Inclusive Design for Maximum Accessibility: A Practical Approach to Universal Design

Elizabeth I. Hanna

Using testing and assessment to promote learning

Pearson Educational Measurement is the largest commercial processor of student assessments. We provide services and products in support of assessment programs to local and state education agencies as well as other assessment organizations and publishers. We understand how assessment activities can promote learning and benefit students, teachers, parents and schools.

PEM Research Reports provide dissemination of PEM research prior to publication. PEM reports in .pdf format may be obtained at:

Abstract

The purpose of this article is to briefly review the literature related to Universal Design for Learning (UDL) and Universal Design for Assessment (UDA), and outline an approach for combining these two philosophies in evaluating large-scale assessment programs. The article begins with a brief history of universal design, followed by a discussion of a planning approach to UDL and UDA in assessment programs that is divided into three categories: the construct of the assessment, the use of the assessment, and the accommodations provided for the assessment. Finally, the article discusses some of the psychometric implications of UDL and UDA, specifically those related to test scaling and comparability.

Inclusive Design for Maximum Accessibility: A Practical Approach to Universal Design

Introduction

Federal law requires states to design and implement large-scale assessment programs in which all but a small percentage of students must participate. Measurement data from these assessments assist in determining the effectiveness of instructional programs in schools, districts, and states. These large-scale assessments require the participation of populations with unique educational needs, varying cultural experiences, diverse linguistic backgrounds, and numerous disability challenges. The *Standards for Educational and Psychological Testing* (AERA, APA, and NCME, 1999) state that the goal of standardized assessment is "to provide accurate and comparable measurement for everyone, and unfair advantage to no one. The degree of standardization is dictated by that goal, and by the intended use of the test." Assessment professionals face the challenge of ensuring that test performance by all students, including those with disabilities and English language learners (ELL), is a valid and reliable measure of their knowledge and skills (Johnstone, 2003). No quick-fix methods can satisfy this complex goal, but several logical and moderate approaches to design and implementation can improve assessment accessibility for all students. Two educational philosophies that organize these approaches and promote inclusion of all students are Universal Design for Learning (UDL) and Universal Design for Assessment (UDA). Examining and implementing these two theories in large-scale assessments will expand their accessibility to all students.

The purpose of this article is to briefly review the literature related to UDL and UDA and to outline an approach for combining these two philosophies in evaluating large-scale assessment programs. The article begins with a brief history of universal design, followed by a discussion of a planning approach to UDL and UDA in assessment programs that is divided into three

1

categories: the construct of the assessment, the use of the assessment, and the accommodations provided for the assessment. Finally, the article discusses some of the psychometric implications of UDL and UDA, specifically those related to test scaling and comparability.

Universal Design—A Brief History

UDL and UDA are derived from the universal design principles that began in the field of architecture. Passage of various laws, most notably the Americans with Disabilities Act, brought an awareness of the environmental limitations experienced by some individuals. Designers began to recognize that some people have difficulty opening doors, reaching light switches, or accessing restrooms. Efforts to improve the disparity between individuals with and without disabilities have been underway since federal legislation began in the 1960s. Rose and Meyer (2000) noted that "a movement grew around the idea of designing buildings from the outset to be accessible to everyone." The term "universal design" was coined by Ron Mace to describe this movement. Universally designed structures are built from the beginning to accommodate the widest spectrum of users, including those with disabilities (Rose and Meyer, 2000). Many features we are now accustomed to in our buildings and public streets came out of this movement. Such innovations include curb cuts on sidewalks, advanced warning of elevator approaches, and improved lighting in buildings. Rose and Meyer (2000) conclude that "Universal Design does not imply 'one size fits all' but rather acknowledges the need for alternatives to suit many different people's needs." This movement has influenced many other fields, as Thompson and Thurlow (2002) noted, such as in environmental initiatives, recreation, the arts, health care, and education.

Universal Design for Learning (UDL)

The application of universal design to education started with exploring instructional methods and the impact of those methods on student learning:

> To many people the term seems to imply that UDL is a quest for a single, one-size-fits-all solution that will work for everyone. In fact, the very opposite is true. The essence of UDL is flexibility and the inclusion of alternatives to adapt to the myriad variations in learner needs, styles, and preferences (Rose and Meyer, 2000, p.4).

Rose (2001) defined the three basic principles of UDL. Each principle aims to minimize barriers and maximize learning by flexibly accommodating individual differences in recognition, strategy, or effect.

1. To support diverse recognition networks, providing multiple, flexible methods of presentation.

2. To support diverse strategic networks, providing multiple, flexible methods of expression and apprenticeship.

3. To support diverse affective networks, providing multiple, flexible options for engagement.

The above principles cannot be directly transferred to assessment, but it is possible to find value in the essence of the principles. The overall goal of UDL is a flexible approach to access learning. Rose and Meyer (2000) write that "the 'universal' in Universal Design for Learning does not imply a single solution for everyone, but rather it underscores the need for inherently flexible, customizable content, assignments, and activities."

Universal Design for Assessment (UDA)

As large-scale assessment programs strive to include the broadest range of students possible, flexibility in the assessment instrument and testing environment is a necessity. However, this approach presents challenges: "Each step toward greater flexibility almost inevitably enlarges the scope and magnitude of measurement error. However, it is possible that some of the resultant sacrifices in reliability may reduce construct irrelevance or construct underrepresentation in an assessment program" (AERA, APA, and NCME, 1999, p. 26).

UDA proposed by the National Center on Educational Outcomes (NCEO) is one approach to flexibility in large-scale assessments: "'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" (Thompson, Johnstone, and Thurlow, 2002, p. 6).

All students can "sit" for an assessment, but the design of the assessment may preclude students from meaningful, valid participation to demonstrate knowledge and skills according to the state standards. NCEO suggests seven elements that assessment professionals should consider when creating universally designed assessments. In combination the seven elements provide a framework for examining tests and their level of accessibility. Individual items may not have direct links to all elements, but UDA as described below is a general framework for improving the design (and thereby accessibility, comprehensibility, and validity) of tests (Johnstone, 2003).

The elements of universal design adapted from Thompson and Thurlow (2002) and Thompson, Johnstone, and Thurlow (2002) include:

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. Assessments can measure the performance of students with a wide range of abilities, allowing opportunities to demonstrate competence on the same content.

Precisely Defined Constructs—The specific constructs tested must be clearly defined so that all construct-irrelevant cognitive, sensory, emotional, and physical barriers are removed. Assessments should measure what they are intended to measure. Item design offers the broadest range of success within the determined constructs.

Accessible, Non-Biased Items—Accessibility is built into items from the beginning, and bias review procedures ensure that quality is retained in all items. The purpose of bias review is to examine items for advantages or disadvantages in presentation or content which invalidate the item's contribution to a test score.

Amendable to Accommodations—Test design facilitates the use of needed accommodations (e.g., all items can be brailled) and reduces threats to validity and comparability of scores.

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 kept to a minimum) for readable and comprehensible text. Student background, sentence

difficulty, and organization of text all contribute to readability of the text, and these points can be considered in item development.

Maximum Legibility—Characteristics that ensure easy decipherability are applied to text, tables, figures, and illustrations, and to response formats. Legibility is the actual appearance of text which enables people to read it easily.

Current assessment programs utilize many of the UDA elements proposed by NCEO in varying degrees. The following sections outline how large-scale assessment programs can combine the philosophies of UDL and UDA to evaluate current programs for preexisting, inclusive testing elements, identify areas where barriers still exist for examinees, and set priorities for maximum inclusiveness.

Combining Universal Design for Learning and Universal Design for Assessment—An Approach for Planning

The approach for planning is divided into three categories: construct of the assessment, use of the assessment, and accommodations for the assessment. This flexible design allows the participation of students in a standardized measure while maintaining state testing program priorities. To allow for inclusive design of standardized tests that is valid and reliable, policy planning before the operational administration is advised.

Construct of the Assessment

One principle of UDL—engagement—and one of the elements of UDA—precisely defined constructs—combine to form a broad category called construct of the assessment. An opportunity to demonstrate mastery offers students a way to engage with the assessment, and the

interpretation of scores from these assessments can show that a majority of students are engaged in the assessment program. As stated in professional testing standards,

> Standard 1.2. The test developer should set forth clearly how test scores are intended to be interpreted and used. The population(s) for which a test is appropriate should be clearly delimited, and the construct that the test is intended to assess should be clearly described (AERA, APA, and NCME, 1999, p. 17).

Assessment design involves two tasks: it determines the constructs to be tested (what to assess), and it determines how the constructs will be assessed to provide maximum accessibility to the population of required test takers. The assessed curriculum is not defined based on the needs of persons with disabilities or ease of mastery, but the skills to be assessed are derived from the state curriculum. It is important to recognize that portions of the assessed state curriculum may be biased against a particular group, and such bias may be difficult to avoid if the curriculum to be measured is mandated by state law. An example of this would be a curriculum strand related to visual representation that could potentially exclude students with low vision or blindness. *How* these constructs are assessed can help reduce bias and possible construct irrelevancy. Construct irrelevancy can occur when an item contains excess components that are irrelevant to the construct being assessed. Assessments should be designed to allow the largest number of persons an opportunity to demonstrate mastering of skills and understanding of concepts contained in educational goals and objectives (American Foundation for the Blind, 2003).

This process of construct validity and accessibility broadly progresses through the following categories:

- Determination of the state curriculum to be assessed.

- Determination of how this content will be assessed, with careful attention to the design principles most accessible to special populations of students.

- After the first two steps have been combined to form actual items, the content deemed appropriate to assess in the first step is still present in the items.

Smisko et al., (2000) elaborated on these broad categories with six detailed steps to plan the content and curricular validity evidence of score interpretations resulting from an assessment.

Step 1: Review curriculum: The first step in the test development process is the review of the state-mandated curriculum.

Step 2: Develop objectives: After reviewing the state curriculum, committees of educators can work with the state to draft test objectives.

Step 3: Refine objectives: Based on input from the review of the draft test objectives, the draft test objectives are refined and the final test objectives are determined.

Step 4: Write sample items: Sample test items are constructed to serve as exemplars of how the objectives will be measured.

Step 5: Develop item guidelines: Using the test objectives and the sample items as a guide, committees of educators assist the state in developing guidelines for assessing each objective.

Step 6: Develop preliminary blueprint: With educator input, a preliminary test blueprint is developed that sets the length of the test by grade and subject and lists the number of items per objective.

To help avoid problems such as excess components in an item and construct-irrelevancy, and to increase accessibility, particular attention should be given to Steps 4 and 5. Committees of educators, experts in the field of special education, English language learners, content, and

assessment, as well as representatives of special interest groups, can convene to discuss the item guidelines or specifications and provide input into designing items based on issues regarding special populations. Having this step occur after review of the curriculum and objectives will help reduce construct-irrelevance and defend construct validity evidence of score interpretations: "Careful review of the construct and test content domain by a diverse panel of experts may point to potential sources of irrelevant difficulty (or easiness) that require further investigation" (AERA, APA, and NCME, 1999, p. 12).

However, as these groups convene to determine specifications for developing more accessible items, therefore decreasing certain types of construct irrelevance, certain changes in item types (such as the use of plain language) may inadvertently add construct irrelevancy. Two types of irrelevance can occur: construct-irrelevant difficulty and construct irrelevant easiness.

In general, construct-irrelevant difficulty leads to construct scores that are invalidly low for those individuals adversely affected (e.g., knowledge scores of poor readers). In contrast, construct-irrelevant easiness occurs when extraneous clues in item or test formats permit some individuals to respond correctly in ways irrelevant to the construct being assessed (Messick, 1989, p. 34).

As states decide to augment the format or "feel" of items to make them more accessible to the largest group of individuals, a balance should be maintained between the desire to change items to be inclusive and the desire to make credible inferences from test data. Construct-related evidence is based on the accumulation of empirical evidence that:

1. the hypothetical construct being measured actually exists, and

2. the assessment device in use does, in fact, measure that construct (Popham, 2000).

It is advised that states receive continuous technical support when planning and developing new or expanded assessments to provide thorough guidance in construct validity.

Use of the Assessment—Recognition

Recognition, to perceive clearly, is one of the three main principles of UDL. To recognize something, representation of the object must be clearly brought before the mind. For a student to use an assessment to show knowledge, the assessment should be easily perceived and clear to the test-taker. Four of the seven elements of UDA fall into this category. They include accessible and non-biased items; simple, clear, and intuitive instructions and procedures; maximum readability and comprehensibility; and maximum legibility. A brief description and a set of decision flowcharts follow for each element.

Accessible, Non-Biased Items

The first element of UDA that enables a student to have recognition is the accessible and non-biased nature of items. Assuming that a student has had adequate preparation, the items should be accessible to the student. Items unfairly penalize a group of students if those students perform less well on the item than another group of students, though both groups are at the same achievement level with respect to the knowledge or skills being tested (Popham, 2000). Students come to the assessment experience with disabilities, language differences, and varied cultural views. Ideally these differences and unique attributes should not hinder a student's ability to interact with an item.

Regardless of the purpose of testing, fairness requires that all examinees be given a comparable opportunity to demonstrate their standing on the construct(s) the test is intended to measure (AERA, APA, and NCME, 1999. p. 74).

Accessibility and biases cannot be completely eliminated due to the variance of student experiences, but through thoughtful attention, items can be written as clearly and fairly as possible. Standard 7.4 states that

> Test developers should strive to identify and eliminate language, symbols, words, phrases, and content that are generally regarded as offensive by members of racial, ethnic, gender, or other groups, except when judged to be necessary for adequate representation of the domain (AERA, APA, and NCME, 1999).

It is important for states to build and maintain relationships with individuals and groups capable of guiding the development of fair test items. Items should undergo accessibility and bias review before appearing on field or pilot tests, and these reviews should occur early enough in the item development process to allow changes to items. Smisko, Twing, and Denny (2000) explain that review committees of educators can revise items where necessary and judge items in terms of content, difficulty, language appropriateness, gender and ethnic role stereotypes, and cultural familiarity.
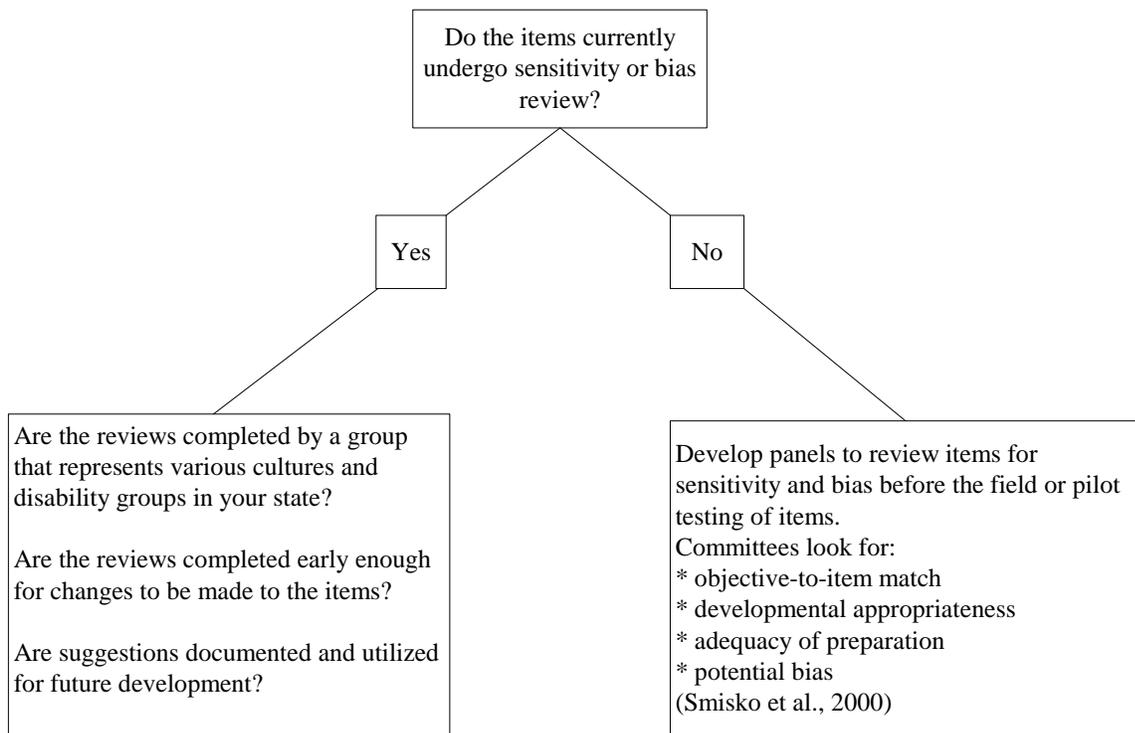
To accomplish this, each item is rated in four areas:

1. objective-to-item match,

2. developmental appropriateness,

3. adequacy of preparation (i.e., the intended examinees have received instruction on this topic), and

4. potential bias (Smisko, et al., 2000).

Smisko, et al., (2000) conclude that this type of review helps confirm that test items are fair for students and maximizes the evidence of content and curricular validity.

By combining early review processes and cultivating individuals with varied backgrounds in culture, language, and disability to review items, sensitive and fair items can be produced. Popham (2000) concludes that if review committees are carefully selected, oriented, and assisted during the review process most biased items can be identified and eliminated.

Decision Flowchart–Accessible, Non-Biased Items

```
                    ┌─────────────────────┐
                    │  Do the items currently │
                    │ undergo sensitivity or bias │
                    │        review?       │
                    └─────────────────────┘
                       /              \
                ┌──────┐            ┌──────┐
                │ Yes  │            │  No  │
                └──────┘            └──────┘
               /                          \
```

┌────────────────────────────────────┐   ┌────────────────────────────────────┐
│ Are the reviews completed by a group │   │ Develop panels to review items for   │
│ that represents various cultures and │   │ sensitivity and bias before the field or pilot │
│ disability groups in your state?   │   │ testing of items.   │
│   │   │ Committees look for: │
│ Are the reviews completed early enough │   │ * objective-to-item match   │
│ for changes to be made to the items? │   │ * developmental appropriateness   │
│   │   │ * adequacy of preparation   │
│ Are suggestions documented and utilized │   │ * potential bias   │
│ for future development?   │   │ (Smisko et al., 2000)   │
└────────────────────────────────────┘   └────────────────────────────────────┘

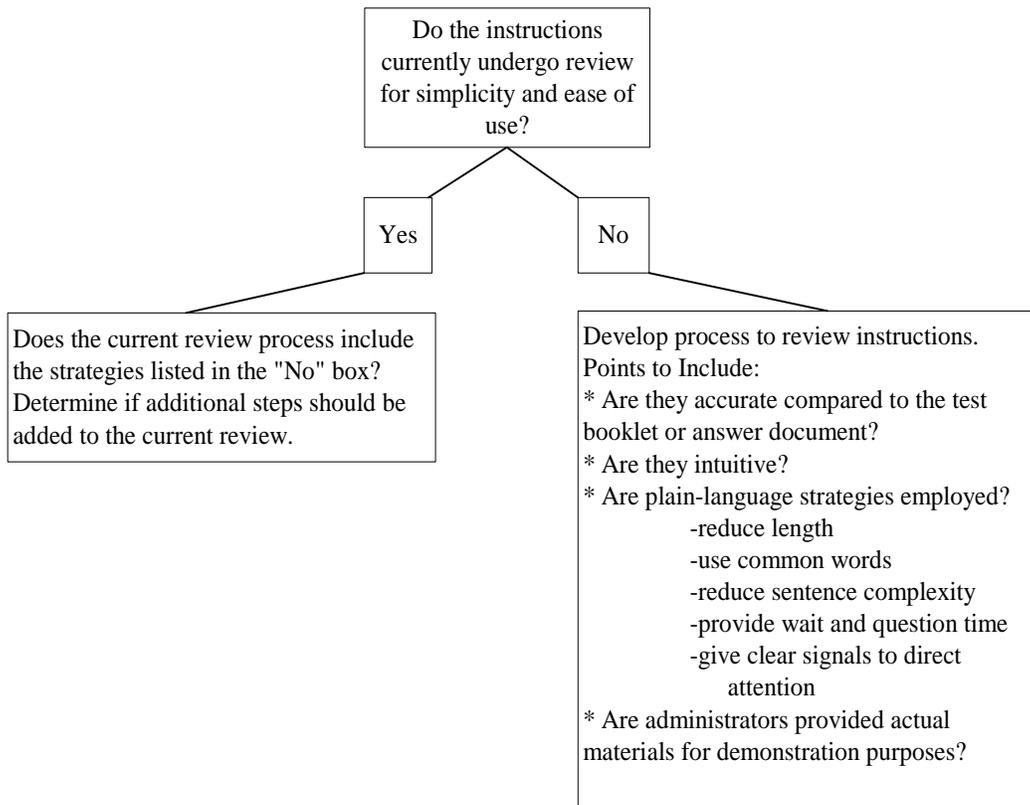**Simple, Clear, and Intuitive Instructions and Procedures**

Instructions and procedures must be simple, clear, and intuitive, allowing the maximum number of students to recognize what to do during a test:

Standard 3.20. The instructions presented to test takers should contain sufficient detail so that test takers can respond to a task in the manner that the test developer intended. When appropriate, sample material, practice or sample questions, criteria for scoring, and representative items 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 (AERA, APA, and NCME, 1999, p. 47).

A review process should be implemented to ensure that all materials and directions associated with a test administration are simple and easy to use. Instructions should be reviewed for accuracy compared to the test booklet and/or answer document. Implementing plain-language strategies should also be considered. Such strategies, adapted from Brown (1999), are to reduce length of the instructions, use common words, and decrease sentence complexity and length. Appropriate "wait time" should be in the script to allow students time to comprehend what is being said. Students should also have multiple opportunities for questions regarding administration instructions throughout the scripted presentation. When directing attention to particular portions of the test booklet or answer document, signals should be clear, and administrators should be allowed to demonstrate with actual test materials, when appropriate.

Decision Flowchart—Simple, Clear, and Intuitive Instructions

```
                    ┌─────────────────────┐
                    │ Do the instructions │
                    │ currently undergo   │
                    │ review for          │
                    │ simplicity and ease │
                    │ of use?             │
                    └─────────────────────┘
                       ╱              ╲
                  ┌───────┐       ┌───────┐
                  │  Yes  │       │  No   │
                  └───────┘       └───────┘
```

Does the current review process include the strategies listed in the "No" box? Determine if additional steps should be added to the current review.

Develop process to review instructions. Points to Include:
* Are they accurate compared to the test booklet or answer document?
* Are they intuitive?
* Are plain-language strategies employed?
      -reduce length
      -use common words
      -reduce sentence complexity
      -provide wait and question time
      -give clear signals to direct
        attention
* Are administrators provided actual materials for demonstration purposes?

Maximum Readability and Comprehensibility

Standard 7.7 states that,

In testing applications where the level of linguistic or reading ability is not part of the construct of interest, the linguistic or reading demands of the test should be kept to the minimum necessary for the valid assessment of the intended construct" (AERA, APA, and NCME, 1999).

One approach to achieving maximum readability and comprehensibility is the use of plain-language strategies. Brown (1999) lists multiple strategies used in the plain-language style.

Such strategies include:

- reduce excessive length,

- eliminate unusual or low-frequency words and replace with common words,

- avoid ambiguous words,

- avoid irregularly spelled words,

- avoid proper names,

- avoid inconsistent naming and graphic conventions,

- avoid unclear signals about how to direct attention, and
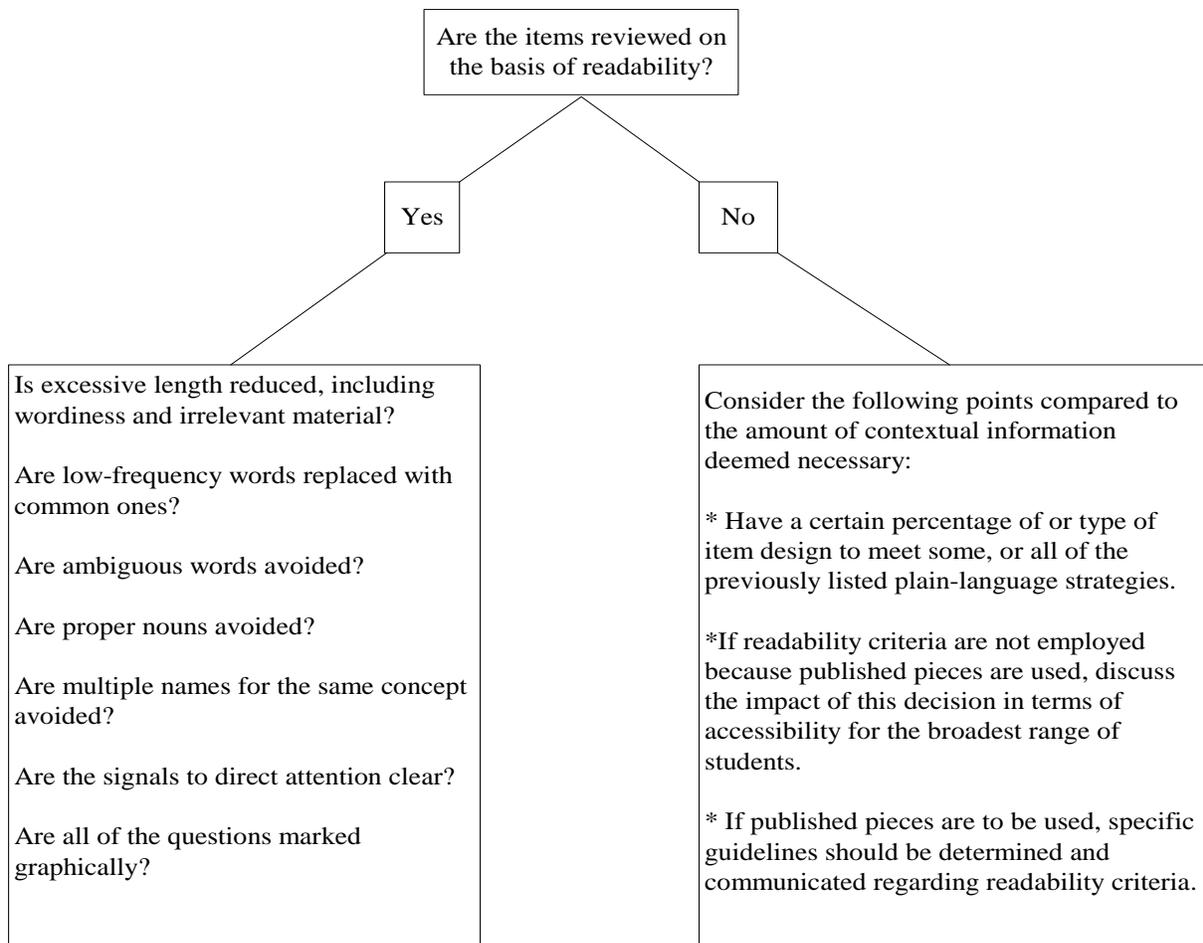
- mark all questions.

Research suggests that plain-language strategies will not hurt students, but a comprehensive look for consistent benefits is needed. In 2002, Stansfield completed a review of literature on simplified English as an accommodation for limited English proficient (LEP) students. At the end of the review, he concluded,

> The results of this study suggest that if test developers and researchers are careful in carrying out linguistic simplification, the resulting assessment could address the linguistic needs of the LEP students without compromising the comparability of the scores obtained on the assessment by taking the standard English version.

Rivera and Stansfield (2001) studied the effects of linguistic simplification of science items on limited English proficient students and monolingual students. The LEP samples were too small to provide "generalizeable results," but the non-LEP groups "were more than adequate for analysis and interpretation." The results of the study support the conclusion that among fully

English-proficient students, linguistically simplified items normally do not help students taking a test (Rivera & Stansfield, 2001). Rivera and Stansfield (2001) summarize that "the results of the process of linguistic simplification must be to make the items accessible to ELLs while not altering the difficulty of the content." This balance is the true challenge, and it can be met through detailed planning and item development specifications. Readability and comprehensibility are affected by many characteristics, including student background, sentence difficulty, and text organization. These features need to be considered in assessment development (Thompson and Thurlow, 2002).

Decision Flowchart—Maximum Readability

```
                    ┌─────────────────────┐
                    │ Are the items reviewed on │
                    │  the basis of readability? │
                    └─────────────────────┘
                         /            \
                    ┌───────┐      ┌──────┐
                    │  Yes  │      │  No  │
                    └───────┘      └──────┘
                      /                  \
```

┌─────────────────────────────────┐   ┌─────────────────────────────────┐
│ Is excessive length reduced, including │   │ Consider the following points compared to │
│ wordiness and irrelevant material?   │   │ the amount of contextual information │
│   │   │ deemed necessary: │
│ Are low-frequency words replaced with │   │   │
│ common ones?   │   │ * Have a certain percentage of or type of │
│   │   │ item design to meet some, or all of the │
│ Are ambiguous words avoided?   │   │ previously listed plain-language strategies. │
│   │   │   │
│ Are proper nouns avoided?   │   │ *If readability criteria are not employed │
│   │   │ because published pieces are used, discuss │
│ Are multiple names for the same concept │   │ the impact of this decision in terms of │
│ avoided?   │   │ accessibility for the broadest range of │
│   │   │ students.   │
│ Are the signals to direct attention clear? │   │   │
│   │   │ * If published pieces are to be used, specific │
│ Are all of the questions marked   │   │ guidelines should be determined and │
│ graphically?   │   │ communicated regarding readability criteria. │
└─────────────────────────────────┘   └─────────────────────────────────┘

Maximum Legibility

Thompson and Thurlow (2002) define maximum legibility as "the physical appearance of text, the way that the shapes of letters and numbers enable people to read text easily." Thompson, Johnstone, and Thurlow (2002) listed dimensions to consider when adapting materials to be legible and accessible for readers. These dimensions include contrast, type size, spacing, leading, typeface, justification, line length, blank space, graphs and tables, illustrations, and response formats. It should be noted that additional adaptations may be necessary for persons with low vision. Experts in this field should be consulted in test design for specific issues related to legibility for persons with low vision.

The following questions may assist states to determine the legibility of their items based on recommendations from Thompson, Johnstone, and Thurlow (2002):

What color type and paper are used? Recommendation–Use black type on pastel or off-white paper.

What type size is used? Recommendation–Use 14-point type to increase readability. Large print for students with visual impairment is at least 18-point.

What spacing is used between letters and words? Recommendation–Use fixed-spaced fonts.

What leading is used for the vertical space between the lines of type? Recommendation–Leading varies depending on the type size; 14-point type needs 3–6 points.

What typeface is used? Recommendation–Avoid italic, slanted, small caps, or all caps.

What justification is used? Recommendation–Use staggered right margins (unjustified).

What line lengths are used? Recommendation–Use line lengths of 4 inches, or 8 to 10 words per line.

How much blank space is used? Recommendation–Occupy about half the page with text.

What are the components of graphs and tables? Recommendation–All symbols should be highly discriminable; labels should be placed directly next to plotlines.

How are illustrations used? Recommendation–Use illustrations only when they contain information being assessed and locate them directly next to the question.

How do students respond? Recommendation–All response options should include larger circles for a bubble response test.

Arditi (2003) includes these additional legibility factors to consider:

Font Family–Avoid complicated, decorative or cursive fonts. When such fonts must be used, reserve for emphasis only. Standard serif or sans serif fonts, with familiar, easily recognizable characters are best. Also, there is some evidence that sans serif fonts are more legible when character size is small relative to the reader's visual acuity.

Paper Finish–Paper with a glossy finish can lessen legibility because many people who are older or who have partial sight also have problems with glare.

Allman (2003) provides further information regarding graphic material when testing students with visual impairments. Certain types of graphic material (maps, charts, graphs, diagrams, and illustrations) cannot be provided in braille or tactile formats. Allman (2003) suggests including verbal descriptions of illustrations and simplified artist features to allow tactile representation for the visually impaired. Thompson, Johnstone, and Thurlow (2002)

18

noted that complex "illustrations also may complicate the use of magnifiers, enlargement, or other assistive technology."

Decision Flowchart—Maximum Legibility

Are items reviewed based on legibility criteria?

Yes

No

Does the current review include the characteristics of maximum legibility?
* contrast of type on paper
* larger type size
* spacing between characters
* leading, or the amount of vertical space between lines of type
* use of standard typeface
* use of unjustified text
* limit length of line
* provide adequate blank space
* graphs and tables use highly discriminable symbols and quickly found labels
* illustrations used when they contain information being assessed
* response formats include larger circles and varied forms of response

Consider employing legibility criteria in item development. Prioritize criteria based on student need and overall feasibility.

Adapted from Brown (1999) and Thompson, Johnstone, and Turlow (2002).

Test developers can compare their current formatting to the above recommendations to determine where changes can be implemented. Recognition and use of the assessment are important aspects of design that can assist in providing more meaningful access to students. However, changes in format may have consequences throughout a testing program, such as item-specification changes leading to equating issues. Format changes may result in time or cost increases related to test booklet length or other product specifications. Assessment professionals

19

must include these issues when considering the positive effects these changes may have on the tested population.

Accommodations for Assessment

Accommodations may be needed for students to effectively express themselves. To express is to show or reflect a true expression of something. To accommodate is to make fit, suitable, or congruous. Dolan and Hall (2001) stipulate that "unless a student with disabilities is provided access during testing to the supports they rely on in the classroom, they may not be able to show their knowledge and understanding." Expression is the third principle of UDL, and being amenable to accommodations is the final element of UDA. The term "expression" is used in UDL to convey the idea that students need multiple ways to express themselves and the knowledge they possess in meaningful ways. As Donlan and Hall state:

> Whether assessment is embedded into teaching (e.g. curriculum-based measurement) or administered separately (e.g. large-scale assessment), it must provide students with adequate and equitable means to express their knowledge and understanding if it is to provide accurate feedback on the performance of students (Dolan and Hall, 2001).

Standardized assessment design does not allow for multiple forms of expression based on the needs of individual students, but the use of accommodations is one way students can express themselves when taking assessments. If a student expressed him or herself in the classroom while employing a routine accommodation, then the same routine accommodation would be necessary for expression on an assessment. If all things were equal, unrestricted accommodations on standardized assessments would be the most advantageous for students. Unfortunately unrestricted accommodations do not fit the parameters of standardization.

Standard 10.1. In testing individuals with disabilities, test developers, test administrators,

and test users should take steps to ensure that the test score inferences accurately reflect

the intended construct rather than any disabilities and their associated characteristics

extraneous to the intent of the measurement (AERA, APA, and NCME, 1999, p. 106).

The goal of an accommodation is to benefit only those students who need it, without effecting

students with no relevant special needs (Shaftel, Belton-Kocher, Glasnapp, and Poggio, 2003).

As noted in the testing *Standards* (AERA, APA, and NCME, 1999), accommodation is not an

easy issue to rectify due to the lack of specific research regarding use on various populations and

the impact on validity and reliability.

There have been few empirical investigations into the effects of various accommodations

on the reliability of test scores or the validity of inferences drawn from modified tests.

Due to a number of practical limitations (e.g., small sample size, nonrandom selection of

test takers with disabilities), there is no precise, technical solution available for equating

modified tests to the original form of these tests (AERA, APA, and NCME, 1999, p.

102).

Because of these issues, some students with disabilities have been excluded from

participation in large-scale assessments (Thompson, Johnstone, and Thurlow, 2002). Much of the

literature regarding universal design expresses an opinion that many accommodation issues

could be resolved by the initial design of the assessment, and that UDA is the most beneficial

type of design. Examining this idea in comparison with the most commonly used

accommodations around the country helps determine whether assessment design will impact the

use of or implementation of certain accommodations. This comparison may help states focus on

areas where they can improve design or policy to accommodate the needs of students. The chart

on the following pages lists commonly used accommodations and factors influencing use of

certain accommodations, such as item design or state policy.

	Use of the Accommodation Affects the Item Design	Use of the Accommodation is Determined by State Policy	Neither Item Design nor State Policy
Accommodations That Alter Test Presentation			
Large Print⸺all parts of the assessment in large print	X		
Braille⸺assessment presented in braille	X		
Read Aloud⸺all or portions of assessment read aloud	X	X	
Sign Interpretation⸺all or portions of assessment presented via sign language	X	X	
Read/Reread/Clarify⸺directions may be clarified		X	
Visual Cues⸺additional visual clues are provided		X	
Administered by Others⸺someone other than regular test administrator			X
One item per page of the test booklet		X	
Equipment and Material Accommodations			
Magnification Equipment⸺enlarges text print size			X
Amplification Equipment⸺increases sound level during the test			X
Light/Acoustics⸺changes lighting or acoustics			X
Calculator	X	X	
Templates/Graph Paper⸺mark location of focus on the test			X
Audio/Video Cassette		X	
Noise Buffer			X
Adaptive or Special Furniture			X
Abacus		X	
Accommodations That Alter Test Response			
Proctor/Scribe⸺student responds verbally; answers translated to answer document			X
Computer or Machine⸺uses device to respond			X
Write in Test Booklet⸺answers transcribed to document			X
Tape Recorder⸺responses recorded then transcribed			X
Communication Device⸺various devices for giving responses			X
Spell-check/Assistance		X	
Brailler⸺generates responses in braille			X
Pointer⸺student points and personnel translates answer			X
Accommodations That Alter Timing or Scheduling of Assessment			
Extended Time⸺more time than typically allowed	X	X	
With Breaks⸺time away from test		X	
Multiple Sessions⸺test broken into more than one session		X	
Time Most Beneficial to Student⸺administered at a time most beneficial to student		X	
Over Multiple Days⸺when usually administered on one day		X	

	Use of the Accommodation Affects the Item Design	Use of the Accommodation is Determined by State Policy	Neither Item Design nor State Policy
Accommodations That Alter the Test Location or Environment			
Individual—assessed separately from other students			X
Small Group—assessed in small group from other students			X
Carrel—student seated in study carrel			X
Separate Room—assessed in separate room			X
Seat Location/Proximity—assessed in a specifically designated seat location			X
Minimize Distractions/Reduces Noise—assessed in quiet environment			X
Student's Home—assessed in home			X
Special Education Classroom—assessed in special education classroom			X

The listed accommodations are adapted from Thurlow, Lazarus, Thompson, and Robey (2002).

The majority of the accommodations listed are not affected by the design of the items, or if the item's design is changed to be more universal, many of the same accommodations are still needed by the student. The design would not eliminate need for the accommodation.

This comparison does not discount the need for more accessible, accommodation-friendly assessments. Rather, it focuses attention on the need to make accommodation decisions—both those affected by design and those unaffected—before the assessment is operational, rather than after the fact when modifying is the only viable option. There should be a balance between testing the construct in as barrier-free a way as possible, allowing for accommodations, and maintaining validity and reliability. If tests are universally designed, they must be created with consideration for a broad student population that includes students with disabilities and limited English proficiency. Otherwise, the benefit of accommodations for students will be limited to what can typically be achieved with retrofit solutions (Dolan and Hall, 2001). Accommodations and the construct of the assessment are interrelated: testing professionals should act now to allow appropriate use of an accommodation later.

Further, if the test delivery systems and the individual components that make up each test item are not accessible, then no amount of accommodations, assistive technology, or time will make the item or the test accessible. It may not be economically feasible to retrofit an existing test to make it accessible. In order to create an accessible test, accessibility must be a consideration from the beginning (Allan, Bulla, and Goodman, 2003, p. 15).

Testing professionals must decide what accommodations they will allow students to use when taking assessments, and these decisions, in tandem with which constructs to assess, should drive the decision-making process regarding test item design. If test developers want to use published works as passages or if complex contextual support for each item is provided, then many of the strategies (such as reduced sentence length) would be difficult to implement. If testing professionals feel it is important for students to use calculators on certain portions or all of the assessment, items should be written with calculator use in mind. Items and accommodations should work together to provide students with an accessible assessment.

Scaling and Comparability

The principles of universal design offer a starting point for discussions regarding accessible assessments for the broadest range of students. The ideal would be to stop current testing, draw up new, more accessible design plans, and implement a redesigned program. This, however, is not practical due to time constraints and funding limitations. Changes to testing programs, such as retrofitting items to match new item specifications, may result in an unwanted impact on the established scale, and interpretations of assessment results may no longer be valid. Small changes in item or test design could result in a change in the trend line that might falsely be interpreted as a variation in instruction or achievement. As the testing standards state,

24

Major shifts sometimes occur in the specifications of tests that are used for substantial periods of time. Often such changes take advantage of improvements in item types or of shifts in content that have been shown to improve validity and, therefore, are highly desirable. It is important to recognize, however, that such shifts will result in scores that cannot be made strictly interchangeable with scores on an earlier form of the test (AERA, APA, and NCME, 1999, p. 59).

Rationale

In most assessment systems, the scaling and equating process is based on the Item Response Theory (IRT), which Popham (2000) describes as the consideration of the difficulty and other technical properties of each item on a test. Mehrens and Lehmann (1991) elaborate on the basic advantage of IRT, which is that if the mathematical model used is maintained, the difficulty values of the scaled items do not depend on the particular sample of students in the standardization group, and the ability estimates of the students are independent of the sample of items administered. Once items are scaled, they can be compared to the scores of students on some characteristics even though all the students did not take the same item.

Using IRT, scores can be statistically adjusted so that they can be used interchangeably across different test forms and years. This adjustment occurs through equating, as described by Kolen and Brennan (2004):

The process of equating is used in situations where such alternate forms of a test exist and scores are earned on different forms are compared to each other. Even though test developers attempt to construct test forms that are as similar as possible to one another in content and statistical specifications, the forms typically differ somewhat in difficulty. Equating is intended to adjust for these difficulty differences, allowing the forms to be

used interchangeably. Equating adjusts for differences in difficulty, not for differences in content (Kolen and Brennan, 2004, p. 3).

Equating adjusts for differences in form difficulty, and form difficulty is determined by item difficulty; therefore, changes in item specifications may make the equating results inaccurate. Any type of change in an item potentially changes how a student interacts with that item. Because of this relationship, item stability is at the foundation of scale equating.

Issues may also result if other content-related factors change, such as the use of time limits or allowing a change in accommodations. The *Standards* (AERA, APA, and NCME, 1999) reiterate that it is not possible to construct conversions between scores on tests that measure different constructs; that differ materially in difficulty, reliability, time limits, or other conditions of administration; or that are designed to different specifications.

Changes in items or testing procedures not only make it difficult to equate tests from year to year, but may invalidate data associated with the existing items in an item bank compared to the data associated with new items. Issues may also arise if a new student group is included in reporting. If the characteristics of respondents change, the difficulty or technical information related to each item and the test as a whole may also change. The effect of these changes may appear minimal, or even trivial, at item level. But the cumulative effect on the reported scaled score may impact examinees in large-scale testing. Prior to altering the population of examinees from which the scaled score is derived, test developers should research the impact on the scale using historical data, and obtain the advice of a technical advisory panel.

The *Standards* (AERA, APA, and NCME,1999) provide clear guidelines related to these issues:

Standard 4.16. If test specifications are changed from one version of a test to a subsequent version, such changes should be identified in the test manual, and an indication should be given that converted scores for the two versions may not be strictly equivalent. When substantial changes in test specifications occur, either scores should be reported on a new scale or a clear statement should be provided to alert users that the scores are not directly comparable with those on earlier versions of the test (AERA, APA, and NCME, 1999, p.59).

Because of the numerous issues related to changes in current testing programs, it is advisable that testing professionals convene a group of technical advisors to assist with balancing these complicated issues.

Conclusion

Inclusive design aims to create assessments that are accessible to the maximum number of students. In the rush to be inclusive, however, the greater goal—accurate and comparable measurement for everyone—must not be jeopardized. Item and test development must also encompass the need for assessment data to be valid and reliable, and the assessments should comply with the law and meet the *Standards for Educational and Psychological Testing*. Resolution of these issues should be moderate and thoughtful.

References

Allan, J. M., Bulla, N., and Goodman, S.A. (2003). *Test access: Guidelines for computer-administered testing*. Louisville, KY: American Printing House for the Blind. Available from: http://www.aph.org

Allman, C.B. (2003). *Making tests accessible for students with visual impairments: A guide for test publishers, test developers, and state assessment personnel.* Louisville, Ky.: American Printing House for the Blind.

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

American Foundation for the Blind. (2003). Comments to the department of education on improving the academic achievement of the disadvantaged. *Federal Register*, March 20, 2003. 68 FR 13795.

Arditi, A. (2003). *Making text legible: designing for people with partial sight.* Lighthouse International. Available from: http://www.lighthouse.org/print_leg.htm

Brown, Pamela J. (1999). *Findings of the 1999 plain language field test* (Publication T99-013.1). Newark, DE: University of Delaware, Delaware Education Research and Development Center.

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

Johnstone, C. J. (2003). *Improving validity of large-scale tests: Universal design and student performance* (Technical Report 37). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved January 9, 2004, from: http://education.umn.edu/NCEO/OnlinePubs/Technical37.htm

Kolen, M. J., and Brennan, R.L. (2004). *Test equating, scaling, and linking: Methods and practices (second edition).* New York: Springer Science & Business Media, Inc.

Mehrens, W.A., and Lehmann, I.J. (1991). *Measurement and evaluation in education and psychology (fourth edition).* Fort Worth: Holt, Rinehart and Winston, Inc.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement (third edition).* New York: American Council on Education: Macmillan.

Popham, W. J. (2000). *Modern educational measurement: Practical guidelines for educational leaders (third edition).* Boston: Allyn and Bacon.

Rivera, C. and Stansfield, C.W. (2001). *The effects of linguistic simplification of science test items on performance of limited English proficient and monolingual English-speaking students.* Paper presented at the annual meeting of the American Educational Research Association, Seattle, WA.

Rose, D. (2001). Universal design for learning: Deriving guiding principles from networks that learn. *Journal of Special Education Technology.* 16 (1). Available from: http://jset.unlv.edu/16.2/asseds/rose.html

Rose, D. and Meyer, A. (2000). Universal design for learning, associate editor column. *Journal of Special Education Technology.* 15 (1). Available from: http://jset.unlv.edu/15.1/asseds/rose.html

Shaftel, J., Belton-Kocher, E., Glasnapp, D. R., and Poggio, J. P. (2003). *The differential impact of accommodations in statewide assessment: Research summary*. Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved February 23, 2004, from:

http://education.umn.edu/NCEO/TopicAreas/Accommodations/Kansas.htm

Smisko, A., Twing, J.S., and Denny, P. (2000). The Texas model for content and curricular validity. *Applied Measurement In Education.* 13(4), 333–342.

Stansfield, Charles W. (2002). Linguistic simplification: A promising test accommodation for LEP students? *Practical Assessment, Research and Evaluation*, 8(7). Retrieved February 23, 2004 from http://PAREonline.net/getvn.asp?v=8&n=7

Thompson, S. J., Johnstone, C. J., and 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 January 9, 2004, from: http://education.umn.edu/NCEO/OnlinePubs/Synthesis44.html

Thompson, S., and 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 February 3, 2004, from: http://education.umn.edu/NCEO/OnlinePubs/Policy14.htm

Thompson, S. J., Thurlow, M. L., Quenemoen, R. F., and Lehr, C. A. (2002). *Access to computer-based testing for students with disabilities* (Synthesis Report 45). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved February 5, 2004, from: http://education.umn.edu/NCEO/OnlinePubs/Synthesis45.html

Thurlow, M.L., Lazarus, S., Thompson, S., and 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 March 23, 2004, from: http://education.umn.edu/NCEO/OnlinePubs/Synthesis46.html