Creating Better Tests for Everyone Through Universally Designed Assessments

Sandra Thompson and Martha Thurlow

University of Minnesota

David B. Malouf

U.S. Department of Education

Abstract

Universally designed assessments are designed and developed to allow participation of the widest possible range of students, in a way that results in valid inferences about performance on grade-level standards for all students who participate in the assessment. This paper explores the development of universal design and considers its application to large-scale assessments. Building on universal design principles presented by the Center for Universal Design (Center for Universal Design, 1997), seven elements of universally designed assessments are identified and described. These elements were derived from a review of literature on universal design, assessment and instructional design, and research on topics such as assessment accommodations (Thompson, Johnstone, & Thurlow, 2002). The seven 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

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 to meet the needs of the widest range of students possible. Challenges and opportunities arising from the application of universally designed assessments are identified.

Creating Better Tests for Everyone Through Universally Designed Assessments

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 (Hitchcock, 2001), 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 (Thurlow, Quenemoen, Thompson, & Lehr, 2001). 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 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.

With the shift to standards-based reform during the past decade, valid assessments for measuring the achievement of all students on grade-level standards 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.

Background

The standard administration of assessments is not appropriate for all students who must participate in state and district assessments today. The use of accommodations – changes in administration procedures or materials – is evidence that there are students who cannot participate in assessments or receive valid scores unless something is changed. Only a very small percentage of students need a completely different assessment, identified in federal special education law as an alternate assessment (Thompson, Quenemoen, Thurlow, & Ysseldyke, 2001). A much larger group of students need changes in the regular assessment.

Because of the emphasis on testing and including all students, the provision of accommodations and decisions about who should participate in alternate assessments has become very complex. There is a great deal of controversy about the “fairness” of many test accommodations and about which students should have access to accommodations and how decisions are made. 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).

Research to validate accommodation use is growing, but the research is difficult to conduct and rarely provides conclusive evidence about the effects of accommodations on validity (Bielinski & Sheinker, 2001; Elliott, Kratochwill, & McKevitt, 2001; Koretz, & Hamilton, 2000; Thompson, Blount, & Thurlow, 2002; Thurlow & Bolt, 2001; Tindal & Fuchs, 1999). States grapple with decisions about which accommodations should be included in school accountability and which invalidate assessment scores. Further, they frequently revise their accommodation policies (Thurlow, Lazarus, Thompson, & Robey, 2002), thereby increasing the likelihood of confusion about what the policies really are, and decreasing the likelihood that the policies will be implemented as intended. It is time to take a more global approach to addressing these testing issues, an approach in which increased access for all students is considered.

Applying Universal Design to Assessments

The concept of universal design is not new. Its use began in the field of architecture, but its application has spread rapidly into environmental initiatives, recreation, the arts, health care, and education. Principles of universal design that traverse all of these areas have been developed (see Table 1). It is reasonable to expect that they can apply equally as well to large-scale assessments.

Table 1. Principles of Universal Design in Architecture and Other Areas

Principle	Explanation
Equitable Use	The design is useful and marketable to people with diverse abilities.
Flexibility in Use	The design accommodates a wide range of individual preferences and abilities.
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.
Perceptible Information	The design communicates necessary information effectively to the user, regardless of ambient conditions or the user's sensory abilities.
Tolerance for Error	The design minimizes hazards and the adverse consequences of accidental or unintended actions.
Low Physical Effort	The design can be used efficiently and comfortably and with a minimum of fatigue.
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.

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

The goal of applying universal design principles to assessments is to be able to design and develop assessments that allow participation of the widest range of students, and result in valid inferences about their performance. The need that many students have for accommodations could be reduced if assessments could be universally designed. Universally designed assessments are not intended to eliminate individualization, or to take away from the Individualized Educational Program (IEP) process. Instead, they could make the IEP process richer by focusing on instructional needs rather than on all the changes that will have to be made for the student to participate in the assessment. Universal design is the best way to increase participation in general state and district assessments.

Universal design is based on the same ethics of equity and inclusiveness that are expected for people with disabilities and others in schools, communities, and on the job – an ethic that values differences in age, ability, culture, and lifestyle. Test performance should not be affected by disability, gender, race, English language ability, or levels of anxiety about tests. On the other hand, it is important to remember that universal design does not address deficiencies in instruction. Students who have not had an opportunity to learn the material tested will be disadvantaged during testing no matter how universal the design of the assessment.

Elements of Universally Designed Assessments

The National Center for Educational Outcome (NCEO) has conducted an extensive review of all research relevant to the assessment development process and the principles of universal design (Thompson et al., 2002). This review produced a set of

seven elements of universal design that apply to assessments (see Table 2). Each of these seven elements is discussed here.

Table 2. Elements of Universally Designed Assessments

Element	Explanation
Inclusive Assessment Population	Tests designed for state, district, or school accountability must include every student except those in the alternate assessment, and this is reflected in assessment design and field testing procedures.
Precisely Defined Concepts	The specific constructs tested must be clearly defined so that all construct irrelevant cognitive, sensory, emotional, and physical barriers can be removed.
Accessible, Non-Biased Items	Accessibility is built into items form the beginning, and bias review procedures ensure that quality is retained in all items.
Amenable to Accommodations	The test design facilitates the use of needed accommodations (e.g., all items can be Brailled).
Simple, Clear, and Intuitive Instructions and Procedures	All instructions and procedures are simple, clear, and presented in understandable language.
Maximum Readability and Comprehensibility	A variety of readability and plain language guidelines are followed (e.g., sentence length and number of difficult words are kept to a minimum) to produce readable and comprehensible text.
Maximum Legibility	Characteristics that ensure easy decipherability are applied to text, to tables, figures, and illustrations, and to response formats.

Based on Thompson, Johnstone, and Thurlow (2002).Inclusive Assessment

Population

When tests are first conceptualized, they need to be thought of in the context of who will be tested (AERA, APA, NCME, 1999; National Research Council, 1999). If the test is designed for state, district, or school accountability purposes, the target population must include every student except those who will participate in accountability through the alternate assessment. 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 accountability for all students.

Precisely Defined Constructs

An important function of well-designed assessments is that they actually measure what they are intended 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). Test developers need to carefully examine *what* is to be tested and design items that offer the greatest opportunity for success within those constructs. Just as universally designed architecture removes physical, sensory, and cognitive barriers to all types of people in public and private structures, universally designed assessments must remove all construct irrelevant cognitive, sensory, emotional, and physical barriers.

Accessible, Non-Biased Items

Items are reviewed through bias review or sensitivity review procedures to ensure that they do not create barriers because of lack of sensitivity to disability, cultural, or other subgroups. But, perhaps more important, items are developed by individuals who understand the varied characteristics of students, and the characteristics of items that might create difficulties for any group of students. Accessibility is incorporated as a primary dimension of test specifications, so that accessibility is woven into the fabric of the test rather than being added after the fact (Kopriva, 2000).

Amenable to Accommodations

Even though items on universally designed assessments will be accessible for most students, there will still be some students who continue to need accommodations. Thus, another essential element of any universally designed assessment is that it is compatible with accommodations and a variety of widely used adaptive equipment and assistive technology. 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.

Simple, Clear, and Intuitive Instructions and Procedures

Assessment instructions should be easy to understand, regardless of a student's experience, knowledge, language skills, or current concentration level. Directions and questions need to be in simple, clear, 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). Knowledge questions that are posed within complex language certainly invalidate the test if students cannot understand how they are expected to respond to a question (Elliott, 1999; Willingham, Ragosta, Bennett, Braun, Rock & Powers, 1988).

Maximum Readability and Comprehensibility

A variety of guidelines exist to ensure that text is maximally readable and comprehensible (Gaster & Clark, 1995). These features go beyond what is measured by readability formulas. Readability and comprehensibility are affected by many

characteristics, including student background, sentence difficulty, organization of text, and others. All of these features need to be considered in developing the text of assessments.

Plain language is a concept now being highlighted in research on assessments. For example, Kiplinger, Haug, and 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. Plain language has been defined as language that is straightforward and concise. Several strategies that have been identified for editing text to produce plain language are shown in Table 3.

Table 3. Plain Language Editing Strategies

Strategy	Description
Reduce excessive length	Reduce wordiness and remove irrelevant material.
Use common words	Eliminate unusual or low frequency words and replace with common words (e.g., replace "utilize" with "use").
Avoid ambiguous words	For example, "crane" should be avoided because it could be a bird or a piece of heavy machinery.
Avoid irregularly spelled words	Examples of irregularly spelled words are "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 heading and graphic arrangement can convey information about the relative importance of information and order in which it should be considered.
Mark all questions	Give an obvious graphic signal (e.g., bullet, letter, number) to indicate separate questions.

Source: Brown (1999).

Maximum Legibility

Legibility is the physical appearance of text, the way that the shapes of letters and numbers enable people to read text easily. As delineated by Schriver (1997), a leading document designer, text that is legible can be read "quickly, effortlessly, and with understanding" (p. 252). Despite a great deal of research on what the characteristics of maximum legibility are, the personal opinions of editors about how they want text to look often prevail.

Bias results when tests contain physical features that interfere with a student's focus on or understanding of the constructs that test items are intended to assess. Dimensions can include contrast, type size, spacing, typeface, leading, justification, line length/width, blank space, graphs and tables, illustrations, and response formats (see Table 4).

Table 4. Dimensions of Legibility and Characteristics of Maximum Legibility

Dimension	Maximum Legibility Characteristics
Contrast	Black type on matte pastel or off-white paper is most favorable for both legibility and eye strain.
Type Size	Large type sizes are most effective for young students who are learning to read, students with visual difficulties, and individuals with eye fatigue issues. The legal size for large print text is 14 point.
Spacing	The amount of space between each character can affect legibility. Spacing needs to be wide between both letters and words. Fixed-space fonts seem to be more legible for some readers than proportional-spaced fonts.
Leading	Leading, the amount of vertical space between lines of type, must be enough to avoid type that looks blurry and has a muddy look. The amount needed varies with type size (for example, 14-point type needs 3-6 points of leading).
Typeface	Standard typeface, using upper and lower case, is more readable than italic, slanted, small caps, or all caps.
Justification	Unjustified text (with staggered right margin) is easier to see and scan than justified text especially for poor readers.
Line Length	Optimal length is about 4 inches or 8 to 10 words per line. This length avoids reader fatigue and difficulty locating the beginning of the next line, which causes readers to lose their place.
Blank Space	A general rule is to allow text to occupy only about half of a page. Blank space anchors text on the paper and increases legibility.
Graphs and Tables	Symbols used on graphs need to be highly discriminable. Labels should be placed directly next to plot lines so that information can be found quickly and not require short-term memory.
Illustrations	When used, an illustration should be directly next to the question for which it is needed. Because illustrations create numerous visual and distraction challenges, and may interfere with the use of some accommodations (such as magnifiers), they should be used only when they contain information being assessed.
Response Formats	Response options should include larger circles (for bubble response tests), as well as multiple other forms of response.

Challenges and Opportunities

The application of universal design to assessments is just beginning as test developers and publishers consider how to apply the elements to assessments. There are both challenges and opportunities that arise as this application occurs.

Among the challenges associated with universally designed assessments is the possibility that development costs will increase at a time when the costs of assessments are already seen by some as excessive. Even though the incorporation of universal design should ultimately save time and money in not having to throw out items later in the test development process, the initial incorporation of universal design elements may seem expensive. Another challenge is that the specific criteria for putting all the universal

design considerations together have not yet been figured out – we do not know when the right balance has been reached to achieve the best item possible. Item review teams need to be cautious. Implementing the principles of universally designed assessments may prompt some reviewers to throw away items that may be usable with minor changes, or that may not have design problems that actually affect a student's response. The desire for authenticity and whether authenticity is more important than universal design is another challenge that the testing community and standards committees need to address.

Perhaps the greatest challenge is the perception that a universally designed assessment is a "cure-all" to the problems of assessment. Just because a test is universally designed does not mean that the test is accessible to all students. Changes that might make a test more accessible to one group of students might actually make it less accessible to another group of students. The principles of universal design can be a useful tool for developing better assessments, but they are not something that can magically make all tests accessible to all students. The challenge of finding the proper balance for universally designed paper and pencil tests translates as well to computer-based assessments. It is difficult to anticipate what accessibility issues will arise when a test is delivered on a variety of different systems with a variety of assistive technologies (e.g., screen readers). Trying to anticipate these issues is important, however, and trying to design computer-based assessments in a more universally accessible manner is an endeavor worth pursuing.

Despite the challenges, the potential opportunities to be gained from developing universally designed assessments are numerous. With the emphasis on universally designed assessments, guidance for item development is becoming clearer and more systematic, with specific criteria for test and item developers as well as item reviewers to consider. The criteria that define good items are easier to understand than item difficulty statistics and more engaging for item reviewers, and present an opportunity for bringing more people to the table in the early stages of test design, including those familiar with disability, language acquisition, and technology. Furthermore, the criteria have research-based support in isolation, and now can be subjected to additional research within assessment contexts. In the end, universally designed assessments should open up assessments so that they are more compatible with accommodations, help make assessments more marketable, and truly make the assessments more inclusive of the entire population to be assessed.

Summary

The concept of universally designed assessments is relatively new, and therefore what it actually means is still undergoing clarification. It is likely that the elements of universally designed assessments will be expanded and become more concrete as they are applied to assessment design and development. 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. Universal design opens the door to ways to rethink assessments to ensure that it is not the assessment itself that produces barriers to improved 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

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.

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.

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

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

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

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

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

Hitchcock, C. (2001). Balanced instructional support and challenge in universally designed learning environments. *Journal of Special Education Technology, 16* (4). Retrieved October 2002, from the World Wide Web: http://jset.unlv.edu/16.4/hitchcock/first.html

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

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.

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.

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

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

Thompson, S. J., Johnstone, C. J., & Thurlow, M. L. (2002). *Universal design applied to large scale assessments* (Synthesis Report 44). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved October 2002, from the World Wide Web:http://education.umn.edu/NCEO/OnlinePubs/Synthesis44.html

Thompson, S.J., Quenemoen, R.F., Thurlow, M.L., & Ysseldyke, J.E. (2001). *Alternate assessments for students with disabilities*. Thousand Oaks, CA: Corwin Press.

Thompson, S., & Thurlow, M., (2002). *Universally designed assessments: Better tests for everyone!* (Policy Directions No. 14). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved October 2002, from the World Wide Web: http://education.umn.edu/NCEO/OnlinePubs/Policy14.htm

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., Quenemoen, R., Thompson, S., & Lehr, C. (2001). *Principles and characteristics of inclusive assessment and accountability systems* (Synthesis Report 40). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved October 2002, from the World Wide Web: http://education.umn.edu/NCEO/OnlinePubs/Synthesis40.html

Thurlow, M.L., Lazarus, S., Thompson, S., & Robey, J. (2002). *2001 state policies on assessment participation and accommodations* (Synthesis Report 46). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved October 2002, from the World Wide Web: http://education.umn.edu/NCEO/OnlinePubs/Synthesis46.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.

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.

Author Note

The National Center on Educational Outcomes is supported 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.

This article is based on information contained in two publications of the National Center on Educational Outcomes: *Universal design applied to large scale assessments* (Synthesis Report 44), and *Universally designed assessments: Better tests for everyone!* (Policy Directions No. 14).

Address correspondence to Sandra Thompson, National Center on Educational Outcomes, University of Minnesota, 350 Elliott Hall, 75 East River Road, Minneapolis, Minnesota 55455, email: thomp178@umn.edu

Author Bios

Sandra Thompson, Ph.D. serves as a Research Associate at the National Center on Educational Outcomes (NCEO) at the University of Minnesota, where she is responsible for a number of research activities that document the participation and performance of students with disabilities in state and district standards and assessments. Dr. Thompson has an extensive background in preparing students with disabilities for successful adult lives, spending 10 years with Minnesota's department of education and 10 years as a special education teacher, working with students with developmental disabilities.

Martha Thurlow, Ph.D. is the Director of the National Center on Educational Outcomes (NCEO) at the University of Minnesota. In this position, she addresses the implications of contemporary U.S. policy and practice for students with disabilities, including national and statewide assessment policies and practices, standards-setting efforts, and graduation requirements. Dr. Thurlow has conducted research involving special education for the past 30 years in a variety of areas, including assessment and decision making, learning disabilities, early childhood education, dropout prevention, effective classroom instruction, and integration of students with disabilities in general education settings.

David Malouf, Ph.D. is a Research Analyst with the Office of Special Education Programs in the U.S. Department of Education. His responsibilities include projects related to large-scale assessment, accountability, accommodations, and alternate assessments. He also works with the technology and media services program for students with disabilities. Before coming to the Department of Education in 1991, Dr. Malouf was a faculty member and researcher at the University of Maryland, and taught school at elementary and secondary levels in Salt Lake City, Utah.