A State Guide to the Development of Universally Designed Assessments

NATIONAL CENTER ON EDUCATIONAL OUTCOMES

In collaboration with:

Council of Chief State School Officers (CCSSO)

National Association of State Directors of Special Education (NASDSE)

Supported by:

U.S. Office of Special Education Programs

A State Guide to the Development of Universally Designed Assessments

Christopher Johnstone • Jason Altman • Martha Thurlow

September 2006

Johnstone, C., Altman, J. & Thurlow, M. (2006). *A state guide to the development of universally designed assessments.* Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

NATIONAL
CENTER ON
EDUCATIONAL
OUTCOMES

The Center is supported through a Cooperative Agreement (#H326G050007) with the Research to Practice Division, Office of Special Education Programs, U.S. Department of Education. This report was supported by a grant (#H324D020050) from the U.S. Department of Education, Office of Special Education Programs, Directed Research Division. The Center is affiliated with the Institute on Community Integration at the College of Education and Human Development, University of Minnesota. Opinions expressed herein do not necessarily reflect those of the U.S. Department of Education or Offices within it.

IDEAs that Work
U.S. Office of Special
Education Programs

NCEO Core Staff

Deb A. Albus
Manuel T. Barrera
Christopher J. Johnstone
Jane L. Krentz
Kristi K. Liu
Ross E. Moen
Michael L. Moore
Rachel F. Quenemoen
Dorene L. Scott
Karen Evans Stout
Martha L. Thurlow, Director

National Center on Educational Outcomes
University of Minnesota • 350 Elliott Hall
75 East River Road • Minneapolis, MN 55455
Phone 612/626-1530 • Fax 612/624-0879
http://www.nceo.info

The University of Minnesota is committed to the policy that all persons shall have equal access to its programs, facilities, and employment without regard to race, color, creed, religion, national origin, sex, age, marital status, disability, public assistance status, veteran status, or sexual orientation.

This document is available in alternative formats upon request.

Acknowledgment

Dr. Sandra Thompson was the inspiration for this manual, through her conceptual and research contributions to all universal design documents published over the past several years, and through her belief that every research project should have a practical outcome. Dr. Thompson was a Research Associate at NCEO until 2005 when she passed away after a courageous battle with cancer. Her presence continues to be felt in our universal design work and in all work that is dedicated to improved educational outcomes for students with disabilities.

Executive Summary

Universal design for assessments is an approach to educational assessment based on principles of accessibility for a wide variety of end users. Elements of universal design include inclusive test population; precisely defined constructs; accessible, non-biased items; tests that are amenable to accommodations; simple, clear and intuitive instructions and procedures; maximum readability and comprehensibility; and maximum legibility. The purpose of this guide is to provide states with strategies for designing tests from the very beginning, through conceptualization and item construction, field-testing, item reviews, statewide operationalization, and evaluation. The objective is to create tests that present an accurate measure of the knowledge and skills of the diverse population of students enrolled in today's public schools. This guide is accompanied by an online supplement, which can be accessed at www.nceo.info/UDmanual/.

Table of Contents

What is Universal Design?

Have you ever tried to get a stroller or cart into a building that did not have a ramp? Or open a door with you hands full? Or read something that has white print on a yellow background, or is printed too small to read without a magnifying glass, or has words from a different generation or culture? Have you ever listened to a speech given without a microphone? All of these challenges have one thing in common—accessibility (or lack thereof). If we can't enter, see, or hear a service, we really can not benefit from what it has to offer.

In an effort to increase accessibility to structures, architects have developed a term called "universal design." In simple terms universal design means design for everyone. The idea behind universal design is to consider access of structures from their initial development (rather than retrofitting later), so that they become accessible to all people—including those with disabilities. By considering all users, then designing structures to be accessible to all, we improve our chances of people benefiting from the services within.

In the educational assessment realm, universal design promotes standard test conditions that are accessible to today's diverse population of test takers. By making general education assessments accessible, we are more able to determine what students know and are able to do. Considerations for universally designed tests include:

- Intended constructs are measured
- Respect for the diversity of the assessment population
- Concise and readable text
- Clear format for test
- Clear visuals
- Changes allowed to format without changing meaning or difficulty

Each of these universal design elements will be described in more detail throughout this guide.

Universal design provides students taking the general education assessment a fair opportunity to learn, and an accurate way to measure what they have learned. The goal of universally designed assessments is to provide the most valid assessment possible for the greatest number of students, including students with disabilities. Considering universal design is important from the beginning of assessment development, and important for the continual refinement and improvement of operational assessments. It has been common in the past to build assessments that are accessible for the majority of students with little thought given to students with disabilities and English language learners before the assessment went to print. An assessment can be much more universal—accessible to a wide range of students—if all students are taken into account at the *beginning* stages of test design.

Universally designed assessments are meant to increase access, but they *do not* change the standard of performance of assessments. Such tests are not intended to water down or make tests easier for some groups. Universally designed assessments are also not meant to replace accommodations or meant to replace the need for an alternate assessment for some students. Even by incorporating the elements for universal design in assessment design, accommodations may still be needed for some students in the areas of presentation, response, setting, timing/scheduling, technical equipment, or an alternate assessment. Universally designed assessments should, however, anticipate some of the common accommodations needed and the design of the general education test should allow for those accommodations to fit into the assessment process.

In addition to students with disabilities, *all* students who take such assessments benefit from having more accessible tests. Just like we all use sidewalk curb cuts for uses that the original laws might not have considered, so all students can benefit from a test being more accessible. Universal design makes tests better for everyone.

Laws

The No Child Left Behind Act of 2001 (NCLB) and other recent changes in federal legislation have placed greater emphasis on accountability in large-scale testing. Included in this emphasis are regulations that require assessments to be as accessible as possible. States are accountable for the success of all students, and tests should be designed in a way that provides all students an opportunity to demonstrate success.

With the reauthorization of the Individuals with Disabilities Education Act in 2004 (IDEA 2004), states are required for the first time to incorporate universal design principles in developing and administering tests, to the extent feasible. The Assistive Technology Act of 2004 defines universal design as “a concept or philosophy for designing and delivering products and services that are usable by people with the widest possible range of functional capabilities, which include products and services that are directly accessible (without requiring assistive technologies) and products and services that are interoperable with assistive technologies.” Because large-scale assessments have such high stakes, it is important to ensure that assessments meet the accessibility requirements outlined in NCLB and IDEA 2004.

Purpose of Guide

The purpose of this guide is to provide states strategies for designing tests from the very beginning, through conceptualization and item construction, field testing, item reviews, statewide operationalization, and evaluation. The end goal is to create tests that present an accurate measure of the knowledge and skills of the diverse population of students enrolled in today's public schools.

Ensuring universal design is not as simple as finding a set of criteria from which to judge test accessibility. Rather, the universal design of assessments is an iterative and on-going process. To this end, multiple on-going steps are necessary to ensure that assessments are as valid, reliable, and accessible as they can be.

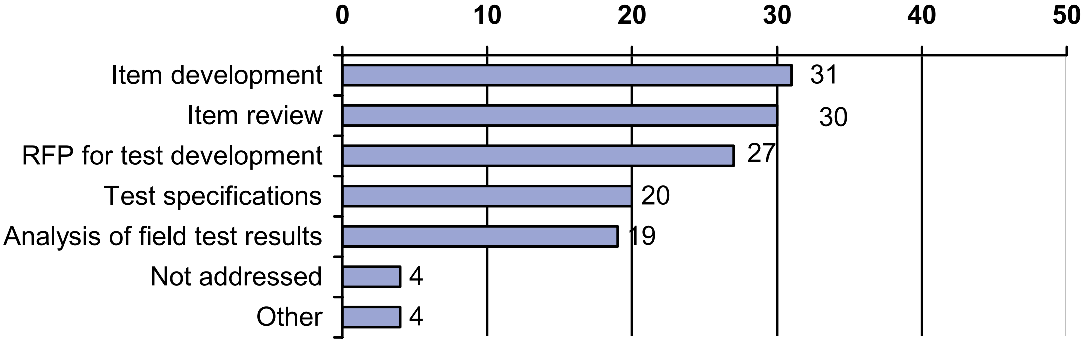
This guide outlines steps that states can use to ensure universal design of assessments from the beginning. The recommendations can be used for both computer- and paper-based assessments. We recommend following these steps in chronological order. Including any of these steps in the design and review of tests, however, may improve the design features of a state assessment. Readers are encouraged to view the online accompaniment to this report, which can be used in conjunction with this guide. See www.nceo.info/UDmanual/.

Research Supporting the Development of this Guide

Over the past several years, the National Center on Educational Outcomes (NCEO) conducted research in three areas that can help validate the universal design of test items. These include a list of considerations to use when constructing and reviewing items, using a "think aloud" process to check student understanding and interpretation of items, and data analysis to make sure that items are comparable across different populations. Details of these research methods can be found in three new NCEO reports (see Johnstone, Thompson, Moen, Bolt, & Kato, 2005; Thompson, Johnstone, Anderson, & Miller, 2005; and Johnstone, Bottsford-Miller & Thompson, 2006).

Including universal design in test construction is already taking place in the majority of states. During the 2004–2005 school year, 43 regular states (86%) and two unique states addressed issues of universal design (Thompson, Johnstone, Thurlow, & Altman, 2005). More than half of the states are addressing universal design at the item development level (n=31, 62%), item review level (n=30, 60%), and by including it in RFPs for test development (n=27, 54%) (see Figure 1).

Figure 1. Number of States Addressing Universal Design in 2005



Because universal design means different things to different people, we have attempted to clarify different ways of addressing issues in a stepwise fashion. This manual is not meant to be the final word on universal design. In fact, we are continuing our research at the time of this writing and we expect states are doing the same. Rather, it provides a systematic approach to examining assessments to ensure accessibility.

Step 1: Ensure the Presence of Universal Design in Assessment RFPs

Many states seek the assistance of external contractors to undertake work related to test development. Typically, such support is solicited in the form of a request for proposal (RFP) issued to suitable agencies or individuals. Test contractors must then find ways to address the language of the RFP in order to be competitive.

RFPs may be broadcasted to test contractors, curriculum and instruction consultants, independent researchers, or other agencies depending on the nature of the requirement. Applicants, often referred to as "bidders," are required to include the full range of students in the definition of the target population that will take the assessments.

Elements of universal design addressed in RFPs focus on the tests themselves. In addition to test design, RFPs generally also include such areas as qualifications of the bidders, reporting requirements, and payment schedules. Universal design does not require states to substantially alter many of the technical issues normally addressed in assessment-related RFPs (e.g., year-to-year equating of tests administered at certain grade-levels, establishing performance standards based on specific standard-setting procedures, etc.).

All students must have the opportunity to demonstrate their achievement of the same content standards. Therefore, to satisfy this section of the RFP, bidders must design state tests that allow the maximum number of students possible—and students with diverse characteristics—to take the same assessments without threat to the validity and comparability of the scores.

To this end, bidders must demonstrate how they will develop "universally designed assessments." If tests are designed from the beginning to allow participation of the widest range of students, these assessments result in valid inferences about the performance of all students, including students with disabilities, students with limited English proficiency, and students with other special needs. Information for bidders is found throughout this guide, for each step of assessment development.

Step 2: Test Conceptualization and Construction

Before any item construction can take place, a test needs to be carefully conceptualized. Conceptualization of tests includes figuring out the number and types of items that will adequately measure each of a state's content standards in a particular subject area (e.g., English language arts, mathematics, science.). The content to be measured must be defined precisely and explicitly, minimizing the effects of irrelevant factors.

In addition, each item needs to be written with accessibility features, that is, items that respect the diversity of the assessment population, are sensitive to test taker characteristics and experiences (gender, age, ethnicity, socioeconomic status, region, disability, language). Avoid content that might unfairly advantage or disadvantage any student subgroup, and minimize the effects of extraneous factors (e.g., avoid unnecessary use of graphics that cannot be presented in braille, use font size and white space appropriate for clarity and focus, avoid unnecessary linguistic complexity when it is not being assessed). Item writers may also need a description of the diverse needs of the population of students tested within a particular state. Bidders must also provide for a full range of test performance to avoid ceiling or floor effects, and must develop an item pool of sufficient size to permit the elimination of items that are not found to be universally appropriate during the test tryout and item analysis.

Item design is a time consuming and challenging practice. Yet, when items are designed from the beginning with accessibility in mind, they may save time and effort later. Well-designed items often move past item review teams (discussed below) with ease.

Step 3: Review Teams

Once the assessment is designed and in a format suitable for previewing, it is important for states to let sensitivity review teams examine the assessment. Such review teams are common practice in states, and are often encouraged by test vendors. When creating bias and content review teams, it is important to involve members of major language groups and disability groups. Grade level experts, representatives of major cultural and disability groups, researchers and teaching professionals all make up an effective review team. States should develop a standard reviewing form and common process and should contract out to experts in the field.

Reviewers will need the following items to perform a careful and comprehensive review:

- Purpose of the test, and standards tested by each item
- Description of test takers (e.g., age, geographic region)
- Field test results by item and subgroup
- Test instructions
- Overall test and response formats
- Information on use of technology
- State accommodation policies

Reviewers should also have a chance to look at response formats presented on the actual test. Are they clear? Will they induce errors not related to the questions being asked? If so, response formats should be adjusted accordingly. The description of what each item is testing should also be given to item reviewers so that they can assess if the construct intended to be tested is actually the one being tested. Also, design elements of the item that might affect the performance of a subgroup of students on that item may be flagged.

Bias and design issues often arise in test development and are not problematic if caught by review teams. Reviewers should look to flag items that may cause certain subgroups to have a disadvantage (or advantage) not related to common educational experiences. An efficient way to “flag” items is to use a review sheet, which provides reviewers an opportunity to mark items with potential issues, thus providing fodder for further discussion among reviewers. By using a structured form, reviewers are more likely to provide specific feedback to test vendors. Such feedback allows for items to be re-examined for design issues, rather than, as is often the case, summarily rejected for unclear reasons. When using structured forms, item reviewers then create a “win-win” situation for advocates and vendors. When reviewers use structured forms, they are able to provide test vendors specific information about items that may have an issue and the extent to which the issue is a problem. Vendors can then determine if changes can be made to items without having to remove items from banks entirely. The forms below are samples that can be used for item reviews. Forms are for both item-specific and whole test reviews.

Whether vendors or states run sensitivity review panels, it is the responsibility of the host organization to ensure that item reviews are conducted responsibly. Some tips to remember when reviewing items for universal design issues are:

1) Universally designed assessments *do not* change the standard of performance—items are not watered down or made easier for some groups. Universally designed assessments make the general education assessment more accessible to *all* students so that there is a better measure of a student's actual skills to make sure the test is measuring what it is supposed to measure.

2) Universally designed assessments are not meant to replace accommodations or alternate assessments. Even by incorporating the elements of universal design in assessment design, accommodations may still be needed for some students in the areas of presentation, response, setting, timing, and scheduling. Furthermore, students with the most significant cognitive disabilities may be eligible to take their state's alternate assessment. Universally designed assessments should, however, anticipate some of the common accommodations needed and the design of the test should allow for those accommodations to fit into the assessment process for which the general education assessment is most appropriate.

3) Some "considerations" were presented earlier in Figure 1. Considerations are simply ideas that should be considered when developing an assessment. They should be discussed openly, weighing the pros and cons of different design elements, and decisions should be made. As the administration of tests changes (e.g., more computer-based testing is used), the universal design considerations are likely to evolve.

4) In addition to English language learners, *all* students benefit from having more accessible general education tests. Just like we all use sidewalk curb cuts for uses that the original laws might not have considered, so all students can benefit from a test being more accessible.

Considerations for Universally Designed Assessment Items
Subject _____ Grade _____ Test form _____ Item #s on this page _1-5_Initials _____

Star (*) areas of strength and Check (√) areas of concern for each item	Pass-age	Item #1	Item #2	Item #3	Item #4	Item #5	Describe Concerns and Suggestions for items and reading passages (include item # with comment)	Recommend review by expert or student in Content Area, Specific Disability, Language, Culture	
Item measures its intended constructs									
Item respects the diversity of the assessment population • Sensitive to test taker characteristics and experiences (gender, age, ethnicity, socioeconomic status, region, disability, language) • Avoids content that might unfairly advantage or disadvantage any student subgroup • Other								Expert review?	Student review?
Item has concise and readable text • Commonly used words (except vocabulary tested) • Vocabulary appropriate for grade level • Minimum use of unnecessary words • Technical terms and abbreviations avoided unless tested • Sentence complexity appropriate for grade level • Question to be answered identifiable • Other								Expert review?	Student review?
Item has a clear format for text • Standard typeface • Twelve (12) point minimum size for all print, • High contrast between text and background • Sufficient blank space • Staggered right margins • Other								Expert review?	Student review?

Star (*) areas of strength and Check (√) areas of concern for each item	Pass-age	Item #1	Item #2	Item #3	Item #4	Item #5	Describe Concerns and Suggestions for items and reading passages (include item # with comment)	Recommend review by expert or student in Content Area, Specific Disability, Language, Culture	
Item has clear visuals (use NA for none) • Visuals are needed to answer the question • Visuals have clearly defined features • High contrast between visuals and background • Visuals are clearly labeled • Other								Expert review?	Student review?
Item allows changes to format without changing meaning or difficulty (check allowed accommodations) • Braille or other tactile format • Sign language interpretation • Oral presentation • Assistive technology • Translation into another language • Other								Expert review?	Student review?
Describe other considerations specific to item on back									

Considerations for Universally Designed Whole Tests
Subject _____ Grade _____ Test form ____ Reviewer Initials _____

	Star (*) areas of strength and Check (√) areas of concern for this test	**Describe Concerns and Suggestions for Improvement.**
This test measures what it intends to measure • Reflects the intended content standards (reviewers have information about the content being measured) • Minimizes knowledge and skills required beyond what is intended for measurement • Other		
Response format for extended response items • Number of points for extended response items is clear • Correct or possible responses are listed • Same amount of credit for written or numerical response (e.g., “explain or show work,” “use words or symbols to describe”)—leaves option of less writing for students who are not skilled writers but can “do the math” • Other		
Response format for multiple choice items • Division between items on response form is clear (change of color or shading) • Response bubbles are sufficiently large • Does test require separate response form (e.g., gr. 8) or do students write directly in test booklet (e.g., gr. 3) For grade 3, student write anywhere on page—besides just in circle • Other		

	Star (*) areas of strength and Check (√) areas of concern for this test	Describe Concerns and Suggestions for Improvement.
Overall comparison of types of items • Number/percent of strong items vs. number/percent with concerns • Number/percent of items with visuals • Number/percent of multiple choice vs. extended response • Number/percent of items with other concerns (e.g., reading passage addressing urban vs. rural settings and other cultures)		
Passages • Appropriate for grade level • Cognitive demands of all passages are balanced (all are not too easy or too difficult) • Visuals related to passages are clear • Format of passages is clear		
Other considerations for this test		

Step 4: Using Think Aloud Methods to Analyze Items That Were and Were Not Flagged During Item Reviews

In Step 3, experts flagged items that they felt may have design issues that could be problematic for students. In an effort to validate the findings of experts, a series of items can be examined by students themselves using cognitive lab, or think-aloud methods.

Think aloud methods were first used in the 1940s and have since been used for a variety of "end user" studies in the fields of ergonomics, psychology, and technology. In the case of statewide assessments, the end users are students who will take tests. Think aloud methods tap into the short-term memory of students who complete assessment items while they verbalize. The utterances produced by students then become the data set.

Researchers believe that the verbalizations produced in think aloud studies provide excellent information because they are not yet in the long-term memory. Once experiences enter our long-term memory, they may be tainted by personal interpretations. Therefore, an excellent way of determining if design issues really do exist for students is to have students try out items themselves.

For example, if experts felt that a particular item was biased against a certain population, an excellent way to validate (or disprove) these suspicions is to have students complete the item while thinking aloud. If a student verbalizes everything he or she is thinking while completing the item, it will be easy to see how the design of the item affects the student's understanding of the item. If the student does have difficulty with the item, it will also be easy to determine if the difficulty is a result of design features or a lack of curricular knowledge.

NCEO typically videotapes all think aloud activities, but states can also either audiotape or have several observers analyze field test notes (inter-rater agreement is important for making decisions based on think aloud activities). In addition, NCEO generally selects students that achieve at both high and low levels on statewide achievement tests. To this end, a sample population might include students without disabilities and students who are from majority cultures as well as students with disabilities, English language learners, and students from low socio-economic status. The process for selecting and conducting think aloud studies is described below, in Vignette 1.

Vignette 1

State X has recently conducted an expert review on its fourth grade mathematics test. Reviewers found that most items only had minor formatting issues that they would like to see improved, but that three of the items had major issues pertaining to bias, presentation, and comprehensible language. State X's assessment director was concerned that these items might cause students with a variety of descriptors to incorrectly answer these items because of design issues, thus reducing the validity of inferences that could be drawn from the test. State X then decided to conduct a think aloud study on the three items in question, as well as three items that generally met the approval of item reviewers.

Overview of Activities

State X opted to conduct the think aloud study with its own staff (alternatively, they may have decided to offer a subcontract to a local university or research organization to conduct the study). The study took place in a quiet room, where State X staff members could videotape the procedures.

Sample

Because State X's assessment director was concerned about the effects of bias, presentation, and language on students with particular disabilities and English language learners, she targeted these students, as well as students who were deemed "typically achieving, non-disabled, English proficient" students. In total, 50 Grade 4 students were contacted. Among these were: 10 students with learning disabilities, 10 students with mild mental retardation (who took the general education assessment), 10 students who were deaf, 10 students who were English language learners (but did not have a disability), and 10 non-disabled, English proficient students.

Procedures

Each student was then individually brought to the quiet room. First, State X staff members explained the process. Then, students practiced "thinking aloud" by describing everything they do when they tie their shoes (sign language interpreters were present for students who are deaf). Once students understood the process, they were asked to think aloud while they answered mathematics items. The only time State X staff spoke was when students were silent for more than 10 seconds, at which time staff encouraged students to “keep talking.” Each item took approximately 10 minutes per student.

After students completed items, State X staff asked post-hoc questions, simply to clarify any issues they did not understand. Data derived from post-hoc questions are not as authentic as think aloud data, but they can help to clarify issues that were unclear to staff.

Analysis

Once all think aloud activities were completed, State X staff reviewed all the videotapes they had taken. Using NCEO’s think aloud coding sheet (found below), staff were easily able to determine if design issues were problematic for particular populations. The data they collected helped them to make recommendations for Step 5.

Universally Designed Assessments – "Think Aloud" Study

Student ID # ___

Grade: 4 8

Researcher Initials ___

Describe Item: ____

Item # ___

Describe Researcher Introduction (if included on video)

Prompts/Assistance

___ No Prompts

___ Other

___ Researcher

___ Teacher

___ Interpreter

___ Point To Item

___ Paraphrase Directions

___ Paraphrase Item

Describe Interaction with Student

Directions at Top of Page

___ NA

___ Other

___ Student Read Aloud

___ Student Read Silently

___ Student Skipped

___ Researcher Read Aloud

___ Signed by Interpreter

___ Reader / Signer Skipped

Describe Reading / Skipping Directions

Item Reading

___ Other

___ Student Read Aloud

___ Student Read Silently

___ Student Skipped

___ Researcher Read Aloud

___ Signed by Interpreter

Describe Reading / Skipping

Reading Fluency

___ NA (Student did not read item aloud)

___ Student read all words correctly

___ Student mispronounced some words

___ Student had difficulty with many words

List words mispronounced

Researcher asked follow-up questions ___ Yes ___ No

Describe follow-up questions and student responses

Problem Solving

___ Incorrect problem solving process

___ Not Apparent

___ Correct process for solving problem

___ Appeared to guess

___ Did not attempt to solve problem

Describe problem solving process

Student was distracted by something on the page ___ Yes ___ No ___ Not Apparent

Describe distraction

**Add observer comment on back (O.C.) and note students to use as examples*

Step 5: Revisit Items Based on Information from Steps 3 and 4

Steps 3 and 4 are likely to produce a rich data set about concerns about particular items or the entire test. Prior to field testing (Step 6) it is important to analyze the data produced in Steps 3 and 4 and make any possible changes that can be made to the test. Some changes may be impossible prior to field testing while others (such as formatting changes) may be quite easy to make. Regardless of whether or not changes are made to tests, data from Steps 3 and 4 are important sources for recommendations and cross-analyzing with field test results.

Step 6: Test Tryout, Analysis, and Revision

It is common practice for states to field test potential exam items well in advance of their actual inclusion in statewide testing systems. Somewhat less common is taking potential exam items and transferring them into accommodated formats and then field testing them for potential differential item functioning. It is important that test administrators are aware of the effects of all or many accommodations on each test item when making decisions on which items to include on exams.

Twenty states (40 percent) currently field test all or some of their potential test items in accommodated formats including braille, large print, audio tape, computer, and oral presentation (Thompson, Johnstone, Thurlow, & Altman, 2005). Many states are currently working out the technical issues of such processes including item selection, accommodation selection, subject selection, small participation numbers, numerous disabilities, and multiple grade levels. Others are conducting tests on a limited set of accommodations at the current time. It is likely that this process will become much more widespread in the next few years. It is critical to include a full range of students in the tryout sample (e.g., students with disabilities, students with limited English proficiency, and other students with special needs). Because there may be constraints in sampling due to the low numbers of students with specific characteristics, states may need to identify over-sampling strategies, (e.g., select groups of items for which additional sampling will occur) or include the use of accommodations during the test tryout.

Field testing using various approaches produce a large data set from which states and contractors can re-examine item qualities. How such data can be used to make decisions is outlined in the next step.

Step 7: Item Analyses

A useful method for ensuring Universal Design of assessments is to conduct large-scale statistical analyses on test item results. This section will describe four categories of statistical techniques currently used in field practice by researchers. The statistical techniques described can help determine items that affect students in different ways, even after researchers factor student latent traits into equations.

Many methods exist for examining assessment data for possible bias or other issues. Several methods, from simple methods based on classical test theory to those with increasing complexity based on more contemporary item response theories (IRT) may be quite valuable. Four categories of analysis recommended by NCEO are: Item Ranking, Item Total Correlation, Differential Item Functioning (DIF) using Contingency Tables, and DIF using Item Response Theory (IRT) approaches.

Item ranking is a procedure that requires the data analyst to compare item ranks from different groups to determine if certain items are more challenging (and potentially biased) toward particular students. Item ranking assumes that every item has a particular degree of difficulty. Statisticians express this level of difficulty with a P (probability) statistic. For example, if 60% of students in a population answered an item correctly, its P TOTAL would be .60. One can also calculate P figures for groups and then rank items from most to least difficult for the total population and for particular subgroups.

A second method for flagging items for possible universal design issues is Item Total Correlation (ITC). ITC analysis examines how items correlate to other items on the same test. Therefore, ITC analysis is a *within* group investigation, that is, ITC determines how well an item's P-value correlates with other items in the test for a particular group. If an item does not correlate with the rest of the test, it may be problematic. A second set of tests will determine if there are statistically significant differences between ITCs for target and comparison groups.

Analysis of Differential Item Functioning (DIF) seeks to determine if a particular item is substantially more difficult for one group than another *after* taking into account the overall differences in knowledge of the subject tested. Analyses are predicated on the notion that items should be of similar difficulty level for students of equal achievement levels across target and reference groups. DIF occurs when one item is substantially more difficult for a particular group after the matching of subjects for achievement levels. One can calculate DIF statistics by computing the proportion of students who answer an item correctly (within a given overall test score range) in target and reference groups. Statistically significant findings may point to an item's problematic nature.

Differential Item Functioning can also be determined using Item Response Theory (IRT) approaches. IRT analyses focus on individual items, but present statistics that couch item difficulty conclusions on the "latent traits" of individual test takers. In theory, all test takers have latent traits in terms of ability. IRT applications require the quantification of these traits. Rather than calculating DIF using results from individual items with assumed equality of achievement, IRT applications allow researchers to investigate item difficulty based on the individual traits of test takers. In other words, researchers weigh items for difficulty in balance with the latent traits of individuals.

The analyses listed above will almost certainly produce disparate results because they are examining slightly different item functions. In fact, between disability groups and analyses, it is likely that most items on a test will be "flagged" at least once. Such a result does not necessarily mean that an entire test is flawed. It is important, however, to understand which items may have universal design issues and for what group of students.

The overwhelming amount of data produced by these analyses will make it difficult to determine which items may have universal design issues. It is impractical to re-examine every item on the test. Therefore, a reasoned approach to sorting through large amounts of data is the rule of halves. If an item is flagged by half of the analyses, that item may be a candidate for re-examination. Other data (expert reviews and think aloud data) may then be revisited for items that are particularly statistically problematic.

By using a series of statistical analyses, state assessment personnel can determine if items function differentially for students with disabilities or English language learners. From there, items can be improved and re-inserted into large-scale tests. Statistical tests, as part of an iterative and on-going design process, will confirm or refute that changes to items have made them more accessible.

Step 8: Final Revision

After experts have reviewed items, students have approached items using think aloud methods, and field test results have been reviewed, states and contractors can discuss the final revisions that need to be made to tests. It is possible that no changes at all will be made. On the other hand, the “final revision” stage is the last time states and contractors can address design issues before tests are distributed with high stakes. This stage is one that should be approached with caution, but in a cooperative spirit that makes sense for all students as well as the needs of the state’s finances and timelines.

Step 9: Testing

This step is the culmination of months (or years) of hard work on the part of both the state and the contractor. During testing periods in states, students take the assessments designed by contractors under standard and accommodated conditions. Results are used for accountability purposes, and are monitored at both the school and district levels. Designing a test for accessibility is a challenging process, and culminates when students take the “live” test.

Step 10: Post-Test Review

Once tests results are available, the process starts again. States can examine results statistically, and begin the expert review and think aloud processes for the following year's test. When contractors develop a test that states deem acceptable for use for more than one year, the universal design process is streamlined, as many of the potential problems with a test were caught during the design and field test stage. Universal design processes can then be used as ongoing item reviews.

It is possible that there will never be a test that is accessible to all students for all items. While a perfectly accessible test may not be possible, a more accessible assessment is always an option. Hard work, cooperation, and following the steps in this guide may help the process. In addition, states may develop their own universal design processes. As universal design research emerges, processes will become more succinct, efficient, and effective. We applaud the states that have made commitments to accessible assessments for all students, and hope our current and future processes will serve you well.

References

Johnstone, C. J., Bottsford-Miller, N. A., & Thompson, S. J. (2006). *Using the think aloud method (cognitive labs) to evaluate test design for students with disabilities and English language learners* (Technical Report 44). Minneapolis, MN: National Center on Educational Outcomes.

Johnstone, C. J., Thompson, S. J., Moen, R. E., Bolt, S., & Kato, K. (2005). *Analyzing results of large-scale assessments to ensure universal design* (Technical Report 41). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Thompson, S. J., Johnstone, C. J., Anderson, M. E., & Miller, N. A. (2005). *Considerations for the development and review of universally designed assessments* (Technical Report 42). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Thompson, S. J., Johnstone, C. J., Thurlow, M. L., & Altman, J. R. (2005). *2005 State special education outcomes: Steps forward in a decade of change*. Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.