Universal Design Applied to Large Scale Assessments

NCEO Synthesis Report 44

Published by the National Center on Educational Outcomes

Prepared by:

Sandra J. Thompson • Christopher J. Johnstone • Martha L. Thurlow

June 2002

Acknowledgments

NCEO extends its sincere thanks to the following individuals who provided us with their reactions to this report. They provided comments reflecting their unique perspectives and made suggestions for changes that improved the report in a variety of ways.

Jamal Abedi, Center for Research on Evaluation, Standards, and Student Testing (CRESST)
Karen Barton, CTB McGraw Hill
Betsy Case, Harcourt Educational Measurement
Lizanne DeStefano, University of Illinois
James Friedebach, Missouri Department of Education
Kevin McGrew, University of Minnesota
Edward Roeber, Measured Progress
Alan Sheinker, CTB McGraw Hill
American Printing House for the Blind
Center for Applied Special Technology (CAST)

In addition to the detailed input of these individuals, NCEO expresses special appreciation to its partners, Eileen Ahearn of the National Association of State Directors of Special Education (NASDSE) and John Olson of the Council for Chief State School Officers (CCSSO), for their input on ways to make the report better. As ever, gratitude is due to our OSEP Project Officer, Dave Malouf, who went the extra mile to remind us of the importance of this report as it went through revision after revision. While the concept of universally designed assessments is still a work in progress, we hope that this report will make the concept more concrete and will move the field forward in assuring that assessments are designed and developed from the beginning to be appropriate for the widest range of students in school today.

Executive Summary

Universal design is a concept that began in the field of architecture, but has been quickly expanding into environmental initiatives, recreation, the arts, health care, and now, education. Despite a slow but steady start in its application to instruction, the potential for dramatically affecting the design of large scale assessments is great. There is a tremendous push to expand national and state testing, and at the same time to require that assessment systems include all students – including those with disabilities and those with limited English proficiency—many of whom have not been included in these systems in the past. Rather than having to retrofit existing assessments to include these students (through the use of large numbers of accommodations or a variety of alternative assessments), new assessments can be designed and developed from the beginning to allow participation of the widest possible range of students, in a way that results in valid inferences about performance for all students who participate in the assessment.

The purpose of this paper is to explore the development of universal design and to consider its application to large scale assessments. Building on universal design principles presented by the Center for Universal Design, seven elements of universally designed assessments are identified and described in this paper. The seven elements are:

1. Inclusive assessment population

2. Precisely defined constructs

3. Accessible, non-biased items

4. Amendable to accommodations
5. Simple, clear, and intuitive instructions and procedures
6. Maximum readability and comprehensibility
7. Maximum legibility

Each of the elements is explored in this paper. Numerous resources relevant to each of the elements are identified, with specific suggestions for ways in which assessments can be designed from the beginning to meet the needs of the widest range of students possible.

With the shift to standards-based reform during the past decade, valid assessments for measuring the achievement of all students are essential. There is no longer an option for test developers to ignore the possibilities that universal design can bring to truly inclusive assessment systems. States that release requests for proposals for their state assessments have a similar obligation – to ensure that any proposal from test developers meets criteria that reflect the elements of universal design highlighted in this paper.

Universal design opens the door to rethinking assessments—to ensure that the assessments themselves are not the barriers to improved learning. Universally designed assessments are a promising approach to providing appropriate assessment conditions for all students, giving each student a comparable opportunity to demonstrate achievement of the standards being tested.

Overview

Large scale assessments are used at local, state, and national levels to measure the progress of schools toward the achievement of educational standards. In this time of educational reform, testing programs are finding that to have accurate and fair measures of progress, all students must be included in the assessment system (Thurlow, Quenemoen, Thompson, & Lehr, 2001). Beyond simply including all students in assessments, there is a need to have their test performance be a valid and reliable measure of their knowledge and skills.

Questions have been raised about whether the administration, procedures, and format of these assessments provide optimal conditions for demonstrating achievement of academic content standards (Hanson, 1997). The standard administration procedures of current large scale assessments may have proven validity and reliability for many students. However, for many other students, standard administrations do not provide appropriate testing conditions, and may actually reduce access for some students, including those from varied cultural backgrounds, those with limited English proficiency, and those with disabilities (Abedi, Leon, & Mirocha, 2001). Students without any special learning needs are just as likely as those with learning needs to have difficulty with unnecessarily confusing or complex formats or design.

The concept of universal design has been developing throughout the world, beginning in the field of architecture and expanding into environmental initiatives, recreation, the arts, health care, and education. A set of principles of universal design has been developed that traverses several of these fields and initiatives (Center for Universal Design, 1997). This paper proposes that assessments based on universal design principles offer a promising approach to providing optimal, standardized assessment conditions for all students, giving each student a comparable opportunity to demonstrate achievement of the standards being tested. The purpose of this paper

is to explore the development of universal design and to consider its application to large scale assessments.

What is Universal Design?

More than 20 years ago, Ron Mace (an architect who was a wheelchair user) began to actively promote a concept he termed "universal design." Mace was adamant that we did not need more special purpose designs that serve primarily to meet compliance codes and may also stigmatize people. Instead, he promoted design that works for most people, from the child who cannot turn a doorknob to the elderly woman who cannot climb stairs to get to a door. 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).

General Universal Design Principles

A cross-disciplinary group working on universal design (architects, product designers, engineers, and environmental design researchers) originally compiled seven general universal design principles:

- Equitable Use
- Flexibility in Use
- Simple and Intuitive Use
- Perceptible Information
- Tolerance for Error
- Low Physical Effort
- Size and Space for Approach and Use

With funding from the National Institute on Disability and Rehabilitation Research, U.S. Department of Education, this group identified the principles, defined each principle, and provided characteristics of each principle (see Table 1). The principles were intended to apply to the evaluation of existing designs, guide the design process, and educate both designers and consumers about the characteristics of more usable products and environments.

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.

In developing principles of universal design, the group recognized that the ultimate objective of universal design is to be as inclusive as possible. Still, it recognized that it is nearly impossible to design all things for all people. 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).

Application of Universal Design to Instruction

No matter how well a test is designed, or what media are used for administration, students who have not had an opportunity to learn the material tested will perform poorly. According to the National Research Council (1999), "High standards cannot be established and maintained merely by imposing them on students" (p. 5). Students need access to the information tested in order to have a fair chance at performing well. Abedi, Hofstetter, Baker, and Lord (2001) found large significant differences in both reading and math performance of students at different schools and of students taught by different teachers. According to Heubert (2002), "Many teachers are not yet teaching students the full range of knowledge and skills that state tests measure, and the gap is probably greatest for students with disabilities, minority students, and English-language learners" (p. 16).

Universally designed instruction provides a way to establish optimal conditions for learning for all students. The goal is not to standardize instruction, but to provide opportunities for maximum access and high expectations for learning for each student, in light of his or her unique characteristics (Meyer & O'Neill, 2000). According to Heubert (2002), "Students with disabilities and minority students are often the victims of low expectations and weak instruction, and stand to benefit from efforts to provide high-quality instruction for all students" (p. 1). This is consistent with the standards movement in American public education, in which the goal is "to enable all students to attain high levels of academic achievement" (Heubert, p. 2).

A first attempt to clarify the meaning of universally designed instruction occurred in 1997 when a stakeholder meeting was convened on the topic. A report on the meeting provided a definition of universal design in learning, and suggested implications for universally designed instruction:

In terms of learning, universal design means the design of instructional materials and activities that makes the learning goals achievable by individuals with wide differences in their abilities to see, hear, speak, move, read, write, understand English, attend, organize, engage, and remember. Universal design for learning is achieved by means of flexible curricular materials and activities that provide alternatives for students with differing abilities. These alternatives are built into the instructional design and operating systems of educational materials – they are not added on after-the-fact (ERIC/OSEP, 1998, ¶ 3).

The importance of these concepts is reflected in the establishment of the National Center on Accessing the General Curriculum by the Office of Special Education Programs. This center is working toward universally designed instruction through the electronic use of digital text, one avenue to providing a variety of accommodations without external changes in materials or procedures (see Center for Applied Research – CAST, 2001).

Application of Universal Design to Assessment

Universal design applied to instruction is likely to advance the need for universally designed assessments. Yet, other factors are also heightening this need. A major impetus is the difficulty

that states and districts are having in attempting to retrofit existing tests to be more inclusive. It seems clear that this difficulty will be eliminated or reduced if tests are developed from the beginning to be inclusive of all students.

"Universally designed assessments" 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. According to the National Research Council (1999), "fairness, like validity, cannot be properly addressed as an afterthought once the test has been developed, administered, and used. It must be confronted throughout the interconnected phases of the testing process, from test design and development to administration, scoring, interpretation, and use" (p. 81). The Standards for Educational and Psychological Testing (AERA, APA, NCME, 1999) also address this need by requiring 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).

Universally designed assessments are based on the premise that each child in school is a part of the population to be tested, and that testing results should not be affected by disability, gender, race, or English language ability. Universally designed assessments are not intended to eliminate individualization, but they may reduce the need for accommodations and various alternative assessments by eliminating access barriers associated with the tests themselves.

Elements of Universally Designed Assessments

The elements of universally designed assessments discussed here were derived from a review of literature on universal design, assessment and instructional design, and research on topics such as assessment accommodations and student test taking. Many of these elements have been applied in various forms by test developers who are continually working on increasing assessment validity, and some reflect widely accepted principles in the field of measurement. According to the National Research Council (1999), "research and development in the field of educational testing is continually experimenting with new modes, formats, and technologies" (p. 202).

All of the elements identified here will undergo further refinement and development as they are addressed within assessment design. These elements are:

1. Inclusive assessment population
2. Precisely defined constructs
3. Accessible, non-biased items
4. Amenable to accommodations
5. Simple, clear, and intuitive instructions and procedures
6. Maximum readability and comprehensibility
7. Maximum legibility

These elements flow from the principles of universal design originally proposed by the Center for Universal Design, although there is not a one-to-one correspondence and in some cases a principle is reflected in all of the elements. Some of the relationships between the principles and elements are shown 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 effieicntly 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).

Element #1. Inclusive Assessment Population

When assessments are first conceptualized, they need to be thought of in the context of the entire population of who will be assessed (AERA, APA, NCME, 1999; National Research Council, 1999). It is sometimes appropriate to limit the population to be included in a test (e.g., a placement test or a selection test), but this is not true for assessments designed for public educational accountability, or for measuring the performance of public schools or conferring benefits that should be generally available to all students. For those assessments, the target population needs to include every student. Assessments need to be responsive to growing demands – increased diversity, increased inclusion of all types of students in the general curriculum, and increased emphasis and commitment to serve and be accountable for *all* students.

The first principle of universal design (Center for Universal Design, 1997) requires equitable use. When applied to large scale assessment, this principle requires opportunities for the participation of all students, no matter what their cognitive abilities, cultural backgrounds, or linguistic backgrounds. Assessments need to measure the performance of students with a wide range of

abilities and skill repertoires, ensuring that students with diverse learning needs receive opportunities to demonstrate competence on the same content.

This does not mean that standards should be relaxed or that contructs to be measured should be changed. Items on a universally-designed standards-based assessment must be aligned to the content and achievement standards with the same depth and breadth of coverage, and the same cognitive complexity as the standards specify. The emphasis can be on accessibility using different formats, technologies, and designs to include all students. It must be clear right from the beginning that in order to be equitable, assessments need to measure the achievement of all students on the same standards. In 1993, Algozzine argued that the principles of "full inclusion" be applied to assessment by avoiding practices that create separation among groups.

A clear implication of Element #1 is that field-tests should sample every type of student expected to participate in the final assessment administration, including students with a wide range of disabilities, students with limited English proficiency, and students across racial, ethnic, and socioeconomic lines. Checking field-test items with a broad range of students will not only help determine whether items are unclear, misleading, or inaccessible for certain groups of students, but will help assure that assessment procedures are applied to all students when the assessment is fully implemented. In order to develop solid item statistics, an additional important step in test development is the use of pilots to gain information about item functioning at a relatively small cost. Pilots should be used with subgroups of interest, to allow item refinement before proceeding to a field-test. This will also improve the likelihood that the items will survive a field-test.

Element #2. Precisely Defined Constructs

An important function of well-designed assessments is that they measure what they actually intend to measure. According to Popham and Lindheim (1980), "a test development project begins with a careful consideration of the skills or attitudinal characteristics proposed for measurement" (p. 3). Just as universally designed architecture removes physical, sensory, and cognitive barriers to all types of people in public and private structures, universally designed assessments remove all non-construct-oriented cognitive, sensory, emotional, and physical barriers. This is referred to as construct-irrelevant variance, "the degree to which test scores are affected by processes that are extraneous to its intended construct" (AERA, APA, NCME, 1999).

Recently, much controversy has surrounded the question of whether the use of particular accommodations invalidates the measurement of the constructs that a test is designed to measure. For example, different groups may define reading comprehension differently. Some may define it as constructing meaning from written text, while others may have a broader construct targeting comprehension and not how the information is obtained. The argument for the latter is especially provocative for students who are visually impaired. With fewer students learning Braille, more students use technological devices that read text. It could be argued that this is the only way for these students to comprehend meaning from text.

Resolution of such issues is hampered by the lack of a clear, generally accepted definition of the construct the test is designed to measure. Because scores on state tests often influence high-stakes decisions about whether a student will be promoted to the next grade or can graduate from high school, the need for clearly defined constructs is more critical than ever. And, once these constructs are defined, they need to be available to people who make decisions about how tests can be administered.

Another common test construct controversy relates to the reading skills required on mathematics assessments. Several research studies have found that students with reading difficulties scored *higher* on math tests when questions were read to them (Calhoun, Fuchs, & Hamlett, 2000; Harker & Feldt, 1993; Koretz, 1997; Tindal, Heath, Hollenbeck, Almond, & Harniss, 1998). This finding implies that the reading requirements of a mathematics assessment may prevent students with marginal reading ability from demonstrating their competency in math. However, problem-solving items tend to require substantial reading. Math educators have mixed feelings about these items and their reading loads. Shorrocks-Taylor and Hargreaves (1999) suggested that the language used in questions on tests that assess subjects other than language become as "transparent" or simplified as possible so that the mathematical demands become clear for most of the students tested. Though these researchers found little mention of the language dimension of testing in the assessment literature, this has been a major concern of test developers for some time.

Element #3. Accessible, Non-Biased Items

According to the Standards for Educational and Psychological Testing (AERA, APA, NCME, 1999), "the quality of the items is usually ascertained through item review procedures and pilot testing. Items are reviewed for content quality, clarity and lack of ambiguity. Items sometimes are reviewed for sensitivity to gender or cultural issues" (p. 39). According to the National Research Council (1999), bias arises when:

Deficiencies in the test itself result in different meanings for scores earned by members of different identifiable subgroups. For example, a test intended to measure verbal reasoning should include words in general use, not words and expressions associated with, for example, particular cultures or locations, as this might unfairly advantage test takers from these cultural or geographical groups. (p. 78)

Kopriva (2000) describes a process for incorporating accessibility as a primary dimension of test specifications. In this process, the dimensions of breadth, depth, and accessibility are considered collectively in the development of tests and test items, making it possible for accessibility to be woven into the fabric of the test, rather than being added after the fact.

One way to reduce bias is to research whether any items are more difficult for students from particular subpopulations. This can be accomplished through the administration of a field-test that can help determine item difficulty and "ability to discriminate among test takers of different standing on the scale" (AERA, APA, NCME, 1999, p. 39). In order to evaluate the quality of the items, studies of differential item functioning (DIF) are often conducted by test developers. Differential Item Functioning occurs when students equated on relevant ability but representing different groups do not have the same probability of responding correctly to test items. DIF is usually investigated by comparing item difficulty. DIF analysis has traditionally been used to detect the differential function of an item according to group identity (e.g., race, gender, disability). Willingham (1988) refers to "comparable validity," defined as a test's ability to yield comparable scores from person to person, subpopulation to subpopulation, and setting to setting. Item response theory (IRT) is used to predict a student's probability of answering an item correctly.

Potentially biasing elements are defined by Popham and Lindheim (1980) as "anything in an item that could potentially advantage or disadvantage any subgroup of examinees within the populations to be tested" (p. 6). Popham (2001, p. 93) lists sample questions that could be asked to avoid potential bias:

- *Curricular congruence*. Would a student's response to this item, along with others, contribute to a valid determination of whether the student has mastered the specific content standards the item is supposed to be measuring?
- *Instructional sensitivity*. If a teacher is, with reasonable effectiveness, attempting to promote students' mastery of the content standard that this item is supposed to measure, is it likely that most of the teacher's students will be able to answer the item correctly?
- *Out-of-school factors*. Is this item essentially free of content that would make a student's socioeconomic status or inherited academic aptitudes the dominant influence on how the student will respond?
- *Bias*. Is this item free of content that might offend or unfairly penalize students because of personal characteristics such as race, gender, ethnicity, [disability,] or socioeconomic status?

There are gray areas of potential bias affected by varying experience. The distinction between deficiencies due to inexperience because of sensory or physical disabilities or instruction-related deficiencies is not clear-cut. The test score "may accurately reflect what the test taker knows and can do, but low scores may have resulted in part from not having had the opportunity to learn the material tested as well as from having had the opportunity and having failed to learn" (AERA, APA, NCME, 1999, p. 76).

It is important not to excuse children with sensory deficits from the academic expectations held for all other children. For example, emphasis on verbal communication that is typical in instruction for deaf students may leave less time for attention to academic content standards. Poor performance on assessments may reveal instructional deficiencies, and would not necessarily indicate item bias. The focus for these students should be on changing instructional practice, thereby giving them opportunities to master important skills and knowledge.

Many states have "sensitivity" or "bias" review panels for their assessments, although most of these do not include a disability or limited English proficiency representative. A review panel made up of people who are trained and knowledgeable in the issues for each of the student subgroups should be invited to screen potential items (Allen, Bulla, Goodman, Henderson, Skutchan, Willis, & Scott, in press; Geisinger & Carlson, 1992; Popham, 2001; Popham & Lindheim, 1980). The purpose of sensitivity/bias reviews is to be sure no child has an advantage or disadvantage because of the presentation or content of an item, which would invalidate that item's contribution to a test score. The item pool should be large enough so that a bias review committee has the flexibility to recommend the elimination of items that appear to be biased. Any items that are biased against particular populations should be replaced or changed to eliminate bias. Careful item development and a full bias review improve the validity of test results.

Kopriva (2000, pp. 67-68), in her discussion of accuracy in testing students with limited English proficiency, recommends the following elements of an expanded bias review:

- Bias reviews provide an opportunity for a wider range of special needs educators and other educational stakeholders to have input throughout the test development process.
- In addition to reviewing test items for offensive content, bias reviews provide an excellent opportunity for test developers and publishers to receive guidance about test formats, working, rubrics, non-text item accessibility, and administration and response conditions.

- Participants should be briefed on all steps taken to ensure accessibility throughout the development and implementation process.
- Publishers need to allocate sufficient time for a thorough, item-by-item review of materials.
- Participants should be able to review mock-ups of some of the assessments to get an idea of layout and presentation, and assess whether these are sufficient.

These elements clearly apply to other populations as well as to English language learners.

Element #4. Amenable to Accommodations

Accommodations have been used to increase access to assessments. Accommodations, which are changes in the way a test is presented or responded to, generally may be used only for students with disabilities and English language learners. Unfortunately, assessment accommodations produce some very complex assessment issues. In addition, there is a great deal of controversy about the "fairness" of some test accommodations, about which students should have access to these accommodations, and about how decisions are made. Research to validate the use of standard and non-standard accommodations is growing, but is difficult to conduct and has yet to provide conclusive evidence about the influence of many accommodations on test scores (Bielinski & Sheinker, 2001; Thompson, Blount, & Thurlow, 2002; Thurlow & Bolt, 2001; Tindal & Fuchs, 1999). Different research studies sometimes produce contradictory findings on the same accommodations (e.g., Hollenbeck, Tindal, Harniss, & Almond, 1999; Russell & Haney, 1997). The absence of clear research evidence means that opinion and expert judgment are the primary basis for decisions about which accommodations are allowable and which, if used, invalidate test scores. As a result, there is little consistency in accommodation policies among states (Thurlow, Lazarus, Thompson, & Robey, 2002).

As a result of issues surrounding assessment accommodations, some students with disabilities have been partially or fully excluded from the benefits associated with participation in large scale assessments. For example, some student assessment scores have been excluded from reports or accountability measures, or controversial procedures such as "out-of-level" testing (i.e., testing students at a grade level lower than the one in which they are enrolled) have been employed. These consequences are significant because they tend to result in these students also being excluded from the benefits of standards-based reforms and instruction.

Even though items on universally designed assessments will be accessible for most students, there will still be some students who continue to need accommodations. The goal of universal design in such cases is to facilitate the use of the appropriate accommodations and to reduce threats to validity and comparability of scores. For example, the use of Braille as an accommodation will be facilitated if the following features are avoided in the design of the test:

- Use of construct irrelevant graphs or pictures
- Use of vertical or diagonal text
- Keys and legends located to the left or bottom of the item, where they are more difficult to locate in Braille formats
- Items that depend on reading of graphic representations (such as blueprints, furniture in a room) that do not also have verbal/textual descriptions that can be translated into Braille

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

These features are also relevant for students with visual disabilities who do not use Braille, and possibly also for many students for whom visual features may create distractions.

Tests can be designed for compatibility with other accommodations as well. For example, two common accommodations are the provision of extended time and the provision of extra breaks in the testing session. These accommodations are more compatible with tests that are not timed and that can be easily broken into brief sessions without compromising validity or security. Performance items that require the use of manipulatives can be more amenable to accommodations if alternate response modes are designed from the beginning and included in field-tests.

Element #5. Simple, Clear, and Intuitive Instructions and Procedures

Assessment instructions and procedures need to be easy to understand, regardless of a student's experience, knowledge, language skills, or current concentration level. Instructions need to be in simple, clear, consistent, and understandable language, so that "test takers can respond to a task in the manner that the test developer intended" (AERA, APA, NCME, 1999, p. 47). For example, a student asked to circle a letter in one section should not be required to cross out the letter in the next section. Numbered paragraphs on reading comprehension tests allow more efficient location of a particular passage.

Instructions in complex language invalidate a test taken by students who cannot understand how they need to respond (ADDA, 2001; Elliott, 1999; Willingham, Ragosta, Bennett, Braun, Rock & Powers, 1988). Some questions to ask when designing assessment instructions that are simple to use are: Will it be possible for all students to work independently throughout this test? Are directions easy to follow? (Tindal & Fuchs, 1999). Grise, Beattie, and Algozzine (1982) found that providing at least one example for each set of items to be tested had positive effects on students. Multiple standardized explanations of instructions could be included with a test. Practice materials may also be helpful. According to the Standards for Educational and Psychological Testing (AERA, APA, NCME, 1999, p. 47):

When appropriate, sample material, practice or sample questions, criteria for scoring, and a representative item identified with each major area in the test's classification or domain should be provided to the test takers prior to the administration of the test or included in the testing material as part of the standard administration instructions.

Guidelines and procedures for test administration also need to be "presented with sufficient clarity and emphasis so that it is possible for others to replicate adequately the administration conditions under which the data on reliability and validity, and, where appropriate, norms were obtained" (AERA, APA, NCME, 1999, p. 47).

Element #6. Maximum Readability and Comprehensibility

Readability is defined by Rakow and Gee (1987) as "an estimate of probability of comprehension by a particular group" (p. 28). Readability is often calculated by considering sentence length and number of difficult words, under the assumption that shorter sentences and easier words make text more readable. Formulas for estimating the readability of continuous text may be inappropriate for single sentence questions or answer choices (Popham & Lindheim, 1980; Rakow & Gee, 1987). Important features of text, such as logical organization of ideas and clarity

of sentence structures are not accounted for in readability formulas. It is possible to write a disorganized text, full of incomprehensible sentences and still obtain a good readability score. Further, dividing long sentences into shorter sentences and substituting familiar words for less familiar words can sometimes make a text more difficult to understand (Anderson, Hiebert, Scott, & Wilkinson, 1985). Lexiles, where text is scaled according to the vocabulary and linguistic difficulty, provide a way of psychometrically judging the difficulty level of text, and might provide a way to gauge readability more accurately (Hanson et al., 1998).

Readability is affected by "students' previous experiences, achievement, and interests, and by text features such as word and sentence difficulty, organization of materials, and format" (Rakow & Gee, 1987, p. 28). Gaster and Clark (1995) recommended these readability guidelines for all print materials:

- 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.

Rakow and Gee expressed concern that the readability of tests is especially important for large scale tests such as the National Assessment of Educational Progress (NAEP) science exam: "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). As a result of this concern, Rakow and Gee developed a checklist to use as a guide for minimizing readability problems in test items (see Table 3). According to Popham and Lindheim (1980), "It is often desirable to limit the number of words and the grammatical complexity of the test materials, as well as the level of difficulty of the individual words themselves" (p. 4). However, it is important to clarify whether the "simplification" of vocabulary would affect the use of content specific words for mathematics, science, and social studies.

Table 3. Rakow and Gee's Test Item Readability Checklist

1. Students would likely have the experiences and prior knowledge necessary to understand what thequestion calls for.
2. The vocabulary is appropriate for the intended grade level.
3. Sentence complexity is appropriate for the intended grade level.
4. Definitions and examples are clear and understandable.
5. The required reasoning skills are appropriate for the students' cognitive level.
6. Relationships are made clear through precise, logical connectives.
7. Content within items is clearly organized.
8. Graphs, illustrations, and other graphic aids facilitate comprehension.
9. The questions are clearly framed.
10. The content of items is of interest to the intended audience.

Source: Rakow & Gee, 1987.

It is critical to define the construct to be measured, and to use plain language when vocabulary level is not part of the construct being tested. Plain language approaches are now being studied by several researchers. A study of language accommodation on math assessments (Kiplinger, Haug, & Abedi, 2000) found that the performance of students on a mathematics assessment with high proportions of word problems was directly related to their proficiency in reading in English. Better performance of English language learners and other students who were not good readers resulted from the simplification of linguistic structures and the addition of a glossary for non-mathematics vocabulary. The results suggest that linguistic simplification or clarification of vocabulary on mathematics word problems can benefit virtually all students. Another study by Abedi, Hofstetter, Baker, and Lord (2001) found that a modified English accommodation on a math test enabled students with limited English proficiency to achieve scores that were comparable to those of non-LEP students.

"The successful performance on tests often depends on students' ability to read, decode, comprehend, and respond to written text" (Hanson, Hayes, Schriver, LeMahieu, & Brown, 1998, p.2). Many students may be "unfairly disadvantaged by achievement tests that place a heavy burden on reading skills." These researchers define "plain language" as "text-based language that is straightforward, concise, and uses everyday words to convey meaning. The goal of plain language editing strategies is to improve the comprehensibility of written text while preserving the essence of its message" (p. 2).

Universally designed assessments reduce the "verbal and organizational complexity of test items while preserving their essential content (i.e., the skills and concepts they were intended to measure)" (Hanson et al., 1998, p.2). The researchers believe that "a Plain Language approach to writing or revising test items in subjects like math and science will simultaneously reduce the language demands placed on students and thus, irrelevant score variance attributable to students' reading skills" (p.2). Brown (1999) lists strategies for plain language editing; these are presented in Table 4.

Table 4. Plain Language Editing Strategies

Strategy	Description
Reduce excessive length.	Reduce wordiness and remove irrelevant material. Where possible, replace compound and complex sentences with simple ones.
Eliminate unusual or low frequency words and replace with common words.	For example, replace "utilize" with "use."
Avoid ambiguous words.	For example, "crane" could be a bird or a piece of heavy machinery.
Avoid irregularly spelled words.	For example, "trough" and "feign."
Avoid proper names.	Replace proper names with simple, common names such as first names.
Avoid inconsistent naming and graphic conventions.	Avoid multiple names for the same concept. Be consistent in the use of typeface.
Avoid unclear signals about how to direct attention.	Well-designed headings and graphic arrangement can convey information about the relative importance of information and order in which it should be considered. For example, phrases such as "in the table below,…" can be helpful.
Mark all questions.	When asking more than one question, be sure that each is specifically marked with a bullet, letter, number, or other obvious graphic signal.

Source: Brown, 1999.

In focus groups on assessment, Hanson (1997) found that both regular and special education teachers indicated that "the linguistic demands (i.e., decoding of vocabulary, comprehension of instructions, writing a response, etc.) of assessment situations often posed the greatest barrier to students' ability to demonstrate their knowledge of test mathematical concepts" (p. 15). A study by Brown (1999) reported that, for students who understood the content of a science test, the plain language version was better able to accurately assess their knowledge. However, for students who did not understand the content, the version made little if any difference in performance.

Element #7. Maximum Legibility

Legibility refers to the capability of being deciphered with ease. The term may be applied to text, to various types of tables, figures, and illustrations, and to response formats. Each of these are very important aspects of legibility in assessments, and therefore are treated separately in this discussion.

Legible Text. Legible text is the physical appearance of text – the way shapes of letters and numbers enable people to read text "quickly, effortlessly, and with understanding" (Schriver, 1997, p. 252). Though a great deal of research has been conducted in this area, the personal opinions of editors often prevail (Bloodsworth, 1993; Tinker, 1963). Bias results from items that contain physical features that interfere with a student's focus on or understanding of the construct an item is intended to assess. Dimensions of legible text include contrast, type size, spacing, typeface, leading, justification, line length/width, and blank space (see Table 5).

Table 5. Characteristics of Legible Text

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)
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.

Legible Graphs, Tables, and Illustrations. Symbols used on graphs need to be highly distinguishable (Schutz, 1961), especially if they are in black and white. Gregory and Poulton (1970) suggested placing labels directly next to plot lines, enabling people to find information more quickly than when a legend or key is used, and reducing the load on short-term memory. According to Schriver (1997), "document designers must give structure to quantitative displays so that readers can construct appropriate inferences about the data" (p. 393). She goes on to state that a well-designed quantitative graphic creates a context for interpreting data.

Shorrocks-Taylor and Hargreaves (1999) described three types of illustrations that appear on assessments:

- 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.

For some students, illustrations result in problems of discrimination due to visual acuity or related challenges. Other students may be unnecessarily distracted due to an inability to shift their focus between the relevant information and extraneous or irrelevant information. For example, illustrations added for interest may draw attention of some students away from the construct an item is intended to assess. Some illustrations use color to attract student attention and maintain student interest. If illustrations use greens and reds, some students may have difficulty due to color blindness. Illustrations need to be directly next to the question to which they refer (Silver, 1994; West, 1997). Black and white line drawings of very simple design are the clearest. Illustrations also may complicate the use of magnifiers, enlargement, or other assistive technology.

Szabo and Kanuka (1998) outlined design principles for computer tests that optimize completion rates and speed of test taking. Principles of unity, focal point, and balance have been shown to reduce the cognitive load of perceiving graphic information, increasing speed of perceiving information, and increasing speed of tests taken with graphic material.

Illustrations need to be meaningful to students participating in the assessment. According to Schriver (1997), "In evaluating graphics, it is essential to explore their appropriateness in relation to the readers' knowledge and cultural context" (p. 375). Cultural norms, beliefs, and customs need to be respectfully reflected in illustrations (Schiffman, 1995).

Legible Response Formats. Tests that have small print, small bubbles to fill in for answers, and small diagrams are inherently biased against people with low vision and people who have difficulty with fine motor skills. Marking in the test booklet is recommended for both Braille and large print users. One of the characteristics sometimes associated with learning disabilities is lack of body awareness and poor directionality. Response mechanisms thus should allow larger circles for bubble responses or multiple forms of response (Willingham et al., 1988).

Several studies have been conducted on the use of different response formats. Grise et al. (1982) found that placing answer options in a vertical format with flattened, horizontal elliptical ovals for answer bubbles was useful. By placing answer bubbles on the same sheet as questions, the opportunity to miss one bubble and mismark many items was greatly diminished. Other research has had mixed results at different grade levels. For example, Rogers (1983) reported that separate answer sheets resulted in invalid scores for both typical and hearing impaired students in grades 1–3 and only resulted in valid scores for students in grades 4 and 5 when special instructions and practice were provided. Tindal et al. (1998) found separate answer sheets to provide invalid scores for students in grades 1–3, but no difference for fourth graders. Veit and Scruggs (1986), however, found that fourth grade students with learning disabilities took significantly more time to complete tests that required bubbling on a separate answer sheet. Muller, Calhoun, and Orling (1972) found that grade 3–6 students made significantly fewer errors when allowed to answer questions directly on the test. If students answer directly in the test booklets, it is important to acknowledge that there will be some financial consequence for the practice because test booklets will be used only once rather than multiple times. Wise, Plake, Eastman, and Novak (1987) found no significant impact on reading or math scores on the California Achievement Test due to the use of different response formats, even for third graders. The authors cautioned, however, that there was little geographic or ethnic diversity in the population, and little diversity in ability.

Computer-Based Assessments

Even though the universal design elements reviewed in this paper have been discussed primarily in reference to paper/pencil assessments, it is important to acknowledge that test developers are

moving quickly to implement computer-based assessments (Trotter, 2001). It should not be assumed that computer-based assessments are universally designed. It is important to maintain each element of universal design when implementing computer-based assessments. Computer-based assessments can be poorly designed and inaccessible in much the same way as paper and pencil assessments.

Computer-based assessment has been viewed as a vehicle to increase the inclusion of students with disabilities in testing programs. A study conducted by Burk (1999) suggested that computerized testing for students with disabilities is a viable medium. She found that accommodations such as large print, audio overlay, and extra spacing were relatively easy accommodations to implement using computer-based testing. Burk cautioned, however, that the use of computer-based testing for students with disabilities requires appropriate equipment, as well as attention to adequate staff and student preparation.

Computer-based assessment may lend itself to universal design in certain ways. For example, it can provide for immediate feedback and efficient results, offer the preferred testing modality for students, and facilitate links to instruction (Baker, 1999, Bushweller, 2000). However, computer-based assessment may introduce its own barriers to accessibility. For example, an article in *Education Week* (5/23/01) highlighted several potential pitfalls of computer-based testing from the perspective of students who reported on their experiences (Trotter, 2001). Concerns included unfamiliarity with answering standardized test questions on a computer screen, using the "next" and "back" buttons excessively to search for specific items, indecision about whether to use traditional tools provided (e.g., hand held calculator) versus computer-based tools, and the inability to see an entire problem on screen at one time (some items required scrolling horizontally and vertically to get the entire graphic on the page). Cole, Tindal, and Glasgow (2000) hypothesized that children in their study performed poorly on a computer-based test because they were not familiar with computer competencies like scrolling with a mouse or using text that is on multiple screens.

Equity may become an issue when different media and equipment, such as computers and calculators, are used in testing. While these tools have increased access and validity for some students, others have had limited practice using them in class or at home prior to assessments (Bridgeman, Harvey, & Braswell, 1995; MacArthur & Graham, 1987). For example, a test that requires the use of a scientific calculator or computer might not be equitable to students who have not had opportunities to use these tools in instructional settings. The computer can be used to access any calculator that the student may need to use for the mathematics test. If a scientific calculator is necessary then that can be available, or if a four-function calculator is needed then it can be accessed.

There are several assistive technology software tools that need to be considered in the design of computer-based assessments. As with other aspects of universal design, people are finding that some of this technology has applications for a wide range of users – not just those with specific disabilities (e.g., low vision). These tools include:

1. Text-to-speech technology or speech synthesis: Software that reads text aloud through an audio format.

2. Electronic reading supports: Software that adds spoken voice, visual highlighting, document navigation, or page navigation to any electronic text.

3. Alternative access devices: These allow the use of devices such as a special mouse, track ball, or other alternate means for keyboard access such as a switch.

It is important to ensure that the student is already familiar with the computer-testing environment. Familiarity with the text-to-speech engine and with screen reading software, for example, is essential before attempting to use this technology in an assessment situation. Universally designed assessments need to be compatible with widely used adaptive equipment. According to the Assistive Technology Act (1998):

The use of universal design principles reduces the need for many specific kinds of assistive technology devices and assistive technology services by building in accommodations for individuals with disabilities before rather than after production. The use of universal design principles also increases the likelihood that products will be compatible with existing assistive technologies.

Conclusion

The purpose of this paper has been to explore the development of universal design and to identify important elements for large scale assessments. We expect that the elements of universal design of assessment will be expanded and become more concrete as they are applied to assessment design and development. For example, the Assessing Special Education Students SCASS (State Collaborative on Assessment and Student Standards), formed by the Council for Chief State School Officers, is planning to create a checklist that states and test developers can use to ensure that their assessments reflect the universal design elements. With the increased emphasis on testing in the nation's schools in response to federal and state mandates, it is essential that this progress occur as rapidly as possible. This will require the consolidation and application of current best practices in assessment along with research and innovation to expand our knowledge in this area.

While universally designed assessments can make tests more equitable, producing results that are more valid for all students, they cannot replace instructional opportunity. No matter how a test is designed, students who have not had an opportunity to learn the material tested probably will perform poorly. Students need access to the information tested in order to have a fair chance to perform. There are significant issues of instruction and opportunity to learn that must also be addressed.

Just as the provision of universally designed assessments does not eliminate the need for good instruction, the provision of adequate standards-based instruction is not going to eliminate the need for well-developed, universally-designed assessments. To the extent that test developers improve assessments, the educational system will have better measures of the performance of all students.

There is no longer the option for test developers to ignore the possibilities that universal design can bring to making assessment and accountability systems truly inclusive (Thurlow, Quenemoen, Thompson, & Lehr, 2001). Universal design can help to ensure that assessments themselves do not produce barriers to learning. The concept of universal design helps us to rethink our basic assumptions about how to create national, state, and district assessments that give a more accurate picture of what all students know and can do so that educators can focus on the critical target of providing universally designed standards-based instruction.

References

Abedi, J. Hofstetter, C., Baker, E., & Lord, C. (2001). *NAEP math performance and test accommodations: interactions with student language background*. Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing.

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.

ADDA (Attention Deficit Disorder Association). (2001). *Accommodations for testing*. Retrieved January, 2002, from the World Wide Web: http://www.adda.org/.

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.

Algozzine, B. (1993). Including students with disabilities in systemic efforts to measure outcomes: Why ask why? In National Center on Educational Outcomes (Ed.), *Views on inclusion and testing accommodations for students with disabilities.* Minneapolis, MN: National Center on Educational Outcomes.

Allen, J., Bulla, N., Goodman, S., Henderson, B., Skutchan, L., Willis, D., & Scott, K. (in press). *Test access: Guidelines for computer administered testing.* Louisville, KY: American Printing House for the Blind.

Anderson, R.C., Hiebert, E.H., Scott, J.A, & Wilkinson, A.G. (1985). *Becoming a nation of readers*. Urbana, IL: University of Illinois, Center for the Study of Reading.

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

Assistive Technology Act. Public Law 105-394. Enacted November 13, 1998.

Baker, E.L (1999). Technology: Something's coming – something good. *CRESST Policy Brief 2*. Los Angeles, CA: UCLA, National Center for Research on Evaluation, Standards, and Student Testing.

Bielinski, J., & Sheinker, A. (2001). *Varied opinions on how to report accommodated test scores: Findings based on CTB/McGraw-Hill's framework for classifying accommodations*. Paper presented at the Council of Chief State School Officers' Large-scale Assessment Conference, Houston, TX.

Bloodsworth, J.G. (1993). *Legibility of print.* South Carolina: ERIC Document Number 335497.

Bridgeman, B., Harvey, A., & Braswell, J. (1995). Effects of calculator use on scores on a test of mathematical reasoning. *Journal of Educational Measurement, 32,* 323-340.

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

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

Bushweller, K. (2000). Electronic exams: Throw away the No. 2 pencils – here comes computerized testing. *Electronic School* (June), 20-24. Retrieved January, 2002, from the World Wide Web: http://www.electronic-school.com/.

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 Applied Special Technology (CAST) (2001). *Universal design for learning: An on-line handout.* Peabody, MA: CAST. Retrieved January, 2002, from the World Wide Web: www.cast.org.

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/.

Cole, C., Tindal, G., & Glascow, A. (2000). *Final report: Inclusive comprehensive assessment system research, Delaware large scale assessment program.* Eugene, OR: Educational Research Associates.

ERIC/OSEP (Educational Resources and Information Clearinghouse & Office of Special Education Programs. (1998, Fall). *Topical report.* Washington, DC: Author. Retrieved January, 2002, from the World Wide Web: www.cec.sped.org/osep/ud-sec3.html.

Elliott, S.N. (1999). *Valid testing accommodations: Fundamental assumptions and methods for collecting validity evidence.* Paper presented at CCSSO Conference, Snowbird, UT.

Fenton, E. (1996). *The Macintosh font book: Typographic tips, techniques and resources* (3rd ed.) Berkeley: Peachpit Press.

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)

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

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.

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.

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

Heubert, J.P. (2002). *Disability, race, and high-stakes testing of students.* Teachers College, Columbia University; Columbia Law School: National Center for Accessing the General Curriculum.

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.

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

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 DC: 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.

MacArthur, C.A., & Graham, S. (1987). Learning disabled students' composing under three methods of text production: handwriting, word processing, and dictation. *Journal of Special Education, 21* (3), 22-42.

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: 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.

Meyer, A., & O'Neill, L. (2000). Beyond access: Universal design for learning. *Exceptional Parent, 30* (3), 59-61.

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.

National Research Council. (1999). *High stakes: testing for tracking, promotion, and graduation* (J. Heubert & R. Hauser editors, Committee on Appropriate Test Use). Washington, DC: National Academy Press.

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: www.healthliteracy.com/oncalljan2001.html.

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.

Russell, M., & Haney, W. (1997). Testing writing on computers: a follow-up study comparing student performance on tests conducted via computer and via paper-and-pencil. *Education Policy Analysis Archives, 5* (3). Retrieved January, 2002, from the World Wide Web: http://www.epaa.asu.edu/.

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.

Schutz, H.G. (1961). An evaluation of methods for presentation of graphic multiple trends – Experiment III. *Human Factors, 31,* 108-119.

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.

Shiffrin, R.M., & Schneider, W. (1977). Controlled and automatic human information processing: Perceptual learning, automatic attending, and a general theory. *Psychological Review, 84* (2), 127-190.

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

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

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

Thurlow, M., & Bolt, S. (2001). *Empirical support for accommodations most often allowed in state policy.* (Synthesis Report 41). Minneapolis, MN: National Center on Educational Outcomes.

Thurlow, M., Lazarus, S., Thompson, S., & Robey, S. (2002). *State participation and accommodation policies for students with disabilities: 2001 update.* Minneapolis, MN: National Center on Educational Outcomes.

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.

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.

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

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

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.

Willingham, W.W. (1998). Testing handicapped people – the validity issue. In H. Weiner & H.I. Brown (eds.) *Test validity* (pp. 89-103). Hillsdale, NJ: Lawrence Erlbaum.

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.

Wise, S.L, Plake, B.S., Eastman, L.A., & Novak, C.D. (1987). Introduction and training of students to use separate answer sheets: Effects on standardized test scores. *Psychology in the Schools, 24,* 285-288.

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.

Large-scale Assessments in Education is a joint publication of the International Association for the Evaluation of Educational Achievement (IEA) and Educational Testing Service (ETS). The articles in this journal contribute to the science of large-scale assessments, help disseminate state-of-the-art information about empirical research using these databases and make the results available to policy makers and researchers around the world. Articles suitable for publication in the journal focus on improving the science of large-scale assessments and make use of data collected by programs, such as Fortunately, there is a field within the broader universal design movement-called universal design for assessment (UDA)-that applies the principles of universal design specifically to assessments, helping to ensure that they are accurate and equitable for all students.Â Given the high-stakes nature of large-scale assessments grounded in Common Core State Standards (CCSS) (National Governors Association Center for Best Practices & Council of Chief State School Officers, 2010) (see www.corestandards.org), assessments that often have significant consequences for students, teachers, and schools, educators must ensure that tests are an accurate measure of the knowledge and skills of all. students. (LSA) International Large Scale Assessments. (ILSA) Recent Developments. The Challenges.Â International Large Scale Assessments. (ILSA). ensure a positive impact on education. Recent Developments. Â§ For this to happen, assessments should be more complete, more authentic and fully integrated into the learning and teaching process. The Challenges.Â The SimScientists program investigates design principles to guide the creation of effective simulations as learning and assessment tools. In addition, the SimScientists program studies how science simulations can be used at different levels of the educational system - classroom, district, and state in balanced state science assessment systems.