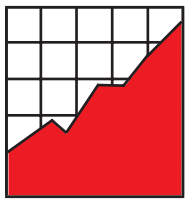
Technical Report 37

Improving Validity of Large-scale Tests: Universal Design and Student Performance



NATIONAL CENTER ON EDUCATIONAL OUTCOMES

In collaboration with:

Council of Chief State School Officers (CCSSO)

National Association of State Directors of Special Education (NASDSE)

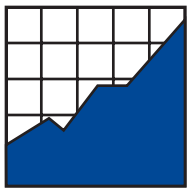
Technical Report 37

Improving Validity of Large-scale Tests: Universal Design and Student Performance

Christopher J. Johnstone

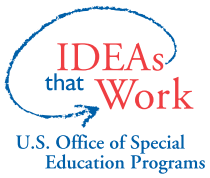
December 2003

Johnstone, C. J. (2003). *Improving validity of large-scale tests: Universal design and student performance* (Technical Report 37). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.



NATIONAL CENTER ON EDUCATIONAL OUTCOMES

This research was funded through the United States Department of Education, Office of Special Education Programs (Student Initiated Grants Award #H324B020025). All research took place under the supervision of staff at the National Center on Educational Outcomes, which is supported through a through a Cooperative Agreement (#H326G000001) with the Research to Practice Division, Office of Special Education Programs, U.S. Department of Education. Opinions expressed herein do not necessarily reflect those of the U.S. Department of Education or Offices within it.



NCEO Core Staff

Deb A. Albus
Mike Anderson
Ann T. Clapper
Jane L. Krentz
Kristi K. Liu
Jane E. Minnema
Ross Moen
Michael L. Moore
Rachel F. Quenemoen
Dorene L. Scott
Sandra J. Thompson
Martha L. Thurlow, Director

National Center on Educational Outcomes
University of Minnesota • 350 Elliott Hall
75 East River Road • Minneapolis, MN 55455
Phone 612/624-8561 • Fax 612/624-0879
http://education.umn.edu/NCEO

Executive Summary

This paper reports the theoretical background and research results for a study conducted using Universal Design of assessment features. A sample of 231 sixth grade students from traditionally under-performing schools and populations took two tests (groups were randomly assigned to order of tests taken). One test (traditionally designed) was drawn from released large-scale assessment items and presented in standard format. The second test was created using the constructs of the traditionally designed test, but included features of Universal Design elements, as explained by Thompson, Johnstone, and Thurlow (2002). One-to-one correspondence of item constructs was determined by a content area expert. Results of this experimental research demonstrated that students scored significantly higher on the universally designed test. Post-test interview data confirmed that students perceived that they scored higher on the universally designed test and preferred Universal Design features. Findings have implications for the validity of testing students with disabilities and English language learners.

Acknowledgements

The author extends his sincere thanks to all of the individuals who assisted in the planning, research, and reporting stages of this project. The unique perspectives of all involved helped improve this project dramatically. First and foremost, thanks are extended to Martha Thurlow for supervising and supporting the project from its inception. Such a project would not be possible without her encouragement.

Second, the author wishes to thank Sandy Thompson and Dave Malouf for their support, interest, and expertise in the area of Universal Design. Although this concept is relatively new to the field of education, Sandy and Dave already have much to offer others in terms of background knowledge and thoughtful advice.

Third, the local Advisory Board made up of Tine Hayes, Verna Jim, Debbie Johnson, and Glenda Lopez enriched the research. This board's unique perspectives were essential in understanding the cultural and disability-related dimensions of assessment. Such perspectives were bolstered by the mathematical construct knowledge of Donald Opitz.

While conducting studies, challenges often arise. The author wishes to thank Edward Monaghan and Karen Barton for their assistance in overcoming logistical difficulties. Additionally, sincere thanks are extended to all the principals, staff, and students who participated in this study. Taking time out of your busy school schedule for an outside researcher was challenging, to be sure, but handled with absolute professionalism. Although the names of the schools and staff must remain anonymous for confidentiality purposes, those reading this report should know that their efforts have not gone unnoticed.

Last, the final product would not be possible without the graphic talents of Michael Moore, who can arrange documents skillfully even under the tightest of deadlines.

Table of Contents

Overview

The field of education in the new millennium has witnessed an upsurge of attention focused on large-scale assessments. Such attention is found both in scholarly journals and the popular press. Much of the attention paid to large-scale assessment in the popular press relates to the high-stakes nature of some large-scale tests and the perceived deleterious effects such stakes have on schools (Henriques, 2003). In academic literature, questions have been raised about whether the administration, procedures, and format of these large-scale assessments provide optimal conditions for demonstrating achievement of academic content standards (Hanson, 1997).

Specifically, standard administrations of norm- and criterion-referenced tests have been challenged for their ability to appropriately assess particular populations of students, including those from diverse cultural backgrounds, those who are English language learners, and those with disabilities (Abedi, Leon, & Mirocha, 2001). These challenges are particularly germane to the current era of educational reform when states are required to include all students in their assessment systems (Thurlow, Quenemoen, Thompson, & Lehr, 2001).

The inclusion of all students in accountability systems is a challenge to schools, but also a challenge to the assessment community to ensure that test performance by students with disabilities and English language learners is a valid and reliable measure of their knowledge and skills. One method of increasing reliability and validity in assessments is to examine what may cause variance in student scores not related to what is being tested (a test item's construct) (AERA, APA, NCME, 1999). Information that produces variance in scores not based on constructs is what Messick (1994) called "construct irrelevant variance." Analysis of construct irrelevant variance helps determine what type of knowledge (i.e., knowledge outside of the domain of the item) is necessary in order to demonstrate proficiency on a particular item.

The potential for construct-irrelevant bias is present in all items, but particular themes have emerged in assessment research related to this issue. One type of construct irrelevant variance is vocabulary (unnecessary to the construct) that confuses students. When students cannot glean the processes necessary to demonstrate knowledge from an item itself, then the assessment is unlikely to validly assess that student's level of skill.

Related to confusing language, the readability of a particular item may unduly penalize certain students. Because English language reading is only one of many subject areas assessed in school, students who do not read well are unfairly disadvantaged in all subject areas when tests are linguistically complex (Hanson, Hayes, Schriver, LeMahieu, & Brown, 1998). For example, Rakow and Gee (1987) argued that some students may be invalidly assessed on linguistically complex science assessments. The authors noted that "Unless you have worded your test items so students are sure to understand what you are asking them, you may be challenging their reading ability rather than their grasp of scientific concepts" (p. 28).

Another type of construct-irrelevant knowledge that is sometimes necessary to demonstrate proficiency on test items is cultural knowledge. Heubert (1999), for example, warned that words and expressions closely associated with particular cultures or locations may give some groups an unfair advantage over others.

Additional item features that may compromise the validity of tests are the instructions given to students. Instructions that are complex, or in a language that students cannot understand, for example, do not provide students a valid opportunity to demonstrate knowledge (ADDA, 2001; Elliott, 1999; Willingham, Ragosta, Bennett, Braun, Rock, & Powers, 1988). Furthermore, when instructions are vague or misleading, test proctors may direct students in ways that are not standardized. Such practice invalidates tests (AERA, APA, NCME, 1999).

Construct irrelevant bias may always occur, but can be addressed in the early phases of test and item design, by reviewing research about comprehensibility and usability, and by conducting careful bias reviews and pilot tests (Kopriva, 2000). Research throughout the 1980s and 1990s demonstrated that designing tests for diverse users affects how such users perform. For example, in an effort to remedy language issues, Gaster and Clarke (1995) and Kiplinger, Haug, and Abedi (2000) demonstrated that simplified language (sometimes called Plain English) greatly increases chances of success on assessment for English language learners.

In their work with students with learning disabilities, Grise, Beattie, and Algozzine (1982) found that providing one example for each new skill tested; staggering right margins of text; and providing enlarged, vertical, elliptical 'bubbles' for student response had a positive effect on student achievement results for statewide tests. Grise et al. and the researchers mentioned above each made substantial contributions to the field of test design and its effects on diverse populations. However, each report mentioned strategies that were effective for particular populations. To address increased demands for testing of *all* students, a comprehensive review of literature was conducted to synthesize results from several fields and to propose a succinct set of design features that could improve assessment validity for all students.

Universal Design: A Brief History

In 2002, the National Center on Educational Outcomes (NCEO) synthesized research from various fields to comprise a list of assessment elements that may improve assessment for all students (Thompson, Johnstone, & Thurlow, 2002). Although this endeavor was based largely on literature from special education and English language learner sources, the inclusive nature of the research indicated that a more broad-based term was needed to describe the improvements in assessment proposed.

NCEO's attempt to identify ways to improve assessment for all students was inspired by the

traditions of an unrelated field—architecture—in its attempt to broaden the focus of practitioners and policymakers. Twenty years before, the field of architecture experienced a dramatic change in how it viewed design when Ron Mace (an architect who was a wheelchair user) actively promoted a concept he termed "Universal Design." Mace was adamant that structures did not need special purpose designs that serve primarily to meet compliance codes but may also stigmatize people. Instead, he promoted design that works for most people, from young children who cannot reach doorknobs to people with low vision who need extra natural light. Universal design was defined by the Center for Universal Design (1997) as "the design of products and environments to be usable by all people, to the greatest extent possible, without the need for adaptation or specialized design" (Universal Design, ¶1). The over-riding concept of Universal Design is to create spaces where everyone can both access and benefit from use. North Carolina State University's Center on Universal Design has published seven general principles that guide architects, product designers, engineers, and environmental design researchers in their work. Each principle is listed and described in Table 1.

While the principles in Table 1 provide a framework for how to make structures more inclusive, they were created with the caution that even Universally Designed areas may be inaccessible to some and may need additional changes to be fully inclusive. Mace (1998) verified this in his statement: "I'm not sure it's possible to create anything that's universally usable. It's not that there's a weakness in the term. We use that term because it's the most descriptive of what the goal is" (What is Universal Design? ¶3). The groundbreaking (yet realistic) set of Principles established by the Center for Universal Design provide a framework for researchers, policymakers, and practitioners in the field of assessment to consider the opportunities and constraints of Universal Design.

The application of philosophies from one field to another is never an exact transfer. Rather, general ideas are often manipulated by the receiving field in order to facilitate useful recommendations. Nevertheless, a clear connection exists between Universal Design of architecture and Universal Design of assessments. As stated above, the field of architecture had, before Universal Design, an egregious history of building a structure, then only later retrofitting it to be more inclusive when forced by law. The assessment community has followed suit in its practices, by often creating tests that need to be changed for students with diverse linguistic backgrounds or students with disabilities.

Changes made to tests, or accommodations, are extremely controversial. Testing accommodations are defined by Tindal and Fuchs (1999) as "changes in standardized assessment conditions introduced to level the playing field for students by removing the construct-irrelevant variance created by their disabilities. Valid accommodations produce scores for students with disabilities that measure the same attributes as standard assessments measured in non-disabled individuals." (p.8).

Table 1. Principles of Universal Design

Principle One: Equitable Use: The design is useful and marketable to people with diverse abilities.
1a. Provide the same means of use for all users: identical whenever possible; equivalent when not.
1b. Avoid segregating or stigmatizing any users.
1c. Provisions for privacy, security, and safety should be equally available to all users.
1d. Make the design appealing to all users.

Principle Two: Flexibility in Use: The design accommodates a wide range of individual preferences and abilities.
2a. Provide choice in methods of use.
2b. Accommodate right- or left-handed access and use.
2c. Facilitate the user's accuracy and precision.
2d. Provide adaptability to the user's pace.

Principle Three: Simple and Intuitive Use: Use of the design is easy to understand, regardless of the user's experience, knowledge, language skills, or current concentration level.
3a. Eliminate unnecessary complexity.
3b. Be consistent with user expectations and intuition.
3c. Accommodate a wide range of literacy and language skills.
3d. Arrange information consistent with its importance.
3e. Provide effective prompting and feedback during and after task completion.

Principle Four: Perceptible Information: The design communicates necessary information effectively to the user, regardless of ambient conditions or the user's sensory abilities.
4a. Use different modes (pictorial, verbal, tactile) for redundant presentation of essential information.
4b. Provide adequate contrast between essential information and its surroundings.
4c. Maximize "legibility" of essential information.
4d. Differentiate elements in ways that can be described (i.e., make it easy to give instructions or directions).
4e. Provide compatibility with a variety of techniques or devices used by people with sensory limitations.

Principle Five: Tolerance for Error: The design minimizes hazards and the adverse consequences of accidental or unintended actions.
5a. Arrange elements to minimize hazards and errors: most used elements, most accessible; hazardous elements eliminated, isolated, or shielded.
5b. Provide warnings of hazards and errors.
5c. Provide fail safe features.
5d. Discourage unconscious action in tasks that require vigilance.

Principle Six: Low Physical Effort: The design can be used efficiently and comfortably and with a minimum of fatigue.
6a. Allow user to maintain a neutral body position.
6b. Use reasonable operating forces.
6c. Minimize repetitive actions.
6d. Minimize sustained physical effort.

Principle Seven: Size and Space for Approach and Use: Appropriate size and space is provided for approach, reach, manipulation, and use regardless of user's body size, posture, or mobility.
7a. Provide a clear line of sight to important elements for any seated or standing user.
7b. Make reach to all components comfortable for any seated or standing user.
7c. Accommodate variations in hand and grip size.
7d. Provide adequate space for the use of assistive devices or personal assistance.

Source: The Center for Universal Design, North Carolina State University, 1997.

A theoretical assumption in the field is that accommodations will validate student test results. This assumption is hinged on the premise that the design of the test was such that a student's disability or language ability would hinder her or his ability to demonstrate knowledge. This being the case, accommodations would then, as Tindal and Fuchs noted, "level the playing field." Bielinski, Thurlow, Ysseldyke, Friedebach, and Friedebach (2001) found, however, that the use of accommodations may alter a test to an extent that the accommodated and un-accommodated items are no longer comparable.

A challenging situation then arises about how to create equal opportunities for success on large-scale tests without changing constructs so much that student results are not comparable. Returning to architecture, Mace described similar issues when he first advocated the idea of Universal Design. Were Ron Mace a test designer, he may have suggested that tests be designed better from the start, rather than having to be retrofitted for individual users. Such was the philosophy adopted by NCEO when exploring Universal Design of assessments

Universal Design of Assessments

The perplexing question of how to make assessments more accessible to all students required a vast literature review that relied heavily on information from psychometric, special education and English language learner sources, but also looked outside of assessment-related literature for answers. Information from the fields of literacy, vision, and graphic design demonstrates that the way in which printed documents, instruments and materials are designed have great impact on how people perceive, interact, and perform on them.

Universally designed assessments, as described by Thompson, Johnstone, and Thurlow (2002) are "designed and developed from the beginning to allow participation of the widest possible range of students, and to result in valid inferences about performance for all students who participate in the assessment. Universally designed assessments add a dimension of fairness to the testing process" (p. 5). According to the Heubert (1999), such fairness and front-end design considerations must be continued throughout the assessment process from design to administration to scoring and interpretation.

The philosophical background for Universal Design of assessments is found in the Standards for Educational and Psychological Testing (AERA, APA, NCME, 1999), which state that "all examinees be given a comparable opportunity to demonstrate their standing on the construct(s) the test is intended to measure. Just treatment also includes such factors as appropriate testing conditions and equal opportunity to become familiar with the test format, practice materials, and so forth... Fairness also requires that all examinees be afforded appropriate testing conditions" (p. 74, cited in Thompson, Johnstone, & Thurlow, 2002).

Similar to the Center on Universal Design's Principles, NCEO's Thompson, Johnstone, and Thurlow (2002) developed a structure for guiding policy and practice of assessment called elements. The authors' Elements of Universally Designed Assessments are compared to the Center on Universal Design's (1997) Principles in Table 2.

Table 2. Relationship Between Principles of Universal Design and Elements of Universally Designed Assessments

Universal Design Principle	Elements of Universally Designed Assessments
Equitable Use – design is useful and marketable to people with diverse abilities.	Reflected in all elements.
Flexibility in Use – design accommodates a wide range of individual preferences and abilities.	Especially reflected in elements #1 (inclusive assessment population), #3 (accessible, non-biased items), #4 (amenable to accommodations), and #6 (maximum readability and comprehensibility).
Simple and Intuitive Use – design is easy to understand, regardless of user's experience, knowledge, language skills, or current concentration level.	Especially reflected in elements #5 (simple, clear, intuitive instructions and procedures), #6 (maximum readability and comprehensibility), and #7 (maximum legibility).
Perceptible Information – design communicates necessary information effectively to the user, regardless of ambient conditions or the user's sensory abilities.	Especially reflected in elements #4 (amenable to accommodations), #5 (simple, clear, intuitive instructions and procedures), and #7 (maximum legibility).
Tolerance for Error – design can be used efficiently and comfortably and with a minimum of fatigue.	Reflected in elements #2 (precisely defined constructs) and #5 (simple, clear, intuitive instructions and procedures).
Low Physical Effort – design can be used efficiently and comfortably and with a minimum of fatigue.	Primarily reflected in element #7 (maximum legibility).
Size and Space for Approach and Use – appropriate size and space is provided for approach, reach, manipulation, and use regardless of user's body size, posture, or mobility.	Primarily reflected in elements #4 (amenable to accommodations), and #7 (maximum legibility).

Source: Thompson, Johnstone, and Thurlow (2002).

The elements are drawn from best practices that are accepted in the field of measurement but demonstrate a succinct collection of recommendations that relate to existing best practices in architecture. As noted by Thompson, Johnstone, and Thurlow (2002), the elements are likely to undergo changes as research increases the knowledge base of the field. Such changes are natural because, according to the Heubert (1999), "research and development in the field of educational testing is continually experimenting with new modes, formats, and technologies" (p. 202). A description of each element of Universally Designed assessments follows.

Element #1. Inclusive Assessment Population: NCEO's first element replicates the inclusive nature of the Center for Universal Design's principle of equitable use. In the case of large-scale assessments, current legislation dictates that all students need to be included in statewide accountability systems. Such inclusion ensures that students will have equal access to the public and private benefits conferred by schooling.

The American Educational Research Association, American Psychological Association, National Council on Measurement in Education (1999), and National Research Council (see Heubert, 1999) called for tests that are conceptualized for the contexts of the entire population to be assessed. In practice, this translates into piloting, field testing, and norming tests with the explicit goal of including students with a wide range of disabilities, students with limited English proficiency, and students across racial, ethnic, and socioeconomic lines. When doing so, constructs and standards should be maintained, but design features selected should reflect end users.

Element #2. Precisely Defined Constructs: Fidelity to constructs is important for two reasons. First, when test designers remain true to constructs, they are able to remove all non-construct oriented sensory, emotional, and physical barriers to students demonstrating competence in a particular area. By removing such barriers, states are better able to determine where needs for educational improvement lie. Constructs that are clearly defined are essential for making sound and valid educational decisions based on assessment results.

Examples to the contrary demonstrate the continued need for explicitly labeling constructs to be tested. In the area of mathematics, research has demonstrated that students are often limited in their ability to demonstrate mathematical knowledge because of the reading level of items. Calhoun, Fuchs, and Hamlett (2000), Harker and Feldt, (1993), Koretz (1997), and Tindal, Heath, Hollenbeck, Almond, and Harniss (1998) all found that students with reading difficulties scored *higher* on tested constructs in math when questions were read to them than when they read tests independently (all students in these studies had reading difficulties). Shorrocks-Taylor and Hargreaves (1999) suggested that the language used in questions on tests that assess subjects other than language become as "transparent" as possible to reduce construct irrelevant bias, or the "degree to which test scores are affected by processes that are extraneous to its intended construct" (AERA, APA, NCME, 1999).

Element #3. Accessible, Non-Biased Items: A biased item is one that is presented in a way that disadvantages particular test takers. Popham and Lindheim (1980) suggested that items can minimize unfair advantage to particular students by:

- Ensuring that required student responses are a valid demonstration that students have mastered content (*curricular congruence).*

- Ensuring the likelihood that students will have had the opportunity to learn the desired outcome of the item (*instructional sensitivity).*

- Ensuring that the socioeconomic status or inherited academic aptitudes the student possesses are not the dominant influence on how the student will respond to an item (*out of school factors).*

- Ensuring that the item's content does not offend or unfairly penalize students because of personal characteristics such as race, gender, ethnicity, [disability,] or socioeconomic status *(bias).*

Bias in items can be examined both qualitatively and quantitatively. Geisinger and Carlson (1992), Popham (2001), and Popham and Lindheim (1980) all suggested that sensitivity review panels are utilized to determine whether particular items are biased against a particular group. Such panels consist of representatives from various ethnic and socioeconomic groups. In addition to such groups, sensitivity panels are strengthened by members that represent persons with disabilities and English language learners.

Bias can also be examined using *post hoc* analyses such as Differential Item Functioning (DIF) and Item Response Theory (IRT). Differential Item Functioning tests determine whether students with equal ability but representing different groups do not have the same probability of responding correctly to test items. DIF analysis has traditionally been used to detect the differential function of an item according to group identity (e.g., race, gender, disability). IRT is used to predict a student's probability of answering an item correctly or incorrectly based on particular student characteristics.

Element #4. Amenable to Accommodations: Mace (1998) clarified that in Universal Design there still may be a need for post-design retrofitting to meet the needs of particular consumers. Likewise, while Universal Design of assessments may reduce the need for accommodations, it will never fully eliminate the need to make certain changes to the standard administration of tests. One example is the use of Braille tests. While Braille tests have obvious text differences, certain design features of the standard test improve the correspondence of Braille and standard tests. Features that improve correspondence include:

- Avoiding the use of construct irrelevant graphs or pictures.

- Avoiding the use of vertical or diagonal text.

- Not placing keys and legends at the left or bottom of the item where they are more difficult to locate in Braille formats.

- Avoiding items that depend on reading of graphic representations (such as blueprints, fur-

niture in a room) that do not also have verbal/textual descriptions that can be translated into Braille.

- Removing items that include distracting or purely decorative pictures, which draw attention away from the item content.

Furthermore, a test is amenable to accommodations if it has "built in" accommodations for all students. For example, two common accommodations for students are extended time on tests and the use of manipulatives. If a test is not a speeded test, then time limits can be removed for all students. Second, if the construct of the test does not call for manipulative-free computation, then all students can be given the opportunity to use manipulatives. The above-mentioned examples demonstrate how different elements of Universal Design at times overlap to create more accessible testing for all.

Element #5. Simple, Clear, and Intuitive Instructions and Procedures: Tests that attempt to gauge student knowledge or performance on a particular task are invalidated when students are unclear of the tasks they must perform to demonstrate knowledge (ADDA, 2001; Elliott, 1999; Willingham, et al., 1988). Clarity of instructions is improved when sample items, examples, and criteria for scoring are made clear to students from the very beginning (AERA, APA, NCME, 1999; Grise, Beattie, & Algozzine, 1982). Tindal and Fuchs (1999) suggested that test designers should question whether it will "be possible for all students to work independently throughout the test?" And "are directions easy to follow?" as preliminary guides for determining whether instructions are simple, clear, and intuitive.

Element #6. Maximum Readability and Comprehensibility: Readability is a much-disputed term because formulae are often calculated to quantify the readability of a text or passage without considering the qualitative elements of that passage. Rakow and Gee (1987) treaded carefully on the description of readability, calling it an "estimate of probability of comprehension by a particular group" (p. 28). According to such estimates, readability generally increases when sentence lengths are reduced, words per line are minimized, and multi-syllabic words are used sparingly. While such estimates are generally reliable for texts and passages, they are less so for test items with only one or two sentences. In the case of items, Popham and Lindheim (1980) and Rakow and Gee (1987) suggested concentrating on features of text such as logical organization of ideas and clarity of message. These features are difficult to quantify, but Rakow and Gee (1987) suggested determining students' prior experiences, achievement, and interests as well as text features when determining readability of a passage.

Gaster and Clark (1995) recommended eight considerations when attempting to improve readability. These were:

- Use simple, clear, commonly used words, eliminating any unnecessary words.

- When technical terms must be used, they should be clearly defined.
- Compound complex sentences should be broken down into several short sentences, stating the most important ideas first.
- Introduce one idea, fact, or process at a time; then develop the ideas logically.
- All noun-pronoun relationships should be made clear.
- When time and setting are important to the sentence, place them at the beginning of the sentence.
- When presenting instructions, sequence steps in the exact order of occurrence.
- If processes are being described, they should be simply illustrated, labeled, and placed close to the text they support.

Many of Gaster and Clark's recommendations mirror the philosophies of the recent "Plain language Movement." Plain Language approaches have been applied to testing situations by Brown (1999) in an effort to reduce construct irrelevant variance based on student reading ability. Brown suggested that editors: reduce excessive length, eliminate unusual or low frequency words, avoid ambiguous words, avoid irregularly spelled words, avoid unclear signals about how to direct attention, and mark all questions clearly.

Element #7. Maximum Legibility: Legibility refers to the capacity with which items are able to be deciphered with ease. In terms of testing, three general categories of legibility are necessary to lead to "maximum" legibility. The first category is the legibility of text. Table 3 demonstrates characteristics of legible text culled from research on vision, graphic design, and literacy. Each dimension of legible text (contrast, type size, spacing, leading, type face, justification, line length, and blank space) is defined and examples are provided to illustrate characteristics.

The second category of legible text refers to the comprehensibility of graphs, tables, and illustrations. Best practices in graphic design guide suggestions for improving the legibility of visual materials on tests. For example, Gregory and Poulton (1970) suggested information is more quickly found on graphs when plot lines and axes are clearly labeled. Shriver (1997) concurs, noting that clearly-designed quantitative graphic information creates a context for appropriate interpretation of data.

Criteria of utility and design lead to the appropriate use of illustrations in text. Sharrocks-Taylor and Hargreaves (1999) found that there are three different types of illustrations found in tests:

Table 3. Characteristics of Legible Type

Dimension	Characteristics of Legible Text
Contrast (degree of separation of tones in print from the background paper)	White or glossy paper should be avoided to reduce glare (Menlove & Hammond, 1998). Blue paper should not be used. Black type on matte pastel or off-white paper is most favorable for both contrast and eye strain (Arditi, 1999; Gaster & Clark, 1995). Avoid gray scale and shading, particularly where pertinent information is provided.
Type Size (standard measuring unit for type size is the point)	The point sizes most often used are 10 and 12 point for documents to be read by people with excellent vision reading in good light (Gaster & Clark, 1995). Fourteen point type increases readability and can increase test scores for both students with and without disabilities, compared to 12-point type (Fuchs, Fuchs, Eaton, Hamlett, Binkley, & Crouch, 2000). Large print for students with vision impairments is at least 18 point. Type size for captions, footnotes, keys, and legends need to be at least 12 point also. Larger type sizes are most effective for young students who are learning to read and for students with visual difficulties (Hoerner, Salend, & Kay, 1997). Large print is beneficial for reducing eye fatigue (Arditi, 1999). The relationship between readability and point size is also dependent on the typeface used (Gaster & Clark, 1995; Worden, 1991).
Spacing (the amount of space between each character)	Letters that are too close together are difficult for partially sighted readers. Spacing needs to be wide between both letters and words (Gaster & Clark, 1995). Fixed-space fonts seem to be more legible for some readers than proportional-spaced fonts (Gaster & Clark, 1995).
Leading (the amount of vertical space between lines of type)	Insufficient leading makes type blurry and gives the text a muddy look (Schriver, 1997). Increased leading, or white space between lines of type makes a document more readable for people with low vision (Gaster & Clark, 1995). Leading should be 25-30 percent of the point (font) size for maximum readability (Arditi, 1999). Leading alone does not make a difference in readability as much as the interaction between point size, leading and line length (Worden, 1991). Suggestions for leading in relationship to type size: • 12-point type needs between 2 and 4 points of leading. • 14-point type needs between 3 and 6 points of leading. • 16-point type needs between 4 and 6 points of leading. • 18-point type needs between 5 and 6 points of leading (Fenton, 1996)

Table 3. Characteristics of Legible Type (continued)

Typeface (characters, punctuation, and symbols that share a common design)	Standard typeface, using upper and lower case, is more readable than italic, slanted, small caps, or all caps (Tinker, 1963). Avoid font styles that are decorative or cursive. Standard serif or sans serif fonts with easily recognizable characters are recommended. Text printed completely in capital letters is less legible than text printed completely in lower-case, or normal mixed-case text (Carter, Dey & Meggs, 1985) Italic is far less legible and is read considerably more slowly than regular lower case (Worden, 1991). Boldface is more visible than lower case if a change from the norm is needed (Hartley, 1985).
Justification (text is either flush with left or right margins – justified – or staggered/ ragged – unjustified)	Staggered right margins are easier to see and scan than uniform or block style right justified margins (Arditi, 1999; Grise et al., 1982; Menlove & Hammond, 1998). Justified text is more difficult to read than unjustified text – especially for poor readers (Gregory & Poulton, 1970; Zachrisson, 1965). Justified text is also more disruptive for good readers (Muncer, Gorman, Gorman, & Bibel, 1986). A flush left/ragged right margin is the most effective format for text memory. (Thompson, 1991). Unjustified text may be easier for poorer readers to understand because the uneven eye movements created in justified text can interrupt reading (Gregory & Poulton, 1970; Hartley, 1985; Muncer, Gorman, Gorman, & Bibel, 1986; Schriver, 1997). Justified lines require the distances between words to be varied. In very narrow columns, not only are there extra wide spaces between words, but also between letters within the words (Gregory & Poulton, 1970).
Line Length (length of the line of text; the distance between the left and right margin)	Longer lines, in general, require larger type and more leading (Schriver, 1997). Optimal length is 24 picas - about 4 inches (Worden, 1991). Lines that are too long make readers weary and may also cause difficulty in locating the beginning of the next line, causing readers to lose their place (Schriver, 1997; Tinker, 1963). Lines of text should be about 40-70 characters, or roughly eight to twelve words per line (Heines, 1984; Osborne, 2001; Schriver, 1997).
Blank Space (Space on a page that is not occupied by text or graphics)	Use the term “blank space” rather than “white space” because the background is not always white (Schriver, 1997). Blank space anchors text on the paper (Menlove & Hammond, 1998). Blank space around paragraphs and between columns of type helps increase legibility (Smith & McCombs, 1971) A general rule is to allow text to occupy only about half of a page (Tinker, 1963). Too many test items per page can make items difficult to read.

Source: Thompson, Johnstone, and Thurlow, 2002.

- Decorative illustrations that are not related to the questions and serve no instructional purpose.
- Related illustrations that have the same context as the questions and are used to support text and emphasize ideas.
- Essential illustrations that are not repeated in the text, but the text refers to them, and they have to be read, or worked with, to answer the question.

Because certain students are easily distracted or may be led astray by decorative illustrations, such illustrations may create barriers to success and should therefore be removed from tests. Silver (1994) and West (1997) recommended that when related and essential illustrations are used, they be placed directly next to the question to which they refer. Similar principles apply to computer-generated images. Szabo and Kanuka (1998) suggested that computer graphics should comply with principles of unity, focal points, and balance in order to increase completion rates on tests.

Finally, even when useful to the item and well-designed, illustrations are most effective when they match the cultural schema of the end user. Not only should graphics be culturally appropriate in terms of experience (Shriver, 1997), they should also be communicated respectfully, considering the student's cultural norms and belief systems (Schiffman, 1995).

The final category of legibility is the legibility of response formats. Response formats are the portions of tests where students are asked to respond to items. In many norm-referenced tests, students are asked to respond on computer-ready bubble sheets. Such formats are inherently biased against students with fine motor difficulties or low vision. Thus, marking answers directly on the page is recommended for both people with low vision and Braille users. Larger circles on which to "bubble" are recommended for all users. As Willingham et al. (1988) noted, many students with learning disabilities have a lack of body awareness and poor directionality. Larger bubbles (or answering directly on test) could help to mitigate some of these issues.

Research in the area of response formats continues today, but several trends have emerged over the past 20 years. Grise et al. (1982) found that flattened, horizontal bubbles were more useful for students with learning disabilities than circular bubbles. Rogers (1983) and Tindal et al. (1998) both found that separate answer sheets were problematic for students in grades 1-3. Results from grades 4-6, however, were mixed. Tindal et al. (1998) found no difference in scores for grade 4 students using separate answer sheets. Rogers (1983) determined that grades 4 and 5 students could use separate answer sheets effectively, provided they were given special instructions. Veit and Scruggs (1986), however, found that fourth grade students with learning disabilities took significantly more time to complete tests when bubbling was required. Muller, Calhoun, and Orling (1972) determined that students in grades 3-6 made fewer errors when allowed to answer

directly on answer sheets. Overall, separate answer sheets are a practice that speed scoring time and decrease costs of large-scale tests. There is no research, however, demonstrating that separate answer sheets are superior to answering questions directly on the test.

When the legibility of text, graphic information, and response formats are improved, the overall legibility of the test increases. To do this, format and font as well as graphics and illustrations deserve close attention. Furthermore, the way students are asked to respond must not be taken for granted. The combination of all elements provides a framework for examining tests and their level of accessibility. Individual items may not have direct links to all elements, but Universal Design of assessments is presented here as a general framework for improving the design (and thereby accessibility, comprehensibility, and validity) of tests.

Although Universal Design of assessments is a relatively recent addition to assessment literature, policymakers at the state and national level have begun to mention the philosophy of Universal Design or its elements. References to Universal Design in testing can be found in the State of California's educational policy, which requires educators to employ strategies of "Universal Design" of education in both classrooms and assessments (State of California, 2003) and the State of Vermont's educational assessment Request for Proposal (RFP), which calls for "Universally Designed Assessments" using Thompson et al.'s Elements of Universally Designed Assessments (State of Vermont, 2003).

Every new idea that will be employed in policy and practice carries a burden of proof. Specific and scientific research must be documented on all positions proposed in the field of education. A recent study, funded by the United States Department of Education, Office of Special Education Programs aimed to determine the efficacy of Universal Design. Although Universal Design literature and funding have a specific leaning toward disability issues, Thompson, Johnstone, and Thurlow (2002) stated that Universal Design could improve assessment for every student. Such a proposition was put to the test in United States Department of Education Project # H324B020025, which was a student-initiated award to conduct research on this issue.

Research Methodology

A mixed-method design was chosen to test hypotheses and gain a greater understanding of test-related issues for students. The methods, which combine experimental and descriptive techniques, asked the research questions:

1. Is there an overall difference in test performance between traditionally and universally designed tests?

2. Does the relationship in performance on traditionally and universally designed tests differ between students with and without disabilities?

3. Do students perceive differences in the ease in which they completed the traditionally designed and universally designed tests?

The purpose of the project was to determine whether eliminating construct irrelevant material in tests would provide a more accurate measure of student performance. By eliminating superfluous, non-construct relevant information and requirements, a universally designed test was examined to determine whether it was more accessible to students with disabilities and provided a clearer assessment picture of all students, including those who do not qualify for special accommodations (students with mild disabilities, low performing regular education students who have "fallen through the [service] cracks" and students with limited English proficiency).

Overview of Research

Typically, test scores for students with low socioeconomic status, English language learners, and students in special education programs have been lower than national averages. Factors such as poor teaching, poverty (Willems, 1986), and limited English proficiency have all been suggested as reasons for poor performance. While personal reasons for student failures are often investigated, the design of tests has rarely been scrutinized as a reason for student underperformance.

To address the research questions, the proposed research project was divided into two studies. Study One was a comparison study of results from a cohort of students on two tests. The first (control) test was a practice mathematics test comprised of actual (released) statewide assessment items. The second (experimental) test was a revision of the same test, adjusted to incorporate Universal Design principles. This test was created by a team of experts. Four community members served on this team. Three of the participants were parents of students in special education programming in the school district. The fourth was a teacher in the district who has dyslexia. Each was considered to have unique perspectives that were solicited for the project. Each parent represented one of the major ethnic groups of the area.

All experimental items were generated from the constructs of the control test items. While constructs remained constant in the experimental test, design and administration features were changed in accordance with Universal Design principles. Students with disabilities were examined on both tests according to their IEP guidelines (i.e., with the accommodations indicated in their IEPs).

The second study was a qualitative analysis of the difference in student experiences in the two

versions of the test. Students included in the second study were those who demonstrated a 1.5 standard deviation improvement from control to experimental tests. These students were interviewed about their perceptions of the two tests.

Participants

As stated above, the purpose of this study was to determine whether the principles of universally designed assessment would more validly assess a sample of all students, and to specifically examine students who were economically poor, mostly English language learners, ethnic minorities, and students with disabilities. The sample was taken from four schools in the U.S. Southwest. Two of the schools were in a small town of approximately 20,000 and two schools were in rural areas. All of the schools were adjacent to Native American Tribal Land.

In total, 231 students participated in the study. Demographic data were missing for 18 subjects. Of the data available, 165 were Native American, and 25 were Latino/a, and 23 were Anglo/a/white. Thirty-one students had specific learning disabilities, 109 were English language learners, and 132 were reading below grade level (according to local assessments).

Study 1: Comparison of Traditional and Universally Designed Assessments

Study 1 was divided into two parts, each part using counterbalancing methodology to ensure valid experimentation. Each student in the sample completed two tests, one traditionally designed and one universally designed. Half of the cohort took the traditionally designed test (made up of released statewide assessment items) according to standard testing procedures. Students with IEP accommodations took the test with all specified accommodations. Later, this group took a second test redesigned using Universal Design principles. The other half of the cohort took the tests in reverse (universally designed first and traditionally designed second) to prevent a "practice effect" (Edelman, 1996) and inflation of scores. To maintain inter-test reliability, all items were drawn from other (released) statewide test items or created by a team of experts using the process described below. Table 4 demonstrates the testing schedule.

The creation of the universally designed test followed a systematic procedure that was created for this research project. Such a procedure may be useful to test-designers when attempting to include Universal Design Elements into assessments. The step-wise procedure was performed as follows:

1. Items were selected for revision from released statewide tests.

2. The team of experts (referred to as the "Advisory Board") was provided information and

Table 4. Testing Schedule

Groups 1, 3, 5, 6, 9, 11, 13, 15	Groups 2, 4, 6, 8, 10, 12, 14
Test 1: Traditional Test	Test 1: Universally Designed Test
Test 2: Universally Designed Test	Test 2: Traditional Test

a training session on the Principles of Universal Design. The Advisory Board then examined the items for perceived design flaws. Design problems and suggestions for improvement were re-submitted to the researcher. The Advisory Board was asked to make changes based on their knowledge of Universal Design and their personal insights based on their child's disability, their own disability, or their cultural background.

3. A mathematics expert at a large research university examined all items on the "traditionally designed" test and marked what he perceived the construct to be measured.

4. Experts on Universal Deign in testing from NCEO examined items for design flaws and possible improvements.

5. Test items were re-designed, keeping constructs constant but removing construct-irrelevant information, changing font size, paper color, and re-arranging numeric combinations (e.g., an item that called for students to add 4,533 + 3,205 might be changed to 3,054 + 4,715).

An example of changes made to a particular item is found in Table 5. Note that some changes were superficial (such as font size) while others greatly changed the administration of the test (response formats and timing). None of the changes disrupted constructs. Original tests were not speeded but timed for administration ease.

Tests were administered to students in using the methods described above. Student scores were then analyzed to determine whether: (1) student performance was affected by test design (claims of causation are possible because experimental and control groups were present), and (2) students were affected differently by ethnicity, disability status, English language learner status, and reading ability.

Group means were calculated for both the traditionally-designed test and the universally designed test. Because the sample took both tests on the same day (to prevent sample deviation or unequal opportunities for exposure to test material), a *matched sample t-test* methodology was chosen to determine whether there was a statistically significant difference between group means. T-test results are reported in the Findings section. To boost the explanatory power of statistical results, a *Cohen's d effect size* was also calculated, with results reported in the Findings section.

Table 5. Item Re-design Process

Feature	Universal Design Element
• Inclusive Test Population (both tests featured this)	1. Inclusive Test Population
• Subject area specialist review and determination of constructs. • Removal of construct-irrelevant information.	2. Precisely Defined Constructs
• Sensitivity review for culturally, geographically, or socioeconomically irrelevant information performed by team of parents and adult with disability.	3. Accessible, non-biased items.
• All unnecessary diagrams removed from original test. • Test was un-timed.	4. Amenable to accommodations.
• Instructions simplified. • Sample problem provided. • Consistent block arrows and words "Go on" used to instruct students to move to following page.	5. Simple, clear and intuitive instructions.
• Sentences written in simple English. • Sentence length reduced. • Construct-irrelevant language removed or difficulty reduced. • Review by person with dyslexia and parents of ELL students for readability.	6. Maximum Readability

Although the sample was purposively chosen to reflect populations with high numbers of ethnic minorities, English language learner students, and students with disabilities (all who traditionally under-perform on large-scale tests), samples within schools themselves were random. Because of this, student sub-populations were unequal, making between-group statistical inferences impossible. Data were disaggregated and descriptive results for various sub-groups are presented in the Findings section of this paper.

Study 2: Student Responses to Design of Tests

Study 2 was a qualitative study, intended to give voice to students (Fetterman, 1998) concerning the traditionally and universally designed tests. The methodological intention of this study was to analyze student perceptions about traditionally and universally designed tests. Quantitative data were valuable in determining the effectiveness of different tests, but qualitative data provided perspectives as to why students are performing in particular ways. A series of interviews was conducted with students with and without disabilities in order to gather their perceptions on the two tests.

Following the second test, control and experimental data were compared. A group of 23 students was chosen and individually interviewed by the Student Researcher. Students who demonstrated a 1.5 standard deviation change in score between control and experimental tests were selected as the purposive sample for this study (Patton, 1990).

Interview questions addressed three major themes: comprehension of test material, test design, and overall large-scale testing issues. For the first and second themes, comprehension of test material and design features, students were asked a series of questions about the two tests they took. Students were provided actual protocols to review and were asked to share their perceptions of the two tests.

Questions on the third theme asked students to reflect on their experiences with large-scale tests and to describe their feelings related to such experiences. Building on these feelings, students were asked to provide advice to test makers for ways to improve testing to make it more child-friendly. A sample of all interview questions is found in Appendix A. All interviews were conducted in a structured interview format (Bogdan & Biklen, 1992) to ensure that prompts were standardized for each student.

All interviews were transcribed and read for content. Qualitative themes arose from data. Themes (and quotes to exemplify them) are provided in the Findings section. Qualitative research, as a methodology, cannot provide data to make causal claims. Rather, in this study, qualitative themes are used to illustrate statistical information and to help the reader better understand the end users of tests whose perspectives may be different from those of educators, policymakers, or assessment designers.

Findings

Results of this research demonstrated that applying the elements of Universal Design of assessments does affect student performance on tests. Overwhelmingly, this effect is an increase in achievement. Such a finding proves that, if constructs are held constant, the design of a test can influence how a student performs. Implications of such a finding are discussed in detail in the final section of this report, and raise important issues in the current era of high-stakes testing.

Upon completion of tests, all student scores were tabulated and entered into SPSS software. A simple sign test demonstrated that 155 students of the 231 sampled scored higher scores on the universally designed test. Conversely, 51 students had lower scores on the universally designed tests, and 25 students had the same score on the traditionally and universally designed tests. Student demographics appeared to follow general sample demographics for both students who did and did not show “improvements” from the traditional to universally designed assessments.

Of the 155 students with improved scores, all were statistically significant, yet only 17 students had statistically significant negative scores (this is most likely an effect of the relatively small number of decreased scores).

Next, mean scores were tabulated for subgroups. Target subgroups were those mentioned in No Child Left Behind guidelines and those of particular interest to the researcher. Subgroup means that were tabulated were: Navajo students; Latino/a students; White, non-Hispanic students; students with disabilities; non-disabled students; English language learner students; English language proficient students; students who read at or above grade level; students who read below grade level; and students who are both English language learners and have disabilities. Because students often belonged to more than one subgroup, statistical inferences were impossible. Every subgroup demonstrated improved performance on the Universally Designed test. Mean scores for both tests as well as standard deviations are presented below.

Table 6. Subgroup Results on Traditionally and Universally Designed Tests

Sub-group	n	Mean Score: Traditional Test	Standard Deviation	Mean Score: Universally Designed Test	Standard Deviation	P-value (alpha = .05)**
Total Sample	231	6.32	3.29	7.59	3.44	.000
Navajo	165	5.56	2.69	6.98	3.15	.000
Latino/a	25	7.08	3.17	7.72	3.61	.000
Anglo/a	23	10.70	3.66	11.70	3.13	.005
Students with Disabilities	31	5.03	2.64	6.23	3.65	.000
Non-disabled Students	175	6.65	3.32	7.94	3.43	.000
English Language Learners	109	5.20	2.53	6.43	2.94	.000
English Language Proficient Students	104	7.43	3.47	8.79	3.61	.000
Students Reading Below Grade Level	132	5.45	2.40	6.69	2.98	.000

*Demographic information is missing for 18 subjects.
**P-value is a biased estimator for subgroups because of overlap, but is reported for descriptive purposes only.

Although statistical inferences could not be made for subgroups, the overall group did not have overlap because every student took both tests. Therefore, a *matched sample t-test* was performed to determine whether the difference in means between tests was statistically significant. The mean for the traditionally designed test was 6.31 with a standard deviation of 3.29. The Universally Designed test yielded a mean of 7.59 with a standard deviation of 3.45. The difference in means produced a t-statistic of 7.466. When tested at (.05) significance on a two-tailed matched sample test, this statistic was found to be significant. Thus, there was a statistically significant difference between mean scores on traditionally and universally designed tests for this sample.

In an effort to report meaningful findings to the research community and to policymakers, an effect size was then calculated. Effect size calculations provide the reader an opportunity to understand findings on a standardized scale. *Cohen's d*, for example is calculated using the formula $d=\mu 1-\mu 2/\sigma$. Because actual population means and variances are not known, in the case of this research the formula (Sample $Mean_1$-Sample $Mean_2/s_p$) was used (Hedges, 1982). *d* statistics report in terms effect in standard deviation units.

A *Cohen's d* of .39 was found for this study, meaning that there was a roughly (.4) Standard Deviation effect on scores as a result of Universal Design. This finding represents a small to moderate effect on test scores as a result of design. Partial Standard Deviation effects may be quite meaningful when situated in educational environments in which points on tests are tied to funding, programs, and Adequate Yearly Progress.

When testing was completed and scores tabulated, 23 students were interviewed to gauge their perceptions on design issues and large-scale testing. Students were those available for interviews from the 29 students who scored 1.5 Standard Deviations higher on the Universally Designed test than the traditionally designed test (only two students scored 1.5 Standard Deviations worse on the traditional test). The interviewees were selected from an "extreme case" sample (Patton, 1990) that is thought to be illustrative when attempting to understand qualitative issues.

Themes that emerged from interviews included:

1. Recognition of Previously Taught Material
2. Readability
3. Time
4. Output-oriented comments
5. Advice for Test Makers
6. Test anxiety

Such themes were culled from student responses to questions found in Appendix A. Overall, students commented that they recognized material better (from their coursework) when it was presented using Universal Design principles. Students also found the vocabulary and print of the universally designed test more readable. Also, students noted that having unlimited time was helpful while solving problems (anecdotal reports from teachers demonstrated that students did not take a substantially longer time to finish either test). Finally, students noted their preference for answering questions directly on the answer sheet, claiming that bubble sheets are often confusing and cause students to erroneously mark answers.

When asked questions about what advice students would give to test-makers, students asked test makers to create items that are clear and concise. This may decrease the anxiety that all students in the interview sample reported. According to the students, high-stakes testing is a stressful experience that causes self-doubt and fear. A list of direct quotes related to themes is found in Table 7.

Overall, data demonstrated that Universal Design principles that are applied *en masse* to test items have a positive effect on student performance. The effect size of .39 shows an overall effect, while descriptive statistics demonstrate that all subgroups analyzed achieved at a higher level on universally designed tests than traditionally designed tests. Reasons for why the change in scores may have occurred were provided by students. Student perceptions were coded and reveal that timing, readability, and recognizable materials are most important for high achievement on tests. Both quantitative and qualitative data point to the importance of examining how test design affects both student performance and student perceptions of tests. Implications for these data are found in both policy and practice and are discussed below.

Implications

Howell (1999) noted that quantitative research can be tested for statistical significance, but should also be considered for practical significance. The practical significance of this project relates to its findings and how they are situated in current policy environments. Current No Child Left Behind requirements dictate that all schools must use large-scale assessments to monitor student progress. Furthermore, the population of students examined on such tests must include students of diverse backgrounds and abilities, including students with disabilities and English language learners.

The findings of this study demonstrate that students who typically under-perform on tests score .39 standard deviations higher on tests that are designed using Universal Design principles. This has two meaningful outcomes. First, it provides the educators with a better understanding of the capabilities of students who have diverse learning needs. Second, findings reinforce the

Table 7. Student Interview Responses by Theme

Theme 1: Recognition of Previously Taught Material
Students stated that they scored better on the universally designed test because…
"…it was everything I knew from Mr. Kenworth's* from math"
"…I understood the questions and I knew them"
"…I remembered some of the questions better…On the (statewide test) they had stuff that we didn't even go over in the classroom."

Theme 2: Readability
Students stated that they scored better on the universally designed test because…
"…the words in this one were a little easier to understand."
"…it had only five or six words in a sentence and really explained it. This one (traditional test) had longer words and it was harder to understand."
"…it didn't have much hard words."

Theme 3: Time
Students stated that they scored better on the universally designed test because…
"… I didn't have to worry about the time and I could think real hard on the test."
"…on the (traditionally designed test) I felt rushed."
"…I had more time to think and I wasn't really confused. I took my time and all that."

Theme 4: Output Oriented Comments
Students preferred answering questions directly on test because…
"…on this one (universally designed test) you could just circle but on that one you had to bubble. I worry (when I bubble) to get the wrong one."
"…I get out of line on the bubbles."
"…you didn't have to switch back and fourth."

Theme 5: Advice for Test Makers
Students advised test makers to create tests similar to the universally designed tests. Specifically, they advised test makers…
"Don't make the questions too long."
"Make the (universally designed) one I guess. Like, make them more descriptive."
"(The universally designed test)…is more understandable and you're not going to get your problems wrong if you take your time."

Theme 6: Test Anxiety
Students expressed test anxiety in various ways. Examples are:
"Scary. Because I felt like I messed up."
"I'm nervous. I'm afraid I might fail."
"I worried about getting a bad grade."
"Like I didn't want to take it. I don't like tests."

*=pseudonym

importance of ensuring that all item designers have been trained in the principles of Universal Design.

The future directions recommended after this research are based on the study's strengths and limitations. This study's greatest strength was that it demonstrated that particular design features positively affect student performance and attitudes of students toward tests. Further research should be conducted to verify these findings or to expand the definition of universal design. As such, Universal Design of assessments has the potential to be a guiding philosophy in all test design, with specific features and elements in a state of constant revision based on research results.

Further research should also continue in an effort to supplement the shortcomings of this study. One limitation of this study was that the only items addressed were multiple choice items. Multiple choice items are very common in large-scale assessments, especially on norm-referenced tests. As some states move to criterion-referenced testing, however, constructed response items also need to be examined for design features that affect student performance.

A second area of future concern is the examination of non-mathematics items. This research, like other studies before it (Calhoun, Fuchs & Hamlett, 2000; Harker & Feldt, 1993; Koretz, 1997; Tindal et al., 1998) studied mathematical items. Mathematics was chosen for this study because of the facility of determining constructs clearly so that design features can be changed. Construct fidelity in the face of design changes may be more difficult in other subject areas. Further research needs are present in the areas of language arts, social studies, and science to understand the interaction between items that are intended to be "authentic" and those that merely contain construct-irrelevant material intended to pique student interest.

Currently, most constructs are undefined by item designers. Researchers can contribute to the ongoing large-scale assessment dialogue by investigating what assessments are appropriate, authentic, and valid for the largest population possible by pushing for information on what is actually being tested in America's schools. As student participation rates grow in large-scale assessment, so must continued efforts to ensure that design features of tests are clear, accessible, and non-biased. In short, this means that tests can be better designed if they are universally designed, but also if they contain other features found in today's literature. This research has added a very small piece to the broader inclusive assessment puzzle. Further work must continue so that a complete collection of best practices in assessment of inclusive populations can emerge.

References

Abedi, J., Leon, S., & Mirocha, J. (2001). *Validity of standardized achievement tests for English language learners*. Paper presented at the American Educational Research Association Conference, Seattle, WA.

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

Arditi, A. (1999). *Making print legible.* New York: Lighthouse.

Bielinski, J., Thurlow, M., Ysseldyke, J., Friedebach, J., & Friedebach, M. (2001). *Read-aloud accommodation: Effects on multiple-choice reading and math items* (Technical Report 31). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Available at: http://education.umn.edu/NCEO/OnlinePubs/Technical31.htm

Bogdan, R.C., & Biklen, S.K. (1992). *Qualitative research for education. An introduction to theory and methods.* Allyn and Bacon: Boston.

Brown, P.J. (1999). *Findings of the 1999 plain language field test*. University of Delaware, Newark, DE: Delaware Education Research and Development Center.

Calhoun, M.B., Fuchs, L., & Hamlett, C. (2000). Effects of computer-based test accommodations on mathematics performance assessments for secondary students with learning disabilities. *Learning Disability Quarterly, 23,* 271-282.

Carter, R., Dey, B., & Meggs, P. (1985). *Typographic design: Form and communication.* New York: Van Norstrand Reinhold.

Center for Universal Design (1997). *What is universal design?* Center for Universal Design, North Carolina State University. Retrieved January, 2002, from the World Wide Web: http://www.design.ncsu.edu

Edelman, S. (1996). A review of the Wechsler Intelligence Scale for Children-Third Edition (WISC III). *Measurement and Evaluation in Counseling and Development, 28,* 219-224.

Fetterman, D.M. (1998). *Ethnography: Step by step.* Thousand Oaks, CA: Sage.

Fuchs, L., Fuchs, D., Eaton, S., Hamlett, C., Binkley, E., & Crouch, R. (2000). Using objective data sources to enhance teacher judgments about test accommodations. *Exceptional Children, 67* (1), 67-81.

Gaster, L., & Clark, C. (1995). *A guide to providing alternate formats.* West Columbia, SC: Center for Rehabilitation Technology Services. (ERIC Document No. ED 405689)

Geisinger, K.F., & Carlson, J.F. (1992) *Assessing language-minority students*. Washington, DC: ERIC Clearinghouse on Tests, Measurement, and Evaluation. (ERIC Document No. ED 356232)

Grise, P., Beattie, S., & Algozzine, B. (1982). Assessment of minimum competency in fifth grade learning disabled students: Test modifications make a difference. *Journal of Educational Research, 76,* 35-40.

Gregory, M., & Poulton, E.C. (1970). Even versus uneven right-hand margins and the rate of comprehension in reading. *Ergonomics, 13* (4), 427-434.

Hanson, M.R. (1997). *Accessibility in large-scale testing: Identifying barriers to performance*. Delaware: Delaware Education Research and Development Center.

Hanson, M.R., Hayes, J.R., Schriver, K., LeMahieu, P.G., & Brown, P.J. (1998). *A plain language approach to the revision of test items*. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA, April 16, 1998.

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

Hartley, J. (1985). *Designing instructional text (2nd Edition).* London: Kogan Page.

Hedges, L.V. (1982). Estimation if effect size from a series of independent experiments. *Psychological Bulletin, 92,* 490-499.

Henriques, D.B. (2003). Rising demands for testing push limits of its accuracy. *New York Times On-Line,* September 2, 2003. Retrieved September, 2003 from the World Wide Web: http://www.nytimes.com

Heines (1984) *An examination of the literature on criterion-referenced and computer assisted testing*. (ERIC Document Number 116633)

Hoerner, A., Salend, S., & Kay, S.I. (1997). Creating readable handouts, worksheets, overheads, tests, review materials, study guides, and homework assessments through effective typographic design. *Teaching Exceptional Children, 29* (3), 32-35.

Howell, D.C. (1999). *Statistical methods for psychology*. Pacific Grove, CA: Duxbury.

Kiplinger, V.L., Haug, C.A., & Abedi, J. (2000). *Measuring math–not reading–on a math assessment: A language accommodations study of English language learners and other special populations*. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA, April 24-28, 2000.

Kopriva, R. (2000). *Ensuring accuracy in testing for English language learners*. Washington D.C.: Council of Chief State School Officers.

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

Mace, R. (1998). *A perspective on universal design.* An edited excerpt of a presentation at Designing for the 21st Century: An International Conference on Universal Design. Retrieved January, 2002, from the World Wide Web: http://www.adaptenv.org/examples/ronmaceplenary98.asp?f=4

Menlove, M., & Hammond, M. (1998). Meeting the demands of ADA, IDEA, and other disability legislation in the design, development, and delivery of instruction. *Journal of Technology and Teacher Education. 6* (1), 75-85.

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

Muller, D., Calhoun, E., & Orling, R. (1972). Test reliability as a function of answer sheet mode. *Journal of Educational Measurement, 9,* (4), 321-324.

Muncer, S.J., Gorman, B.S., Gorman, S., & Bibel, D. (1986). Right is wrong: An examination of the effect of right justification on reading. *British Journal of Educational Technology, 1* (17), 5-10.

Heubert, J.P. (1999). *High stakes: Testing for teaching, promotion, and graduation.* Washington, DC: National Research Council.

Osborne, H. (2001). "In Other Words…Communication across a life span…universal design in print and web-based communication. *On Call* (January). Retrieved January, 2002, from the World Wide Web: http://www.healthliteracy.com/oncalljan2001.html

Patton, M.Q. (1990). *Qualitative evaluation and research methods.* Newbury Park, CA: Sage.

Popham, W.J. (2001). *The truth about testing: An educator's call to action.* Alexandria, VA: Association for Supervision and Curriculum Development.

Popham, W.J., & Lindheim, E. (1980). The practical side of criterion-referenced test development. *NCME Measurement in Education, 10* (4), 1-8.

Rakow, S.J., & Gee, T.C. (1987). Test science, not reading. *Science Teacher, 54* (2), 28-31.

Rogers, W.T. (1983). Use of separate answer sheets with hearing impaired and deaf school age students. *BC Journal of Special Education, 7* (1), 63-72.

Schiffman, C.B. (1995). *Visually translating materials for ethnic populations*. Virginia. (ERIC Document Number ED 391485)

Schriver, K.A. (1997). *Dynamics in document design.* New York: John Wiley & Sons, Inc.

Sharrocks-Taylor, D., & Hargreaves, M. (1999). Making it clear: A review of language issues in testing with special reference to the National Curriculum Mathematics Tests at Key Stage 2. *Educational Research, 41* (2), 123-136.

Silver, A.A. (1994). Biology of specific (developmental) learning disabilities. In N.J. Ellsworth, C.N. Hedley, & A.N. Barratta, (Eds.), *Literacy: A redefinition*. Mahwah, New Jersey: Erlbaum Associates.

Smith, J.M., & McCombs, M.E. (1971). Research in brief: The graphics of prose. *Visible Language, 5* (4), 365-369.

State of California (2003). Education Law Section 60061.8. Enacted February 23, 2003.

State of Vermont (2003). *Request for proposals: Statewide assessment*. Retrieved May, 2003 from the World Wide Web: http://www.state.vt.us/educ/new/html/pgm_assessment/rfp_revised_5_22_03.html

Szabo, M., & Kanuka, H. (1998). Effects of violating screen design principles of balance, unity, and focus on recall learning, study time, and completion rates. *Journal of Educational Multimedia and Hypermedia, 8* (1), 23-42.

Thompson, D.R. (1991). *Reading print media: The effects of justification and column rule on memory.* Paper presented at the Southwest Symposium, Southwest Education Council for Journalism and Mass Communication, Corpus Christi, TX. (ERIC Document Number 337 749)

Thompson, S.J., Johnstone, C.J., & Thurlow, M.L. (2002). *Universal Design applied to large scale assessments* (Synthesis Report 44). Minneapolis, MN: National Center on Educational Outcomes. Available at http://education.umn.edu/nceo/OnlinePubs/Synthesis44.html

Thurlow, M., Quenemoen, R., Thompson, S., & Lehr, C. (2001). *Principles and characteristics of inclusive assessment and accountability systems* (Synthesis Report 40). Minneapolis, MN: National Center on Educational Outcomes. Available at http://education.umn.edu/nceo/OnlinePubs/Synthesis40.html

Tindal, G., & Fuchs, L.S. (1999*). A summary of research on test changes: An empirical basis for defining accommodations.* Lexington, KY: University of Kentucky, Mid-South Regional Center.

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

Tinker, M.A. (1963). *Legibility of print.* Ames, IA: Iowa State University Press.

Viet, D.T., & Scruggs, T.E. (1986). Can learning disabled students effectively use separate answer sheets? *Perceptual and Motor Skills, 63,* 155-160.

Willems, J.D. (1986). Social class segregation and its relationship to pupils' examination results in Scotland. *American Sociological Review, 51,* 224-242.

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

West, T.G. (1997). *In the mind's eye: Visual thinkers, gifted people with dyslexia and other learning difficulties, computer images, and the ironies of creativity.* Amherst, NY: Prometheus Books.

Worden, E. (1991). *Ergonomics and literacy: More in common than you think*. Indiana. (ERIC Document Number 329 901)

Zachrisson, G. (1965). *Studies in the legibility of printed text*. Stockholm: Almquist and Wiskell.

Appendix A

Interview Questions

1. Look at both tests, which one was easier for you? Why?

2. What made test ( ) easier for you than test ( )?

3. Which test was easier to read? Why do you think that?

4. Did having more time help you? Why?

5. If you could tell something to the people who make tests, what would you tell them?

6. How do you feel when you take tests every year? Would anything help you to feel (even) better?

7. You did a lot better on Test ( ) than on Test ( ). Tell me why you think that is true.