage state participation in cross-state collaborative research projects like the CTEAG

4. **Implementing Cross-State Research: Recommendations from Project States, March, 2010**
   Results of a series of discussions with project states about implementation – what worked well and what could be improved for future studies

Evaluation Results: Lessons for Collaborative Research

The results of evaluation activities are synthesized below according to key components of the project: planning and design, implementation, and dissemination. This synthesis is based on aspects of the study that contributed to its success and on results of discussions with and surveys of participating and non-participating state members of TILSA and ASES. The discussion highlights factors that are relevant to future implementation of cross-state research; for each component, recommended practices are followed by a description of how the CTEAG addressed the component.

Planning and Design

State staff involvement in planning and designing the study and the choice of topic contribute to the strength of a cross-state collaborative study and states' willingness to participate in the study. The topic should be relevant to a current state need, relate to federal mandates and/or be forward looking, and the results should affect the way states will implement their assessment systems.

The purpose and potential utility of outcomes of the research need to be clearly stated. This will help states in deciding whether to join in the study and assist state staff in communicating with policymakers who have the final say in whether the state participates. In determining whether to participate in the study, states should consider their student populations and assessment systems and policies to evaluate whether participation will be practical and beneficial. The quality of external researchers and advisors and the degree to which appropriate state staff will have input into the design and implementation are important considerations. If available state staff members do not have the requisite research or psychometric background to assist in the planning and design, the state may decide against participating in the study.

Another critical factor is resource allocation – states are more likely to participate in studies that receive full outside funding and require little state-level staff time to implement the research; most state departments of education do not have staff members to devote even part-time to research projects, regardless of the project's importance. In the planning phase, practical issues should be laid out insofar as possible and the roles of responsibilities of all research partners should be spelled out so that states have a realistic idea of what participation will entail. An option is for the project to provide funding for additional state staff time or consultants to work at the state DOE if necessary.

Issues related to state policy and practice affect whether states can participate in collaborative research. For example, some states have policies prohibiting the release of student-level assessment information for all students or for students in special education or have requirements regarding preparing data for release to researchers that take more time than the state has to devote to the task. In some cases, states do not collect the data required to implement the research design (e.g., disability category); either funding must be sufficient to collect that information or the state must be excluded from participation. The timing of the research should take into account each state's assessment schedule, and flexibility in the timing (e.g., the ability to conduct the study in either the fall or spring or the provision of a two-month window for data collection) should be incorporated to the extent possible. If possible, studies should incorporate designs in which all students in a classroom are included; the logistics of student-level sampling provides difficulties in research coordinated at the state level.

A strong research design that is well-described and flexible enough that it can be implemented across state contexts is critical. Before implementation, the design should be reviewed both by researchers who can critique the study design and identify potential threats to the generalizability and validity of findings and by state participants so that adjustments can be made as needed. Participants need to have clear expectations about inferences, actions, and decisions that can be made based on research outcomes, and consider the implications of possible research outcomes in advance. If the research involves field-based implementation, the study should include processes for monitoring and evaluating the implementation to assist in interpreting the results.

The CTEAG proposal development was coordinated by CCSSO, and potential state participants had the opportunity to provide input to project planning and design components of the proposal. The topic was selected to provide information that states could use to improve their assessment systems and meet the assessment requirements of Title I of ESEA. Practical considerations were part of proposal development discussions and of presentations to TILSA and ASES in the recruitment phase of the study. Project funding covered most costs, including the development of materials, analysis of results, and writing of interim and final reports; SDOE staff were expected to contribute their time and expertise to planning activities, recruiting schools, developing forms, overseeing study logistics implementation, and interpreting results. Some SDOEs did not participate in the study because they did not have the staff time to contribute to the project. In other cases, staff could not persuade their leadership of the value of participating, given competing priorities, and some states did not collect the demographic data needed to implement the study.

The design of the study was reviewed by state participants and advisors as well as the project management team, and some changes were made to improve the potential utility of the results. The project reports include recommendations based on the study and study limitations, but the potential outcomes and interpretations were not an explicit part of planning discussions. Three of the four states (four of the five studies) used the planned design, and it was modified to meet the needs of the fourth state in the study. Study timelines were flexible enough to allow for implementation at various times in the participating states. There was no monitoring plan included in the design.

Implementation

Consistent and public support from educational leaders in the state can help in the implementation of the research, particularly in soliciting volunteer schools and districts for the study; incentives in addition to money, such as professional development for participating schools and districts or access to data for classroom use, can also help in soliciting volunteers. Teachers and administrators should see the value of the research, and the anticipated improvements that could be made in the state assessment program based on the research should be explained.

Policymaker support should extend to providing time and resources for state staff to meet commitments related to the research project. Clear, straightforward communication with policymakers should be supported by the project, with sample communication strategies and materials developed by project partners.

Other conditions during implementation that can contribute to the smooth running of cross-state projects include having a specific but flexible timeline and well-structured communication channels between project researchers and state staff. Research activities should be scheduled with the understanding that there will likely be competing priorities that affect state time commitments, and time should be built in so that activities can be rescheduled as necessary. A practical aspect of implementation that can affect the project is logistical support on the part of the primary contractor to the states. Logistics involving delivery of funds, convening project meetings, and facilitating communication among project partners must be well-planned with defined procedures and channels for states to use in arranging for financing and similar issues.

As noted earlier, staff members at state departments of education have little spare time to work on tasks that are not directly related to running state programs, even when they have the support of educational leaders. Cross-state projects should be organized to use state staff members' time efficiently, with clear agendas for meetings, goal-driven communications, and the use of technology such as web-based meeting software and teleconferencing. In-person meetings involving state staff should either be important enough to justify a stand-alone meeting or held in conjunction with other meetings the staff members would likely attend. Having a consistent point of contact in each state can maximize the efficient use of state staff time, eliminating the need to have new staff members spend time learning the project particulars and allowing communication to flow more effectively between project and state staff.

State-level staff or consultants need to be available to work with schools as they implement the study. A clear communication procedure for schools to use as they have questions is critical. This becomes more important the more the study procedures deviate from typical test administration procedures. Because of limitations on state staff members' time, when needed, external consultants who know the state's assessment practices and policies should be employed to work with schools in implementing the study and to monitor the implementation.

The CTEAG was organized around a general timeline, with specific activities scheduled based on state timelines. Communication with state participants was efficient, with researchers available to discuss general design and administration issues with participating states as a group and specific issues with each state individually. In-person meetings of participating states were held in conjunction with ASES and TILSA meetings so that no additional travel was required. Project management team meetings were held via bi-monthly teleconference and each meeting was structured around an agenda. Funds were not available to hire consultants to work within each state; therefore state staff had to recruit schools, develop materials, and manage all logistics of the study. Most states worked with their assessment vendor(s) to get the work done. Stipends were paid by the project to help states defray the additional costs of their vendors. Project staff assisted in state-level forms development and data collection design, and all analyses were conducted by project staff with resources dedicated to the project.

The project researchers were flexible and responsive during the implementation phase. For example, because strict comparability of the half-forms used was not a feature of the study, researchers developed methods for investigating the forms' difficulty and correcting for differences as far as possible. When results of the first CT study revealed some issues that could be included in a follow-up study, the CTDOE implemented a second study. When one state had a research question that required a different design, the researchers worked with the state to fit the study into the project.

Dissemination

Ongoing dissemination of interim results can be useful both to states that are actively participating in the research and to other states. This type of in-progress reporting is likely to highlight planning and implementation issues and solutions that may not be discussed in the final project report. If a goal of the project is practical implementation or application of research results, written reports should be aimed at a variety of audiences, including state-level staff members who implement effected programs, educational leaders, and researchers. Presentations at meetings and conferences that the target audiences are likely to attend can spread awareness of the research. These presentations should include information about where to easily obtain more information about the project; a project website or webpage should be established.

Reports and presentations should stress the implications of the research and carefully describe the limitations of the findings. Suggestions for follow-up studies and practical ways that states can look at their extant data are helpful. If the data collected for the project can be made available for secondary analyses, this should be done.

Throughout the course of the project, progress reports, preliminary results, and final results have been disseminated to various important audiences. Researchers have made periodic presentations to the TILSA and ASES SCASS groups, both to provide information and solicit advice. Presentations were made at the annual National Conference on Student Assessment (NCSA) in 2008 and 2009, and one is scheduled for 2010, and a presentation was made at the 2009 annual meeting of the Northeastern Educational Research Association, reaching audiences of researchers and practitioners.

The CTEAG project is producing a variety of written reports, aimed at audiences with different purposes for reading about the project (e.g., for developing accommodations policy, for determining research needs) and different levels of technical and substantive expertise. For example, the *Guidebook for Studies of the Validity of Test Results for Test Accommodations* (Olson, and CTEAG PMT, 2010) is aimed at "all people in state education agencies, local districts, and special education, as well as policymakers and other key stakeholders" (p. 5) and is written in non-technical terms. The *Technical Report* (Olson et al., 2010) is aimed at the research community and includes details about study design and results more appropriate for that group. *Lessons Learned from State Validity Studies* (Andrada and Amenta, 2010) was developed to highlight practical considerations in designing and implementing accommodations research, with a focus on computer-based readers. The *Technical Report, Lessons Learned report, and Project Database, which can be used by researchers interested in replicating the research or conducting additional analyses,* will be distributed at the NCSA in June 2010, and the Guidebook has been distributed to state DOEs and, in some states, to schools participating in the research study in early 2010. In addition, all products will be put as PDFs on a project website and made available, for free, to anyone who wants to download them.

APPENDIX C.1
CTEAG Project Participants

CTEAG Project Management Team

Barbara Beaudin, Connecticut DOE and Project Director

Mohamed Dirir, Connecticut DOE and Project Analyst

Joe Amenta, Connecticut DOE

Gil Andrada, Connecticut DOE

Janet Stuck, Connecticut DOE

John Olson, Principal Investigator and Project Coordinator

Doug Rindone, TILSA Advisor

Sandra Warren, ASES Advisor

Martha Thurlow, Advisor/Partner

Eileen Ahearn, Advisor/Partner

Phoebe Winter, Project Evaluator

Adam Petermann, CCSSO Liaison

State Validity Study Project Leaders

Joe Amenta, Connecticut

Gil Andrada, Connecticut

Mohamed Dirir, Connecticut

Ken Draut, Kentucky

Roger Ervin, Kentucky

Joseph Martineau, Michigan

Steve Viger, Michigan

Adam Wyse, Michigan

Carol Crothers, Nevada

Cindy Sharp, Nevada

Andrew Parr, Nevada

APPENDIX C.2
Evaluation Reports

CTEAG -- Evaluator Notes and Comments/Recommendations
Submitted by Phoebe Winter, February 14, 2008

Background

The CTEAG project staff met with representatives of the six states that have expressed interest in or committed to joining the study. During the meeting, state representatives and project staff discussed various issues that arose after staff presentations (see power points and handouts provided at the meeting). This set of notes contains a list of the critical issues (in the evaluator's judgment) that arose, brief notes on the discussion, and evaluator comments/recommendations. The issues and comments from the state meeting are followed by general issues the evaluator would like to see discussed and addressed (either by dismissing the issue or by amendments/clarifications to procedures).

Most of these comments and recommendations have been discussed with members of the project team, during Management Team meetings or in individual discussions. The comments focus on potential threats to validity of the results and, where possible, provide some approaches to reducing these threats.

Issues Arising in State and PMT Meetings with Comments/Recommendations

1. How to select an SWOD group that matches the SWD group on demographics – This was discussed briefly at the state meeting, with suggestions such as using SWODs from the same school as SWDs. The idea is that the samples should match, as a whole, as much as possible, on as many relevant demographic variables as possible. Some of these variables may be collected only at the school (not student) level.

   Comment – Research reports will include a description of matching variables and procedures, as well as the demographic characteristics of the groups. The research team should think about if/how study results may be affected by the degree and nature of matching, and they should discuss the types of caveats that should be attached to the interpretation of results.

2. How to select SWDs for the study if not all eligible students are used – For logistical reasons, it may be difficult to obtain SWDs from outlying schools or other types of schools; this may result in the SWD sample being not representative of the state's SWDs eligible for the accommodation. It was suggested that states may wish to use schools with relatively large concentrations of SWDs.

   Comment – Again, it's important to consider how sample selection may affect the interpretation of study results. For example, if there is an interaction between an accommodation and a demographic characteristic (e.g., rural vs. urban, high vs. low SES) and students with the demographic characteristic are over- or under-represented, the appropriate interpretation of results could be affected. Care should be taken to consider such issues when developing the

SWD sampling scheme and/or analyses should be included that can rule out such threats to validity.

3. How to ensure students use the accommodation – For this design, it's important that the students use the accommodations, e.g., that both SWD and SWOD attend to the text reader or that both groups use the calculator provided. As noted in the guidelines, both groups should be trained in using the accommodation if it is a new accommodation, and SWOD should be trained in the accommodation regardless of its novelty in the SWD population. Procedures for training will be reviewed and approved by the project team.

   Comment – Meeting this criterion is critical for interpretation of the research results. CT has developed some plans for training students in the accommodation and for encouraging them to use the accommodation. These should be shared as models with other states. The group should consider asking teachers to observe whether the student appeared to be using the accommodation, perhaps with a frequency scale (often, sometimes, etc.) so that results can be interpreted in light of use, if necessary.

4. How to counterbalance – The issue of counterbalancing arose briefly at the meeting. The order of administration for both accommodations and form (half-test) will be counterbalanced. [Please check accuracy of this – if not accurate, we need to discuss.]

   Comment – Counterbalancing procedures should be clearly spelled out for states.

5. Use of a single accommodation rather than a package of accommodations – The guidelines suggest providing only the accommodation under study to students (with necessary supports for the accommodation such as extra time or individual administration). During the discussion with states, it was noted that the effect of the target (studied) accommodation might be dampened if the student does not receive his/her needed support. For example, if a student is typically given a read-aloud accommodation but is given a calculator for purposes of the study without an accompanying read-aloud, the calculator may have no (or even a negative) effect on the score. The student may not be able to read the test well enough to get to the stage in problem-solving where a calculator might be needed; the student may spend so much cognitive capital on decoding the item that performance is affected, despite the calculator; the student may be frustrated by lack of understanding the text, affecting motivation to do well (this has implications for non-accommodated conditions for SWDs as well); etc.

   Comment – The study should include packages of accommodations or be limited to students who need only the target (studied) accommodation. Since the latter condition is difficult to satisfy, I recommend the former. This issue will be discussed with the management team on February 20, 2008.

6. Parallel forms – Study forms (half or short versions of a regular test) will be designed to be as parallel as possible in content and difficulty. The test scores will not be on the same scales.

   Comment – Procedures for comparing scores on the two forms need to be specified. For example, will scores be standardized in some way?

7. Motivation – This issue arose during the 2/13/08 MT call – will motivation differ by condition or by group or by a combination of condition and group? CT has plans to encourage students to do their best (see also 3 above). There was discussion of collecting information about student motivation at the end of each session.

Additional Issues

8. Interaction hypothesis – The study has been designed to test what has been called the "interaction hypothesis": if the accommodation improves the scores for SWDs more than it improves scores for SWODs, then there is support that the accommodation is a "valid" one. Additional analyses that will add to the evaluation of an accommodation's validity are planned, including factor analyses, DIF analyses, and reviews of item characteristics in accommodated and non-accommodated conditions within and across groups.

   Comment – The interaction hypothesis as a way to study the effects of accommodation has been losing support in the field (e.g., Sireci, Li, and Scarpati (n.d), The Effects of Test Accommodation on Test Performance: A Review of the Literature – see especially top of page 48 and pages 60-63; National Academies of Sciences (2004), Keeping Score for All: The Effects of Inclusion and Accommodation Policies on Large-Scale Educational Assessments – see especially pages 87-88, page 96, and page 101).

   One reason for this is that there can be a number of interpretations of the same result from studies relying on demonstrating an interaction effect. For example, it could be that the accommodation **appropriately** removes a construct-irrelevant access barrier for some SWDs and removes the same barrier for some SWODs. The results might show that both groups, on average, increased their scores on the accommodated version of the test, with the inference, according to the interaction hypothesis, that the accommodation is not valid (i.e., it provides an unfair advantage to SWDs). In this case, the accommodation is providing for better (more valid) scores for students in both groups, rather than reducing the validity of scores for SWDs (by providing an unfair advantage). A real-life example could be when IEP teams allow read-alouds for SWDs who do not require them; the read-aloud may benefit part of the group of SWDs (the part that needs read-alouds) and benefit SWODs who are poor readers.

   Another example – If SWODs do not use the accommodation, their scores will be the same under accommodated and unaccommodated conditions (other things being equal). If SWD scores increase from unaccommodated to accommodated, there will be an interaction that can

be interpreted as supporting the validity of the accommodation. However, the SWODs effectively did not receive the accommodated.

The National Academies report puts it this way: "For the most part, existing research has investigated the effects of accommodations on test performance but is not informative about *the validity of inferences* [emphasis added] based on scores from accommodated administrations." (page 101)

Recommendation – Discuss alterations and/or additions to the design that will allow for a stronger focus on valid interpretations of results. The National Academies' book has a chapter on ways to articulate validity arguments that may be useful.

9. Other potential threats to validity of results

Comments – The researchers may wish to review the design in light of Shadish, Cook, and Campbell (2002), Experimental and Quasi-Experimental Designs for Generalized Causal Inference, to identify potential threats and control for them as much as possible. Doing so will also allow the researchers to provide caveats for the interpretation of results.

Connecticut EAG State Team Leader Survey, Spring 2008
Results (4 of 5 states responding), May 25, 2008

Phoebe Winter

Team leaders from the five states participating in the Enhanced Assessment Grant, Establishing the Validity of Test Accommodations and Score Interpretations for Special Education Students: A Collaboration of State-based Research, were asked to complete an online survey in early May, 2008. The purpose of the survey was to gather information that would help in the next phases of the grant. As of May 25, 2008, four of the five states had responded. Two responses were received for some questions for one of these states; it is not clear whether this is because of two respondents or whether the same person responded twice.

Results

The results are described for each question posed in the survey.

1. What information do you expect your state to get from the project and how will you use that information?

   All responses were specific to the accommodations the state is studying. Each state hoped to gain information related to the validity of the accommodation. In addition the states had specific expectations related to validity: whether the accommodation might disadvantage some students; how the accommodation affects student scores; whether the accommodation provides more accurate scores; and the degree to which the accommodation provides scores equivalent to the regular form.

2. Have your goals changed since the beginning of the project? If so, how have they changed?

   Three states responded "No" and respondent was new to the project and was not sure.

<u>Implementation</u>

These results are listed by state, since a summary would not be interpretable.

3. At what stage are you in implementing the study in your state?

   Please check all that apply and include a brief description in the space provided in the next item.

Response Options	State				
	1	2	3	4	4a
Deciding on accommodation to use		X			
Finalizing research plans and procedures		X		X	
Designing sample selection procedures				X	
Selecting samples					
Soliciting schools to participate					
Selecting test form(s) for use in the study					
Developing split forms	X				
Other activities – please specify		X	X		X

Other
State 2: Working with vendor on price estimates.
State 3: Administering the test.
State 4a: In between deciding and finalizing research plans and procedures.

4. Do you have a shareable schedule or time line for implementing the project? If yes, please attach the schedule, indicating completed tasks.

   No state had a project schedule or timeline to share.

   If no: What project activities do you plan to undertake between now and Fall, 2008?

   State 1: Select items and designing test forms; finalize research plans and procedures; select the sample.
   State 2: Publicize the study; finalize agreements with the vendor; explore sample requirements.
   State 3: Conduct preliminary analysis.
   State 4: Select schools; prepare forms.

5. What resources and information do you need from project management to continue with implementation?

   State 1: Assistance with the research plan and sampling.
   State 3: Money to provide schools with a text reader.

<u>Communication</u>

6. In October, 2008, in conjunction with the ASES and TILSA SCASS meetings, we will have the second in-person meeting of the entire CTEAG team. What suggestions do you have for the structure and content of those meetings?

- Two states suggested that we hear a status report from each state. This report should include a discussion of challenges and successes.
- A three-part meeting: large group (status report), small group (similar accommodations), and individual states with the management team.
- A clear agenda specifying the practical information to be provided and the goals of the meeting.
- Hold the meeting with ASES and TILSA so that participants can attend as part of their SCASS meeting. There should not be an empty day for participants. If there is a dinner meeting, the SCASS meeting should start the day after.

7. Between in-person meetings, we will hold web-based meetings of the project team. What suggestions do you have for the structure and content of those meetings?

    - These should be set up individually with states since the needs and timelines are so diverse.
    - Share experiences and lessons learned, roadblocks; leave time for feedback during each call.
    - Keep to the agenda and timelines.

<u>General Suggestions</u>

8. What features of the project need improvement? How should we improve them (please be specific)?

    - Getting money to the states.
    - More focused, agenda-based conference calls (less off-topic talk).

9. What lessons have you learned to date that will help move the project forward?

    - Being practical is important is this type of research – some accommodations may not be able to be studied this way because of practical constraints (e.g., administration time).
    - Planning and establishing deadlines is important.
    - Clearly identifying the purpose of the research in each state is important.

10. Other suggestions and comments

    - Feedback from John and Mohamed has been helpful and thoughtful.

Connecticut Enhanced Assessment Grant
Results of ASES/TILSA Member Survey on Collaborative Research

Phoebe C. Winter, January 8, 2010

The evaluation activities for the Connecticut Enhanced Assessment Grant (CT EAG) are formative and designed to be used to improve the project as it moves forward. Reflecting this future-oriented approach, in September, 2009, a survey about conditions that would facilitate and encourage state participation in cross-state collaborative research projects like the CT EAG was administered to ASES and TILSA SCASS members. The survey questions were based on discussions with participating state members and group discussions with the two SCASS projects.

Administration

With the assistance of CCSSO, the survey was administered via Survey Monkey, a web-based survey system. All state members of the ASES and TILSA SCASS projects were asked to complete the survey. The TILSA and SCASS coordinators sent an email with the following instructions to their members via their SCASS listservs:

> Below is a link to a web-based survey for SCASS ASES and TILSA members. The purpose of this survey is to determine what conditions support state education departments' participating in cross-state collaborative research projects. The survey is part of an enhanced assessment grant project awarded by the US Department of Education to Connecticut and a consortium of SCASS ASES and TILSA states, facilitated by CCSSO (the CT EAG). Results from the survey will be used in designing cross-state projects and in structuring the final report for the CT EAG and will be reported to SCASS members.
>
> We hope to receive responses from all ASES and TILSA state members; please respond even if your state is a member of both SCASS projects. **Please respond by September 25, 2009.**[5]

The survey consisted of an opening question about participation in cross-state research followed by 28 Likert items about specific features of cross-state collaborative research. Respondents rated each feature in terms of how they thought it would affect their State Departments of Education's (SDE) decisions regarding participation in cross-state research. The Likert items were followed by four open-ended questions that probed states further about their possible participation in cross-state research. These questions and Likert items and responses can be found in the Appendix.

[5] The survey plan included a follow-up email to non-respondents to increase the response rate; this was not possible, however, due to a technical difficulty.

Participants

At the time of the survey, the ASES SCASS had 30 member states and the TILSA SCASS had 24 member states, with 38 member states across the two SCASS projects. There were 22 responses to the survey from 16 unique states, with 11 responses from ASES members and 10 from TILSA members; one response was a group response from three people whose membership covered both SCASS projects. Of these 22 responses, 18 were complete; two responses were for the first question only, one was through question 2.6, and one response was through question 2.16.

Most respondents (78%) said that their states had participated in cross-state research as a research site (14) or as a project advisor only (3). One respondent said that his/her state had not participated in cross-state research, and four respondents did not know.

Likert Item Results

Participants responded to 28 questions about how specific features of a cross-state research project would affect their state departments of education's decision regarding participation on a five-point scale ranging from "less likely to participate" to "more likely to participate."

SDE staff involvement

In general, respondents indicated that their SDEs were somewhat more or more likely to participate if state staff were involved in planning, design, implementation, and analysis activities. This response was strongest for planning and design.

Funding

All respondents indicated that their SDEs were somewhat more or more likely to participate if outside funding covered all study costs, and almost all (90%) indicated that their SDEs were somewhat less or less likely to participate in studies that were completely funded by participating SDEs. Responses were split across more and less likely categories for research funded by a mix of SDE and outside funding.

Research topics

Results indicate that SDEs are more likely to participate in research with topics that

- Are clearly related to federal requirements (ESEA and IDEA)
- Address current needs
- Are forward looking
- Are proposed based on a summary of existing research

Respondents were evenly divided about whether their SDEs would be more likely to participate in research about topics that were selected by participants but not necessarily tied to federal requirements.

Research design

Results indicate that respondents are flexible in terms of the structure of the research design, with most respondents choosing "more likely" or "likely" in response to conditions in which each state conducted a replication of a single study or varied components of the study. Using the same materials across states and focusing on original research also received a large majority of responses on the likely end of the scale.

Administration and sampling conditions

Respondents overwhelmingly favored studies which could be conducted in conjunction with a statewide administration of the test (89% more likely and likely) over those that required a stand-alone administration (11%). Questions about sampling levels garnered the most "no effect" responses, with respondents ambivalent about whether the student or school was the unit of sampling and whether the sample was representative or random. Studies that include all students in selected classrooms (56% more likely and likely) appear more practical than those that include selected students (11%).

Research support

Having the project provide resources was seen as increasing the likelihood that an SDE would participate in cross-state research, with a majority of responses of more likely and likely to

- Having the research carries out by personnel external to the SDE (77%)
- Providing monetary incentives to schools (72%)
- Providing professional development to participating teachers (94%)

Relevance

Respondents indicated that research with results that can be used to modify or provide support for SDE testing programs (94%) and for which schools can see a direct effect on the testing program (78%) was more likely or likely to foster SDE participation.

Open-Ended Question Results

Four open-ended questions about participation in cross-state research were included in the survey.

*Please provide any **additional** ideas you have about the best possible (but realistic) conditions that would allow your SDE to participate in a collaborative cross-state research project.*

Ten participants wrote responses directly connected to the question:

Five responses reiterated the need for monetary (full funding) support, with 3 of these adding logistic support.

Four responses noted the need to have clear expectations of SDE staff involved in terms of time commitment and level and nature of involvement (at the same time indicating that current conditions made it difficult to allocate staff time).

Three responses included specific logistics and conditions for the research:

- Flexible timeline
- Teacher and administrator buy-in (2)
- Well-structured communication between researchers and state
- Researchers make all contact with the schools

Two responses suggested that studies be related to the common core standards, with one of the responses adding that common assessments and technology (especially hand-held technology) be incorporated in studies.

One response noted that participation of schools and districts depended on the nature of the study and that the decision to participate typically included many factors, including policy factors.

If your state was recruited to participate in the CT EAG project, but it decided not to join, what were the main reasons for this decision?

<u>Four participants wrote responses directly connected to the question:</u>

Two respondents indicated that they did not have the staff time to commit to the project.

One respondent said that their state did not collect information about the accommodations needed for the study.

One respondent noted that the initial study design seemed weak (but gave no further explanation why he/she thought this).

If your state did participate in the CT EAG, what were the main reasons why it did?

<u>No respondents were from states participating in the study.</u>

What would facilitate your state agreeing to take a lead role in a future EAG, such as the role Connecticut has for the current project?

Fifteen participants wrote responses directly connected to the question, although some responses explained why they *couldn't* take a lead role rather than what would facilitate their taking such a role:

Seven responses specified that the research would have to meet a current state need.

Six responses indicated that financial support would be necessary.

Five responses mentioned having the capacity for state staff involvement, with most noting that their SDE staff time is very limited, precluding their serving as lead state in some cases; one respondent suggested allowing the state to hire outside people to manage the grant.

Two respondents mentioned clear, straightforward communication with/for state educational leaders to explain the benefits and encourage participation.

One respondent mentioned stronger support from CCSSO in logistics and reimbursements.

One response specified research on the common core standards as a condition.

Collaborative Research: ASES/TILSA Member Survey

Questions and Selected-Response Results

State or Organization: 22 responses from 16 unique states; 1 response was a group response. 18 complete responses.
SCASS Affiliation: 11 ASES members, 10 TILSA members, 1 group response with members of both

1. Has your State Department of Education (SDE) participated in cross-state research projects?

Yes, as a research site	14 (64%)
Yes, as a project advisor only	3 (14%)
No	1 (5%)
Don't know	4 (18%)

2. How do you think the following features of a cross-state research project would affect your SDE's decision regarding participation?

1 = Less likely to participate, 2 = Somewhat less likely to participate, 3 = No effect, 4 = Somewhat more likely to participate, 5 = More likely to participate

Number (percent) of responses

N=20 Q2.1-Q2.6; **N=19, Q2.7-Q2.16**; N=18, Q2.17-Q2.29

Feature	**1 [Less]**	**2**	**3**	**4**	**5 [More]**	**DNK**
2.1. The involvement of SDE staff in planning and design	0	1 (5)	1 (5)	6 (30)	11 (50)	1 (5)
2.2. The involvement of SDE staff in carrying out the study	1 (5)	6 (30)	0	6 (30)	6 (30)	1 (5)
2.3. The involvement of SDE staff in analyzing the results	2 (10)	4 (20)	1 (5)	5 (25)	6 (30)	2 (10)
2.4. Outside funding for all costs of the study	0	0	0	3 (15)	17 (85)	0
2.5. Mixed outside and participating SDE funding	4 (20)	5 (25)	2 (10)	6 (30)	2 (10)	1 (5)
2.6. All funding from participating SDEs	15 (75)	3 (15)	2 (10)	0	0	0

Feature	1 [Less]	2	3	4	5 [More]	DNK
2.7. Research topic clearly related to NCLB/ESEA requirements	0	0	1 (5)	10 (50)	8 (40)	0
2.8. Research topic clearly related to IDEA requirements	0	0	1 (5)	9 (47)	9 (47)	0
2.9. Research topic selected by participating SDEs, not necessarily clearly tied to federal requirements	1 (5)	6 (32)	3 (16)	4 (21)	4 (21)	1 (5)
2.10. Research topic addresses current need	0	0	0	5 (26)	14 (74)	0
2.11. Research topic is forward-looking	0	1 (5)	0	5 (26)	13 (68)	0
2.12. Summary of existing research and gaps in research is available before SDE makes decision	0	0	2 (11)	9 (47)	8 (42)	0
2.13. SDEs in the cross-state project each conduct a replication of the study (**for example**, all SDEs investigate the generalizability of their constructed-response items using the same method)	0	0	4 (21)	9 (47)	3 (16)	3 (16)
2.14. SDEs in the cross-state project each conduct a state-selected variant of a common research design (**for example**, each state selects the academic subject, grade level, and focus topic for a study examining consequential validity)	0	2 (11)	1 (5)	13 (68)	1 (5)	2 (11)

Feature	1 [Less]	2	3	4	5 [More]	DNK
2.15. SDEs in the cross-state project all use the same materials (**for example**, a set of items measuring content standards that the states have in common)	0	1 (5)	2 (11)	9 (47)	6 (32)	1 (5)
2.16. The research project focuses on original research rather than replicating earlier research	0	0	3 (16)	10 (53)	5 (23)	1 (5)
2.17. The study can be carried out in conjunction with the statewide assessment administration	0	0	1 (6)	3 (17)	13 (72)	1 (5)
2.18. The study requires a stand-alone administration	1 (6)	11 (61)	4 (22)	2 (11)	0	0
2.19. The requirement that a random sample of ***students*** participate	0	7 (39)	6 (33)	2 (11)	1 (6)	2 (11)
2.20. The requirement that a random sample of ***schools*** participate	1 (6)	3 (17)	8 (44)	4 (22)	0	2 (11)
2.21. The requirement that a representative (rather than random) sample of schools participate	0	3 (17)	8 (44)	4 (22)	1 (6)	2 (11)
2.22. Including ***all*** students in the classroom in the study	2 (11)	0	4 (22)	9 (50)	1 (6)	2 (11)
2.23. Including ***only selected*** students in the classroom in the study	2 (11)	9 (50)	3 (17)	2 (11)	0	2 (11)
2.24. The majority of project work is carried out by researchers and consultants external to the SDE.	0	3 (17)	0	6 (33)	8 (44)	1 (6)

Feature	1 [Less]	2	3	4	5 [More]	DNK
2.25. Schools receive a monetary incentive for participating	0	0	3 (17)	6 (33)	7 (39)	2 (11)
2.26. Professional development related to the study topic is offered to participating teachers as part of the research	0	0	0	7 (39)	10 (56)	1 (6)
2.27. The SDE can use study results to modify or provide support for its testing program	0	0	0	3 (17)	14 (78)	1 (6)
2.28. Schools see a direct effect on state testing based on study results	0	0	1 (6)	7 (39)	7 (39)	3 (17)

3. Please provide any **additional** ideas you have about the best possible (but realistic) conditions that would allow your SDE to participate in a collaborative cross-state research project.
4. Please provide comments about your SDE's decision about whether to participate in the CT EAG.
    - 4.1. If your state was recruited to participate in the CT EAG project, but it decided not to join, what were the main reasons for this decision?
    - 4.2. If your state did participate in the CT EAG, what were the main reasons why it did?
    - 4.3. What would facilitate your state agreeing to take a lead role in a future EAG, such as the role Connecticut has for the current project?

Implementing Cross-State Research: Recommendations from Project States

Phoebe Winter, March, 2010

The results of a series of phone conversations with staff from states participating in the CT EAG research are summarized below. The evaluator talked with two to three staff members from each of the four project states in separate conference calls. The participants were told that the purpose of the call was not to critique the CT EAG project but to gather information about the project that would be useful in planning and conducting future cross-state collaborative research. Comments made in the phone calls that are pertinent to future research efforts are organized by topic.

Discussion Questions

1. What worked well in terms of project design and implementation?
2. What would you advise states in collaborative projects like this one to consider or do differently based on your experience with the CTEAG?
3. Is there anything else you'd like to comment on about the project or discuss that would be helpful for future projects?

State Staff Comments

<u>State DOE Conditions and Support</u>

Commitment to the project within the state department of education is a necessity.

Having the right people at the state department of education involved in the study is an important factor in success: keeping management-level staff involved helps in getting ongoing policy advice and staffing resources; having the formal time commitment of staff with expertise in the area of study and in measurement, as well as staff familiar with the data available in the state, keeps the study going smoothly and ensures that tasks are completed as needed.

Collaboration across offices within the state department of education is necessary.

The state need for the information provided by the study is a motivation to participate and to continue to commit resources throughout the study period. One state had planned to conduct a study in the area before the grant was funded because they needed the results to address a current issue.

Practical outcomes – studies in which the results will be used in decision making – can motivate state policymakers to provide resources and support and provide motivation to districts to participate in the study.

Planning and Preconditions for Participation

States working in collaboration must understand the need for sharing data and have no barriers to sharing the data necessary for the work of the collaborative.

Expectations for participation should be spelled out in as much detail as possible prior to states' making the decision to join.

Before agreeing to participate in the study, state representatives should make sure they understand the requirements of the study and be prepared to commit the necessary resources.

A condition for participation could be a consistent state representative as the study contact, if at all possible. This would prevent some miscommunication.

Flexibility is a plus, but there should be well-defined guidelines as well.

Research proposals should include funding for time for state staff time to work on the project.

When possible, the study should be designed so that results are generalizable to or applicable in a broad range of schools, grade levels, and content areas.

It is easier to persuade schools to participate if the study is embedded in the regular state administration or in another activity, rather than as a stand-alone study.

Management and Logistics

Project management should include a representative from each state involved so that all states can provide input to ongoing decisions and discuss critical issues; in cases where this is unwieldy, it might be necessary to have a two-tiered management structure.

The states should have a contact person who understands the study and can answer questions and provide information about the in-state resources needed for the study (e.g., the state's student data bases; the state's item bank).

Each state needs a system for internal communication regarding the study.

Having a third party (CCSSO) handle financial matters (e.g., sending money directly to schools, paying the test contractor) helped smooth the process, avoiding many bureaucratic slow-downs.

Implementation and Follow-Up

Because of limitations in state staff time, having external researchers (John Olson and Mohamed Dirir) available is invaluable. Like John and Mohamed, these researchers need to be responsive to state needs.

Communication and support from the research team is critical (this was a strong aspect of the current study).

Studies should include provisions for monitoring implementation to avoid irregularities in administration that could affect the results.

Financial incentives are not enough to persuade schools to participate, especially when special populations are included in the study. Schools perceive the time in the study as time lost in instruction. Incentives for participation should include benefits for the education of the students involved or other education-related benefits. For example, schools could receive the data in such a way that it could be used diagnostically, for instructional purposes.

While financial incentives were welcome, schools put money second to instructional needs.

Providing study data to each state for follow-up analyses is helpful.

Suggestions for using or following up on the study results would be helpful.

APPENDIX D
Item and Test Statistics from State Studies

This appendix shows the item difficulties and point-biserial correlations by administration condition for all five studies.

1. Connecticut # 1 item statistics (the prefixes of the items represent, M, P, R, & B represent passages)

FORM B: STANDARD GRP 2; ACCOMMODATION GRP 1								FORM A: STANDARD GRP 1; ACCOMMODATION GRP 2						
	DIFFICULTY			POINT BISERIAL					DIFFICULTY			POINT BISERIAL		
ITEM	Group 2	Group1	Difference	Group 2	Group1	Difference		ITEM	Group 1	Group 2	Difference	Group 1	Group 2	Difference
M1	0.768	0.808	-0.040	0.452	0.440	0.013		R1	0.779	0.713	0.067	0.410	0.396	0.013
M2	0.463	0.554	-0.090	0.442	0.424	0.018		R2	0.646	0.556	0.090	0.198	0.348	-0.150
M3	0.896	0.887	0.009	0.399	0.405	-0.007		R3	0.718	0.763	-0.045	0.447	0.502	-0.055
M4	0.598	0.593	0.004	0.442	0.426	0.016		R4	0.600	0.588	0.013	0.531	0.441	0.090
M5	0.537	0.435	0.102	0.317	0.128	0.189		R5	0.513	0.531	-0.018	0.419	0.363	0.055
M6	0.750	0.757	-0.007	0.384	0.391	-0.007		R6	0.605	0.544	0.061	0.326	0.237	0.089
M7	0.482	0.384	0.098	0.450	0.345	0.105		R7	0.436	0.488	-0.052	0.355	0.221	0.134
M8	0.555	0.610	-0.055	0.367	0.361	0.006		R8	0.733	0.756	-0.023	0.410	0.394	0.016
M9	0.695	0.701	-0.005	0.550	0.415	0.135		R9	0.564	0.669	-0.105	0.523	0.380	0.143
M10	0.530	0.503	0.028	0.323	0.230	0.093		R10	0.754	0.813	-0.059	0.496	0.545	-0.050
P1	0.299	0.311	-0.012	0.211	0.062	0.149		B1	0.579	0.638	-0.058	0.504	0.357	0.147
P2	0.622	0.768	-0.146	0.479	0.318	0.162		B2	0.836	0.831	0.005	0.387	0.286	0.101
P3	0.543	0.559	-0.017	0.370	0.436	-0.066		B3	0.323	0.344	-0.021	0.111	0.197	-0.086
P4	0.213	0.186	0.027	-0.186	0.028	-0.214		B4	0.595	0.750	-0.155	0.450	0.354	0.096
P5	0.201	0.181	0.020	0.051	-0.010	0.062		B5	0.656	0.688	-0.031	0.451	0.590	-0.140
P6	0.628	0.616	0.012	0.286	0.345	-0.059		B6	0.303	0.250	0.053	0.081	0.088	-0.007
P7	0.530	0.588	-0.057	0.541	0.413	0.128		B7	0.641	0.694	-0.053	0.466	0.345	0.121
P8	0.329	0.305	0.024	0.127	0.101	0.026		B8	0.615	0.644	-0.028	0.540	0.575	-0.035
P9	0.341	0.350	-0.009	-0.054	0.116	-0.171		B9	0.554	0.606	-0.052	0.537	0.249	0.289
P10	0.762	0.757	0.005	0.470	0.391	0.079		B10	0.759	0.763	-0.004	0.436	0.416	0.020
Mean	**0.537**	**0.543**	**-0.005**	**0.321**	**0.288**	**0.033**		**Mean**	**0.611**	**0.631**	**-0.021**	**0.404**	**0.364**	**0.040**
Min	**0.201**	**0.181**	**-0.146**	**-0.186**	**-0.010**	**-0.214**		**Min**	**0.303**	**0.250**	**-0.155**	**0.081**	**0.088**	**-0.150**
Max	**0.896**	**0.887**	**0.102**	**0.550**	**0.440**	**0.189**		**Max**	**0.836**	**0.831**	**0.090**	**0.540**	**0.590**	**0.289**
Reliability				**0.759**	**0.721**			**Reliability**				**0.830**	**0.797**	

2. Connecticut # 2 item statistics

FORM A							FORM B						
	Item Difficulty			Point Biserial Correlation				Item Difficulty			Point Biserial Correlation		
Item	Standard	Accomm.	Difference	Standard	Accomm.	Difference	**Item**	Standard	Accomm.	Difference	Standard	Accomm.	Difference
1	0.809	0.814	-0.004	0.548	0.597	-0.049	**1**	0.900	0.915	-0.014	0.370	0.401	-0.031
2	0.834	0.820	0.014	0.543	0.583	-0.040	**2**	0.751	0.780	-0.030	0.521	0.570	-0.049
3	0.772	0.786	-0.013	0.523	0.583	-0.061	**3**	0.841	0.774	0.067	0.538	0.588	-0.050
4	0.729	0.717	0.012	0.560	0.589	-0.029	**4**	0.735	0.649	0.086	0.465	0.561	-0.096
5	0.846	0.835	0.011	0.338	0.396	-0.058	**5**	0.788	0.768	0.020	0.394	0.569	-0.175
6	0.686	0.708	-0.022	0.319	0.463	-0.144	**6**	0.667	0.631	0.036	0.573	0.589	-0.016
7	0.212	0.168	0.045	0.086	0.093	-0.007	**7**	0.595	0.619	-0.024	0.450	0.535	-0.085
8	0.674	0.717	-0.044	0.414	0.534	-0.120	**8**	0.785	0.799	-0.014	0.475	0.521	-0.046
9	0.542	0.556	-0.014	0.524	0.528	-0.004	**9**	0.601	0.595	0.007	0.304	0.297	0.006
10	0.474	0.522	-0.048	0.380	0.350	0.030	**10**	0.548	0.515	0.033	0.346	0.333	0.013
11	0.412	0.432	-0.019	0.222	0.244	-0.022	**11**	0.480	0.409	0.071	0.411	0.408	0.002
12	0.434	0.379	0.055	0.307	0.246	0.061	**12**	0.374	0.345	0.029	0.215	0.090	0.126
13	0.898	0.870	0.029	0.121	0.355	-0.235	**13**	0.972	0.915	0.057	0.182	0.319	-0.137
14	0.714	0.680	0.034	0.397	0.461	-0.064	**14**	0.713	0.720	-0.006	0.543	0.549	-0.007
15	0.674	0.643	0.031	0.420	0.428	-0.008	**15**	0.548	0.573	-0.025	0.480	0.505	-0.025
16	0.671	0.575	0.096	0.227	0.291	-0.064	**16**	0.660	0.713	-0.053	0.487	0.451	0.036
17	0.511	0.512	-0.002	0.468	0.426	0.042	**17**	0.489	0.430	0.059	0.496	0.451	0.045
18	0.418	0.422	-0.004	0.149	0.284	-0.135	**18**	0.374	0.311	0.063	0.289	0.314	-0.025
19	0.775	0.742	0.033	0.509	0.510	-0.001	**19**	0.710	0.747	-0.037	0.509	0.458	0.051
20	0.662	0.736	-0.074	0.578	0.580	-0.002	**20**	0.779	0.741	0.038	0.518	0.549	-0.031
21	0.640	0.658	-0.018	0.401	0.543	-0.142	**21**	0.645	0.582	0.063	0.430	0.454	-0.024
22	0.738	0.770	-0.032	0.461	0.521	-0.060	**22**	0.626	0.634	-0.008	0.319	0.400	-0.080
23	0.437	0.441	-0.004	0.460	0.537	-0.077	**23**	0.654	0.689	-0.035	0.554	0.429	0.125
24	0.815	0.764	0.051	0.438	0.623	-0.185	**24**	0.776	0.716	0.059	0.421	0.405	0.016
25	0.717	0.683	0.034	0.378	0.371	0.006	**25**	0.583	0.515	0.067	0.290	0.395	-0.105
26	0.302	0.398	-0.096	0.297	0.397	-0.100	**26**	0.738	0.726	0.013	0.409	0.396	0.013
Mean	**0.631**	**0.629**	**0.002**	**0.387**	**0.444**	**-0.056**	**Mean**	**0.664**	**0.643**	**0.020**	**0.423**	**0.446**	**-0.022**
Min	**0.212**	**0.168**	**-0.096**	**0.086**	**0.093**	**-0.235**	**Min**	**0.374**	**0.311**	**-0.053**	**0.182**	**0.090**	**-0.175**
Max	**0.898**	**0.870**	**0.096**	**0.578**	**0.623**	**0.061**	**Max**	**0.972**	**0.915**	**0.086**	**0.573**	**0.589**	**0.126**

Reliability	**0.847**	**0.880**			**Reliability**	**0.869**	**0.880**

3. Kentucky item statistics

FORM A							FORM D						
	Item Difficulty			Point Biserial Correlation				Item Difficulty			Point Biserial Correlation		
Item	Standard	Accomm.	Difference	Standard	Accomm.	Difference	Item	Standard	Accomm.	Difference	Standard	Accomm.	Difference
1	0.639	0.777	-0.138	0.573	0.471	0.102	23	0.669	0.701	-0.032	0.390	0.299	0.092
2	0.728	0.926	-0.198	0.527	0.371	0.156	24	0.662	0.816	-0.154	0.483	0.414	0.070
3	0.633	0.743	-0.111	0.506	0.291	0.215	25	0.804	0.857	-0.053	0.501	0.504	-0.003
4	0.639	0.892	-0.252	0.530	0.474	0.056	26	0.669	0.782	-0.113	0.447	0.476	-0.030
5	0.537	0.804	-0.267	0.592	0.447	0.145	27	0.709	0.714	-0.005	0.489	0.037	0.452
6	0.830	0.926	-0.096	0.476	0.345	0.131	28	0.635	0.728	-0.093	0.478	0.425	0.053
7	0.578	0.608	-0.030	0.380	0.333	0.047	29	0.797	0.864	-0.067	0.494	0.581	-0.087
8	0.524	0.709	-0.186	0.445	0.476	-0.031	30	0.615	0.762	-0.147	0.401	0.331	0.070
9	0.510	0.628	-0.118	0.542	0.428	0.113	31	0.730	0.830	-0.100	0.460	0.516	-0.056
10	0.592	0.730	-0.138	0.563	0.531	0.032	32	0.770	0.884	-0.114	0.512	0.488	0.025
11	0.531	0.770	-0.240	0.444	0.500	-0.056	33	0.649	0.701	-0.052	0.488	0.448	0.040
12	0.673	0.770	-0.097	0.346	0.364	-0.018	34	0.439	0.510	-0.071	0.175	0.247	-0.072
13	0.769	0.797	-0.029	0.304	0.416	-0.112	35	0.568	0.619	-0.051	0.451	0.347	0.104
14	0.741	0.932	-0.191	0.586	0.421	0.164	36	0.568	0.599	-0.031	0.465	0.280	0.185
15	0.646	0.831	-0.185	0.578	0.290	0.288	37	0.372	0.476	-0.105	0.111	0.188	-0.077
16	0.673	0.743	-0.070	0.555	0.465	0.090	38	0.824	0.905	-0.080	0.511	0.226	0.286
17	0.401	0.669	-0.268	0.583	0.389	0.194	39	0.750	0.837	-0.087	0.574	0.347	0.228
18	0.469	0.493	-0.024	0.314	0.283	0.031	40	0.615	0.864	-0.249	0.474	0.314	0.159
19	0.497	0.777	-0.280	0.572	0.408	0.164	41	0.676	0.857	-0.181	0.580	0.302	0.278
20	0.639	0.804	-0.165	0.554	0.386	0.168	42	0.628	0.687	-0.059	0.352	0.359	-0.007
21	0.537	0.628	-0.091	0.406	0.349	0.057	43	0.622	0.667	-0.045	0.469	0.406	0.063
22	0.605	0.770	-0.165	0.412	0.279	0.134	44	0.547	0.857	-0.310	0.551	0.515	0.036
Mean	**0.609**	**0.760**	**-0.152**	**0.490**	**0.396**	**0.094**	**Mean**	**0.651**	**0.751**	**-0.100**	**0.448**	**0.366**	**0.082**
Max	**0.830**	**0.932**	**-0.024**	**0.592**	**0.531**	**0.288**	**Max**	**0.824**	**0.905**	**-0.005**	**0.580**	**0.581**	**0.452**
Min	**0.401**	**0.493**	**-0.280**	**0.304**	**0.279**	**-0.112**	**Min**	**0.372**	**0.476**	**-0.310**	**0.111**	**0.037**	**-0.087**
Reliability				**0.891**	**0.831**		**Reliability**				**0.866**	**0.803**	

4. Nevada item statistics

FORM A							FORM B						
	Item Difficulty			Point Biserial Correlation				Item Difficulty			Point Biserial Correlation		
Item	Standard	Accomm.	Difference	Standard	Accomm.	Difference	Item	Standard	Accomm.	Difference	Standard	Accomm.	Difference
1	0.407	0.444	-0.038	0.462	0.560	-0.098	21	0.610	0.591	0.019	0.396	0.469	-0.072
2	0.627	0.541	0.086	0.336	0.314	0.022	22	0.630	0.606	0.024	0.407	0.483	-0.076
3	0.755	0.792	-0.037	0.306	0.298	0.007	23	0.395	0.458	-0.063	0.222	0.180	0.042
4	0.608	0.565	0.043	0.252	0.294	-0.042	24	0.545	0.542	0.003	0.384	0.337	0.047
5	0.515	0.565	-0.051	0.413	0.433	-0.019	25	0.660	0.695	-0.035	0.322	0.375	-0.052
6	0.824	0.826	-0.003	0.308	0.268	0.039	26	0.515	0.483	0.032	0.273	0.277	-0.004
7	0.681	0.715	-0.034	0.318	0.419	-0.101	27	0.340	0.212	0.128	0.481	0.344	0.137
8	0.387	0.449	-0.062	0.278	0.419	-0.140	28	0.385	0.315	0.070	0.389	0.398	-0.009
9	0.490	0.459	0.031	0.264	0.278	-0.014	29	0.505	0.478	0.027	0.394	0.332	0.062
10	0.721	0.768	-0.048	0.414	0.389	0.025	30	0.375	0.473	-0.098	0.483	0.298	0.185
11	0.407	0.362	0.045	0.274	0.324	-0.050	31	0.270	0.340	-0.070	0.307	0.316	-0.009
12	0.373	0.425	-0.053	0.319	0.399	-0.080	32	0.375	0.389	-0.014	0.301	0.267	0.034
13	0.422	0.502	-0.081	0.478	0.409	0.068	33	0.670	0.631	0.039	0.431	0.373	0.059
14	0.475	0.430	0.046	0.383	0.392	-0.008	34	0.625	0.704	-0.079	0.234	0.444	-0.210
15	0.740	0.807	-0.067	0.472	0.408	0.063	35	0.375	0.433	-0.058	0.388	0.434	-0.045
16	0.456	0.478	-0.022	0.283	0.372	-0.088	36	0.575	0.660	-0.085	0.481	0.452	0.030
17	0.667	0.589	0.077	0.447	0.456	-0.008	37	0.590	0.675	-0.085	0.382	0.343	0.040
18	0.529	0.536	-0.007	0.404	0.329	0.076	38	0.625	0.581	0.044	0.148	0.190	-0.042
19	0.549	0.551	-0.002	0.367	0.505	-0.138	39	0.615	0.680	-0.065	0.443	0.465	-0.022
20	0.387	0.372	0.015	0.200	0.206	-0.006	40	0.485	0.581	-0.096	0.226	0.230	-0.004
Mean	**0.551**	**0.559**	**-0.008**	**0.349**	**0.374**	**-0.025**	**Mean**	**0.508**	**0.526**	**-0.018**	**0.355**	**0.350**	**0.004**
Min	**0.373**	**0.362**	**-0.081**	**0.200**	**0.206**	**-0.140**	**Min**	**0.270**	**0.212**	**-0.098**	**0.148**	**0.180**	**-0.210**
Max	**0.824**	**0.826**	**0.086**	**0.478**	**0.560**	**0.076**	**Max**	**0.670**	**0.704**	**0.128**	**0.483**	**0.483**	**0.185**
Reliability				**0.810**	**0.790**		**Reliability**				**0.790**	**0.790**	

5. Michigan item statistics

REGULAR				ACCESS (Enhanced directions)		
Item	Difficulty	Point Biserial		Ite m	Difficulty	Point Biserial
1	0.451	0.432		1	0.671	0.395
2	0.365	0.402		2	0.712	0.259
3	0.254	0.437		3	0.643	0.394
4	0.546	0.325		4	0.709	0.395
5	0.324	0.395		5	0.676	0.413
6	0.385	0.403		6	0.669	0.454
7	0.337	0.440		7	0.707	0.421
8	0.281	0.289		8	0.471	0.373
9	0.294	0.247		9	0.634	0.327
10	0.403	0.356		10	0.687	0.339
11	0.473	0.365		11	0.439	0.328
12	0.378	0.359		12	0.452	0.368
13	0.248	0.409		13	0.671	0.405
14	0.216	0.403		14	0.631	0.363
15	0.203	0.367		15	0.663	0.417
16	0.233	0.409		16	0.594	0.401
17	0.357	0.404		17	0.424	0.251
18	0.323	0.420		18	0.553	0.466
19	0.287	0.462		19	0.586	0.363
20	0.369	0.382		20	0.659	0.400
21	0.341	0.355		21	0.406	0.252
22	0.483	0.339		22	0.589	0.296
23	0.290	0.232		23	0.695	0.409
24	0.429	0.385		24	0.579	0.370
25	1.929	0.414		25	2.023	0.476
Reading						
Mean	**0.337**	**0.384**			**0.603**	**0.371**
Min	**0.203**	**0.247**			**0.406**	**0.251**
Max	**0.546**	**0.462**			**0.712**	**0.466**
Writing MC						
Mean	**0.401**	**0.319**			**0.621**	**0.358**
Min	**0.290**	**0.232**			**0.579**	**0.296**
Max	**0.483**	**0.385**			**0.695**	**0.409**
Reliability		**0.833**		**Reliability**		**0.832**

Appendix E
Examples of Additional and Post Hoc Analyses

Various types of additional analyses were conducted post hoc during the course of the CTEAG project (as noted in the body of the technical report). In this appendix, results are shown for analyses done to examine (a) effects of test accommodations on different student ability levels and (b) effect sizes.

a) Grouping students to re-examine effects

The chart below summarizes the results from the Repeated Measures ANOVA after manipulating the samples in different ways. Please note that all groupings were based on performance on the pretest scores only, which was from the form administered in the standard condition. For SWD, for example, two groups of high scorers and low scorers based on their standard test scores were formed. For the top ¾, students whose scores placed them in the lowest performing quartile were excluded from the analysis while the students in the highest performing quartile were excluded from the bottom ¾ group analysis.

Among the groups, interaction effects were found for all SWD in all four studies that were analyzed. On the other hand, no interaction effects were found for the bottom ¾ of the whole sample.

Connecticut # 2

		Was the effect significant?		
Group	N Size	Accommodation	Interaction	Comments
All	**633**	**X**	√	**This is our initial finding with whole sample.**
Top 3/4	487	√	√	The lowest ¼; mostly SWD, was excluded. Both means went down, but more for SWOD.
Bottom 3/4	473	X	Borderline (0.08)	The top ¼; mostly SWOD, was excluded.
Two halves of SWD	276	X	√	Accommodation effect on high and low SWD.

√ = Yes X = No

Nevada

		Was the effect significant?		
Group	N Size	Accommodation	Interaction	Comments
All	**415**	Borderline (0.05)	**X**	**This is our initial finding with whole sample.**
Top 3/4	335	Borderline (0.06)	X	Both means went down.
Bottom 3/4	316	√	X	Both means went up.
Two halves of SWD	211	X	√	Accommodation effect on high and low SWD. Means went up (bottom) and down (top) .

√ = Yes X = No

Kentucky

		Was the effect significant?		
Group	N Size	Accommodation	Interaction	Comments
All	**295**	√	√	**This is our initial finding with whole sample.**
Top 3/4	222	√	X	Both means went up.
Bottom 3/4	217	√	X	Both means went up.
Two halves of SWD	150	√	√	Accommodation effect on high and low SWD. Mean for bottom half went up more.

√ = Yes X = No

Michigan

		Was the effect significant?		
Group	N Size	Accommodation	Interaction	Comments
All	**693**	√	√	**This is our initial finding with whole sample.**
Top 3/4	539	√	√	Both means went up; more for SWD.
Bottom 3/4	500	√	X	Both means went up.
Two halves of SWD	500	√	√	Accommodation effect on high and low SWD. Both means went up; more for lower half.

√ = Yes X = No

b) Effect sizes

The effect size is presented as partial eta-squared (η^2_p), as recommended for univariate significance tests based on F or t. The formula used is based on the F values and the degrees of freedom for each significant effect. The formula could be written as (Stevens, 1996, p.177):

$$\eta_p^2 = \frac{df_h}{df_h} + df_e$$

Where df_h is the degrees of freedom for the hypothesis and df_e is the degrees of freedom for the error.

Below is a tabulation of η^2_p for the significant effects in four of the studies:

Study	**Significant effect**	**Effect size (η^2_p)**
Connecticut # 2	Disability status × testing condition	0.02
Nevada	Testing condition (borderline)	0.01
Kentucky	Testing condition	0.30
	Disability status × testing condition	0.05
Michigan	Test form	0.59
	Disability status × test form	0.05

Guidelines on interpreting size of η^2_p recommended for univariate tests (e.g. Cohen, 1988):

Small:	0.01
Medium:	0.06
Large:	0.14

APPENDIX F
Examples of Statistical Analysis Procedures Used

In this appendix, a brief summary of the statistical procedures used in the studies is provided. Examples of statistical codes and outputs are presented for the three main analyses that were conducted: (a) score summaries, (b) repeated measures analysis of variance, and (c) factor analysis.

a) Scores summaries

The SAS software was used to compute score summaries (mean, standard deviation, N count) in all studies. Below is an example of the SAS code that was used for computing summary statistics for the Connecticut study #2.

```
  PROC TABULATE DATA=CT2 FORMAT=5.1;  CLASS ACCOMMODATION SPED FORM; VAR TOTAL;
     TABLE FORM*(ACCOMMODATION all)*(SPED all), TOTAL*(N MEAN STD); LABEL SPED='SWD';
RUN;
```

Output of the above SAS code:

			TOTAL		
			N	Mean	Std
Form	Accommodation	SWD			
Form A	With	No	189	18.3	5.4
		Yes	134	13.6	5.4
		All	323	16.3	5.9
	Without	SWD			
		No	177	18.9	4.4
		Yes	148	13.4	4.8
		All	325	16.4	5.3
	All	SWD			
		No	366	18.6	4.9
		Yes	282	13.5	5.1
		All	648	16.4	5.6
Form B	Accommodation	SWD			
	With	No	177	19.2	5.0
		Yes	150	14.0	5.6
		All	327	16.8	5.9
	Without	SWD			
		No	189	19.9	4.5
		Yes	132	13.7	5.1
		All	321	17.3	5.7
	All	SWD			
		No	366	19.6	4.7
		Yes	282	13.9	5.4
		All	648	17.1	5.8

b) Repeated Measures Analysis of Variance (ANOVA)

The test data for all five studies were arranged in a way that fulfils requirements for a repeated measures ANOVA. Scores from the standard administration were used as the pretest, and scores from the accommodated administration were used as the post test. The assumption in the ANOVA was that the pre and post tests were parallel, which was one of the requirements in designing the studies. In general, the framework considers both SWD and SWOD as being tested in standard form, and then an intervention or treatment was provided to each student in the form of test accommodation and students were tested again.

The following SAS code is a sample of what was used in processing the repeated measures ANOVA for the Kentucky study.

```
DATA ANOVA;  SET KY;
  TITLE KENTUCKY - ANOVA RESULTS;
PROC GLM DATA=ANOVA;   CLASS EDPROGRAM;  MODEL STDD ACCOM=EDPROGRAM  / SOLUTION ;
  lsmeans EDPROGRAM ;  repeated ACCOMMODATION;  LABEL STDD='STANDARD'
ACCOM='ACCOMMODATED';
   QUIT;  RUN;
```

Selected parts of the ANOVA output are presented below.

KENTUCKY - ANOVA RESULTS

The GLM Procedure

Class Level Information		
Class	**Levels**	**Values**
EdProgram	2	G S

Number of Observations Read	295
Number of Observations Used	295

Dependent Variable: STANDARD

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	2791.214592	2791.214592	131.13	<.0001
Error	293	6236.805747	21.286026		
Corrected Total	294	9028.020339			

Source	DF	Type I SS	Mean Square	F Value	Pr > F
EdProgram	1	2791.214592	2791.214592	131.13	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
EdProgram	1	2791.214592	2791.214592	131.13	<.0001

Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	10.83333333	B	0.37670525	28.76	<.0001
EdProgram G	6.15287356	B	0.53731464	11.45	<.0001
EdProgram S	0.00000000	B			

Dependent Variable: ACCOMMODATED

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1291.338496	1291.338496	99.62	<.0001
Error	293	3797.895402	12.962100		
Corrected Total	294	5089.233898			

Source	DF	Type I SS	Mean Square	F Value	Pr > F
EdProgram	1	1291.338496	1291.338496	99.62	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
EdProgram	1	1291.338496	1291.338496	99.62	<.0001

Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	14.56666667	B	0.29396259	49.55	<.0001
EdProgram G	4.18505747	B	0.41929440	9.98	<.0001
EdProgram S	0.00000000	B			

Least Squares Means

EdProgram	STDD LSMEAN	ACCOM LSMEAN
G	16.9862069	18.7517241
S	10.8333333	14.5666667

Repeated Measures Analysis of Variance

Repeated Measures Level Information		
Dependent Variable	STDD	ACCOM
Level of ACCOMMODATION	1	2

MANOVA Test Criteria and Exact F Statistics for the Hypothesis of no ACCOMMODATION Effect H = Type III SSCP Matrix for ACCOMMODATION E = Error SSCP Matrix S=1 M=-0.5 N=145.5					
Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.70126671	124.82	1	293	<.0001
Pillai's Trace	0.29873329	124.82	1	293	<.0001
Hotelling-Lawley Trace	0.42599097	124.82	1	293	<.0001
Roy's Greatest Root	0.42599097	124.82	1	293	<.0001

MANOVA Test Criteria and Exact F Statistics for the Hypothesis of no ACCOMMODATION*EdProgram Effect H = Type III SSCP Matrix for ACCOMMODATION*EdProgram E = Error SSCP Matrix S=1 M=-0.5 N=145.5				
Statistic	Value	F Value	Num DF	Den DF
Wilks' Lambda	0.94826829	15.98	1	293
Pillai's Trace	0.05173171	15.98	1	293
Hotelling-Lawley Trace	0.05455387	15.98	1	293
Roy's Greatest Root	0.05455387	15.98	1	293

MANOVA Test Criteria and Exact F Statistics for the Hypothesis of no ACCOMMODATION*EdProgram Effect H = Type III SSCP Matrix for ACCOMMODATION*EdProgram E = Error SSCP Matrix S=1 M=-0.5 N=145.5	
Statistic	**Pr > F**
Wilks' Lambda	<.0001
Pillai's Trace	<.0001
Hotelling-Lawley Trace	<.0001
Roy's Greatest Root	<.0001

Repeated Measures Analysis of Variance
Tests of Hypotheses for Between Subjects Effects

Source	DF	Type III SS	Mean Square	F Value	Pr > F
EdProgram	1	3939.803039	3939.803039	155.62	<.0001
Error	293	7418.020690	25.317477		

Repeated Measures Analysis of Variance
Univariate Tests of Hypotheses for Within Subject Effects

Source	DF	Type III SS	Mean Square	F Value	Pr > F
ACCOMMODATION	1	1114.682252	1114.682252	124.82	<.0001
ACCOMMODATION*EdProgram	1	142.750049	142.750049	15.98	<.0001
Error(ACCOMMODATION)	293	2616.680460	8.930650		

c) Factor Analysis

Below is a simple SAS code that was used to conduct the exploratory factor analysis for the Nevada study. Examples of the results from this analysis are presented as well.

```
TITLE "Factor Analysis: Nevada, Form B";
```

```
PROC FACTOR DATA=FAC
      METHOD=PRIN
      VARDEF=DF
      SINGULAR=1E-08
      MINEIGEN=1.0
      PRIORS=ONE
      ROTATE=NONE
      PLOTS=SCREE;
      VAR CT1 CT2 PD1 PD2 PS1 PS2;        BY TESTING;
RUN;
```

Factor Analysis: Nevada Form B

The FACTOR Procedure

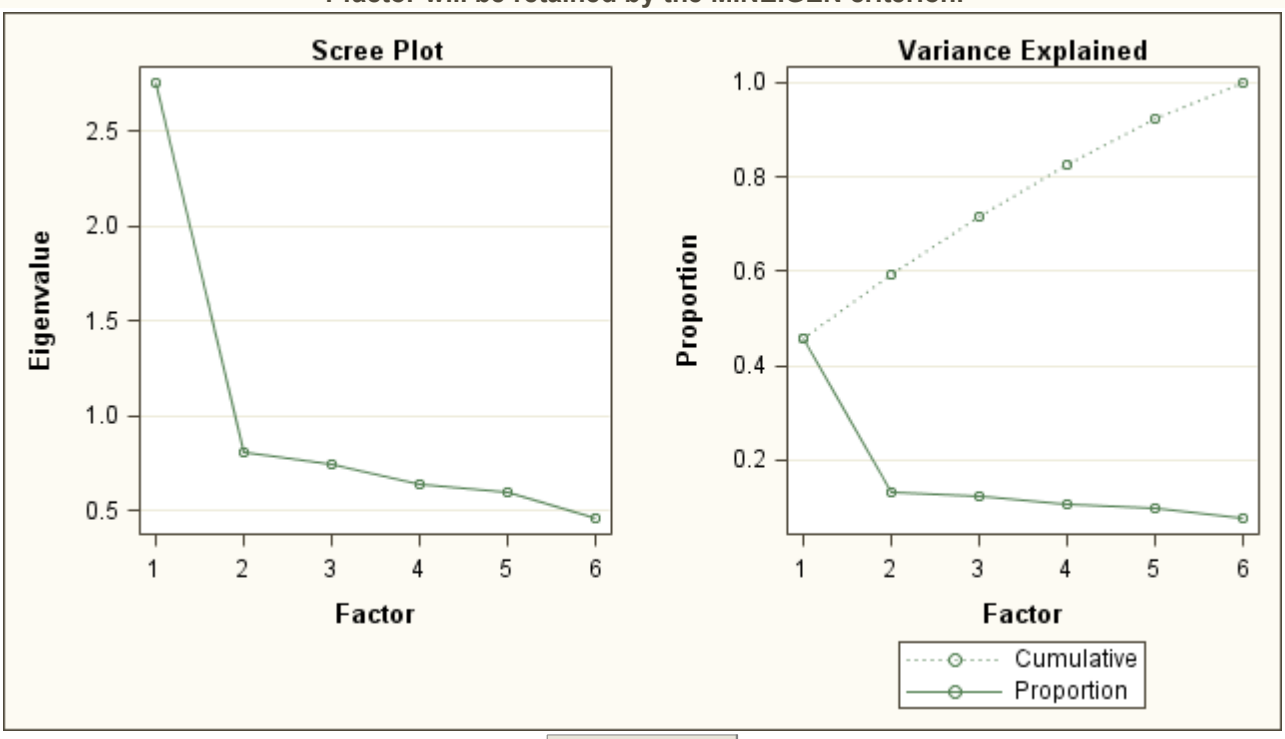
Initial Factor Method: Principal Components

TESTING=ACCOMM

Prior Communality Estimates: ONE

Eigenvalues of the Correlation Matrix: Total = 6 Average = 1				
	Eigenvalue	Difference	Proportion	Cumulative
1	2.75684082	1.95158372	0.4595	0.4595
2	0.80525711	0.05942479	0.1342	0.5937
3	0.74583231	0.10454366	0.1243	0.7180
4	0.64128865	0.04594156	0.1069	0.8249
5	0.59534710	0.13991308	0.0992	0.9241
6	0.45543401		0.0759	1.0000

1 factor will be retained by the MINEIGEN criterion.



Factor Pattern	
	Factor1
CT1	0.67470
CT2	0.60428
PD1	0.67999

Factor Pattern	
	Factor1
PD2	0.78154
PS1	0.68282
PS2	0.63010

Variance Explained by Each Factor
Factor1
2.7568408

Final Communality Estimates: Total = 2.756841					
CT1	CT2	PD1	PD2	PS1	PS2
0.45522593	0.36515238	0.46238736	0.61080349	0.46624541	0.39702626

Factor Analysis: Nevada Form B

The FACTOR Procedure

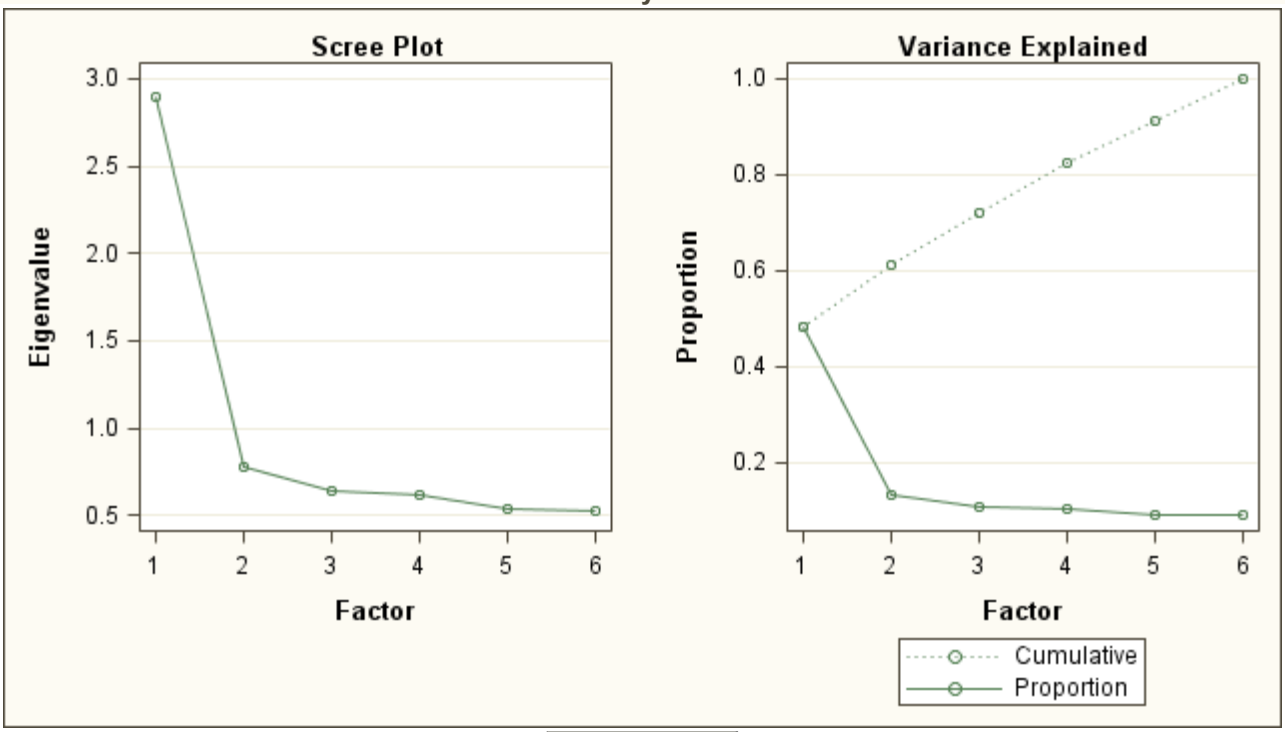
Initial Factor Method: Principal Components

TESTING=STANDARD

Prior Communality Estimates: ONE

Eigenvalues of the Correlation Matrix: Total = 6 Average = 1				
	Eigenvalue	Difference	Proportion	Cumulative
1	2.89594955	2.11531364	0.4827	0.4827
2	0.78063592	0.13600938	0.1301	0.6128
3	0.64462653	0.02838174	0.1074	0.7202
4	0.61624479	0.08289331	0.1027	0.8229
5	0.53335148	0.00415976	0.0889	0.9118
6	0.52919172		0.0882	1.0000

1 factor will be retained by the MINEIGEN criterion.



Factor Pattern	
	Factor1

Factor Pattern	
	Factor1
CT2	0.60025
PD1	0.69059
PD2	0.72114
PS1	0.72337
PS2	0.66816

Variance Explained by Each Factor
Factor1
2.8959496

Final Communality Estimates: Total = 2.895950					
CT1	CT2	PD1	PD2	PS1	PS2
0.56899117	0.36029933	0.47691485	0.52003985	0.52326786	0.44643649

APPENDIX G
Accommodations Bibliography: Experimental Studies

Bishop-Temple, C. (2007). The effects of interactive read-alouds on the reading achievement of middle grade reading students in a core remedial program. Dissertations International, Florida Atlantic University.

Brown, D.W. (2007). The role of reading in science: Validating graphics in large-scale science assessment. Assessment for Effective Intervention.

Cahalan, C., Mandinach, E., & Camara, W.J. (2002). Predictive validity of SAT I: Reasoning test for test-takers with learning disabilities and extended time accommodations. New York, NY: College Entrance Examination Board, 2002–05.

Curtis, H.A. & Kropp, R.P. (1961). A comparison of scores obtained by administering a test normally and visually. Journal of Experimental Education, 29 (3), 249–260.

DiRosa, F. (2008). The impact of testing accommodations on individual postsecondary student test outcomes. Dissertations International, Temple University.

Elliott, S.N. (1999). Valid testing accommodations: Fundamental assumptions and methods for collecting validity evidence. Paper presented at the CCSSO National Conference on Large-Scale Assessment, Snowbird, UT.

Elliott, S.N., Kratochwill, T.R.; McKevitt, B. (2001). Experimental analysis of the effects of testing accommodations on the scores of students with and without disabilities. Journal of School Psychology, 39, 3–24.

Fletcher, J.M., Francis, D.J., Boudousquie, A., Copeland, Young, Kalinoski, & Vaughn (2006). Effects of Accommodations on High-Stakes Testing for Students with Reading Disabilities. Exceptional Children, 72 (2), 136–150.

Fuchs, L.S. (2000). Research Report on Math. The validity of test accommodations for students with learning disabilities. Differential item performance on reading and mathematics test as a function of test accommodations and disability status. Newark, DE: Delaware Education Research and Development Center, University of Delaware, 1.

Fuchs, L.S. (2000). Research Report on Reading. The validity of test accommodations for students with learning disabilities. Differential item performance on reading and mathematics tests as a function of test accommodations and disability status. Newark, DE: Delaware Education Research and Development Center, University of Delaware, #2.

Helwig, R. & Tindal, G. (2003). An Experimental Analysis of Accommodation Decisions on Large-Scale Mathematics Tests. Exceptional Children, 69 (2), 211–225.

Higgins, J., Russell, M., & Hoffman, T. (2005). Examining the Effect of Computer-Based Passage Presentation on Reading Test Performance. The Journal of Technology, Learning, and Assessment, 3 (4).

Hollenbeck, K., Tindal, G., & Almond, P. (1998). Teachers' knowledge of accommodations as a validity issue in high-stakes testing. The Journal of Special Education, 32 (3), 175–183.

Huesman, R. (1999). The validity of ITBS reading comprehension test scores for learning disabled and non learning disabled students under extended-time conditions. Dissertation Abstracts International.

Huesman, R. & Frisbie, D. (2000). The validity of ITBS reading comprehension test scores for learning disabled and non learning disabled students under extended-time conditions. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Kamei-Hannan, C. (2008). Examining the Accessibility of a Computerized Adapted Test Using Assistive Technology. Journal of Visual Impairment & Blindness, 102 (5), 261–271.

Keng, L., McClarty, K.L., & Davis, L.L. (2008). Item-Level Comparative Analysis of Online and Paper Administrations of the Texas Assessment of Knowledge and Skills. Applied Measurement in Education, 21 (3), 207–226.

Ketterlin-Geller, L.R., Yovanoff, P., & Tindal, G. (2007). Developing a New Paradigm for Conducting Research on Accommodations in Mathematics Testing. Exceptional Children, 73 (3), 331–347.

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.

Kettler, R.J., Niebling, B.C., Mroch, A.A., Feldman, E.S., Newell, M.L., Elliot. S.N., Kratchowill, T.R., & Bolt, D.M. (2005). Effects of Testing Accommodations on Math and Reading Scores: An Experimental Analysis of the Performance of Students With and Without Disabilities. Assessment for Effective Intervention, 31 (1), 37–48.

Kim, D.H. & Huynh, H. (2008). Computer-Based and Paper-and-Pencil Administration Mode Effects on a Statewide End-of-Course English Test. Educational and Psychological Measurement, 68 (4), 554–570.

Koretz, D. & Hamilton, L. (2000). Assessment of students with disabilities in Kentucky: Inclusion, student performance, and validity. Educational Evaluation and Policy Analysis, 22 (3), 255–272.

Kosciolek, S. & Ysseldyke, J.E. (2000). Effects of a reading accommodation on the validity of a reading test. Minneapolis, MN: National Center on Educational Outcomes, University of Minnesota, 28.

Lewandowski, L.J., Lovett, B.J., Parolin, R., Gordon, M., & Codding, R.S. (2007). Extended Time Accommodations and the Mathematics Performance of Students with and without ADHD. Journal of Psychoeducational Assessment, 25 (1), 17–28.

Lewis, D., Green, D.R., & Miller, L. (1999). Using differential item functioning analyses to assess the validity of testing accommodated students with disabilities. Paper presented at the National Conference on Large-Scale Assessment, Snowbird, UT.

Lindstrom, J.H. & Gregg, N. (2007). The Role of Extended Time on the SAT for Students with Learning Disabilities and/or Attention-Deficit/Hyperactivity Disorder. Learning Disabilities Research and Practice, 22 (2), 85–95.

Lovett, B.J. (2008). Determinants of postsecondary students' performance on timed examinations: Implications for extended time testing accommodations. Dissertations International, Syracuse University.

Middleton, K.V. (2007). The effect of a read-aloud accommodation on items on a reading comprehension test for students with reading-based learning disabilities. Dissertations International, The University of Iowa.

Robinson, G. & Conway, R. (1990). The effects of Irlen colored lenses on students' specific reading skills and their perception of ability: A 12-month validity study. Journal of Learning Disabilities, 23 (10), 589–596.

Russell, M. & Haney, W. (1997). Testing writing on computers: an experiment comparing student performance on tests conducted via computer and via paper-and-pencil. Educational Policy Analysis Archives, 5 (3).

Schulte, A.G. (2000). Experimental analysis of the effects of testing accommodations on student's standardized mathematics test scores. Dissertation Abstracts International, 61/08, 3056.

Schulte, A.G., Elliott, S.N., & Kratochwill, T.R. (2000). Experimental analysis of the effects of testing accommodations on students' standardized mathematics test scores. Paper presented at the annual meeting of the Council of Chief State School Officers (CCSSO), Snowbird, UT.

Thornton, A.E., Reese, L.M.; Pashley, P.J., & Dalessandro, S.P. (2002). Predictive validity of accommodated LSAT scores. Newton, Pennsylvania: Law School Admission Council.

Tindal, G., Heath, B., Hollenbeck, K., Almond, P., & Harniss, M. (1998). Accommodating students with disabilities on large-scale tests: An experimental study. Exceptional Children, 64 (4), 439–450.

Weaver, S.M. (1993). The validity of the use of extended and untimed testing for postsecondary students with learning disabilities. Dissertation Abstracts International, 55/03, 535.

Weston, T. (2002). The Validity of Oral Accommodation in Testing. Palo Alto, CA: NAEP Validity Studies Panel.

Weston, T.J. (2003). The Validity of Oral Accommodation Testing: NAEP Validity Studies. Washington D.C.: National Center for Education Statistics, Working Paper No. 2003-06, April 2003.

Woods, K. (2004). Deciding to Provide a Reader in Examinations for the General Certificate of Secondary Education (GCSE): Questions about Validity and Inclusion. British Journal of Special Education, 31 (3), 122–124.

Zurcher, R. & Bryant, D.P. (2001). The validity and comparability of entrance examination scores after accommodations are made for students with LD. Journal of Learning Disabilities, 34 (5), 462–471.