NCEO POLICY DIRECTIONS

Number 18 September 2006

Using Systematic Item Selection Methods to Improve Universal Design of Assessments

Background

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 accessible. 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 their knowledge and skills. With the reauthorization of the Individuals with Disabilities Education Act in 2004, states are required for the first time to incorporate universal design principles in developing and administering tests, to the extent feasible.

Applying the concept of universal design to statewide assessment means that assessments are designed from the beginning and continually refined to result in more valid inferences about performance of students with diverse characteristics. These assessments are based on the premise that each child in school is a part of the population to be tested, and that test results should not be affected by disability, gender, race, or English language ability. While universally designed assessments are not intended to eliminate individualization, they may reduce the need for accommodations by eliminating access barriers associated with the tests themselves. At the same time, the intent of the measurement—the intended content and construct of the assessment—are not changed.

Including universal design in test construction is already taking place in the majority of states. According to a survey of states conducted by the National Center on Educational Outcomes (NCEO), during the 2004–2005 school year 43 states addressed issues of universal design. More than half of the states addressed universal design at the item development and review levels, and by including it in RFPs for test development.

There are many elements involved in creating universally designed assessments. They include making sure that students with disabilities are part of field testing, for example. But a major focus of universal design in assessments is making sure that the items included in the assessment are appropriate. There are several methods for selecting items to ensure that they optimize the characteristics of universal design. The purpose of this *Policy Directions* is to provide an overview of these item selection methods, and to suggest that a combination of the methods

NATIONAL CENTER ON EDUCATIONAL OUTCOMES

The College of Education & Human Development

UNIVERSITY OF MINNESOTA

Table 1. Item Analysis Methods: Pros, Cons, and Current Practice

Method	Strengths	Weaknesses	Current Practice	Improvement on Current Practice
Expert Review	Provides structured review of items by experts	Does not provide actual student data	Unstructured "sensitivity" review panels	Provides reviewers with tools to make decisions
Statistical Analyses	Provides significance data on field tested items; flags potentially problematic items	Validity questionable for small populations	Only DIF calculations performed	Multiple analyses conducted, providing patterns of flagged items
Think Aloud Methods	Provides information on why particular items function as they do	Does not provide data across groups	Not currently in widespread use	Provides important design information

will produce the better result. Each method has strengths and weaknesses, may lead to different results, and is in different stages of current practice (see Table 1). Although each method has merits, NCEO recommends states employ all methods systematically and in conjunction with each other.

Expert Review

Once an assessment is designed and in a format suitable for previewing, it is important for states to have sensitivity review teams examine the assessment. The use of review teams to examine items is common practice in many states, and is generally encouraged by test vendors. When creating bias and content review teams, it is important to involve members familiar with disability and language issues. Grade level experts, representatives of major cultural and disability groups—or those who can reflect their needs—researchers, and teaching professionals all make up an effective review team.

NCEO, working closely with experts in the fields of assessment, disability, reading, mathematics, and language acquisition, developed and refined a set of considerations for test developers and item reviewers to use to ensure that tests are universally designed. The considerations are listed in Table 2 (see *Technical Report 42* listed in Resources). They ask six important questions about test items. Test item developers and reviewers can use these questions to help determine the extent to which an item provides appropriate accessibility without changing the intended construct.

There are considerations that item and test developers or reviewers can use to help determine whether a test overall—not just the individual items—is universally designed:

- Do all visuals (e.g., images, pictures) and text provide information necessary to respond to the item?
- Is information organized in a manner consistent with an academic English framework with a left-right, top-bottom flow?
- Can booklets/materials be easily handled with limited motor coordination?
- Are response formats easily matched to question?
- Is there a place for taking notes (on the screen for computer-based tests) or extra white space with paper pencil?

These considerations show that it is not just each item that is important, but also how the whole test is put together that is also an important aspect of universal design.

As states and other testers explore the use of computers for testing, it is important to have ways to judge their appropriateness and universal design features as well. There are additional considerations for computer-based tests to ensure that they are universally designed (see Table 3).

When using Expert Review considerations, NCEO recommends incorporating the following into the review:

- Conduct the review as early as possible in the stages of test development.

Table 2. Considerations for Universally Designed Assessment Items

Does the item...
Measure what it intends to measure? • Reflect the intended content standards (reviewers have information about the content being measured)? • Minimize knowledge and skills required beyond what is intended for measurement?
Respect the diversity of the assessment population? • Sensitive to test taker characteristics and experiences (consider gender, age, ethnicity, socio-economic level, region, disability, and language)? • Avoid content that might unfairly advantage or disadvantage any student subgroup?
Have clear format for text? • Standard typeface? • Twelve (12) point minimum size for all print, including captions, footnotes, and graphs (type size appropriate for age group), and adaptable font size for computers? • High contrast between color of text and background? • Sufficient blank space (leading) between lines of text? • Staggered right margins (no right justification)?
Have clear visuals (when essential to item)? • Visuals are needed to answer the question? • Visuals have clearly defined features (minimum use of gray scale and shading)? • Sufficient contrast between colors? • Color alone is not relied on to convey important information or distinctions? • Visuals are labeled?
Have concise and readable text? • Commonly used words (except vocabulary being tested)? • Vocabulary appropriate for grade level? • Minimum use of unnecessary words? • Idioms avoided unless idiomatic speech is being measured? • Technical terms and abbreviations avoided (or defined) if not related to the content being measured? • Sentence complexity is appropriate for grade level? • Question to be answered is clearly identifiable?
Allow changes to its format without changing its meaning or difficulty (including visual or memory load)? • Allows for the use of braille or other tactile format? • Allows for signing to a student? • Allows for the use of oral presentation to a student? • Allows for the use of assistive technology? • Allows for translation into another language?

- Include disability, technology, and language acquisition experts in item reviews.
- Provide professional development for item developers and reviewers on use of the universal design considerations.
- Present the items in the format in which they will appear on the test.
- Include standards being tested with the items being reviewed.
- Try out items with students (use Think Aloud methods).
- Field test items in accommodated formats.
- Review computer-based items on computers.

Statistical Analysis

A quantitative approach to selecting items that appear to provide access to students with certain characteristics, such as disabilities, is to conduct statistical analyses on test item results. Many statistical methods exist, ranging from simple methods based on

Table 3. Considerations for Universally Designed Computer-based Tests

Layout and design

- Sufficient contrast between background and text and graphics for easy readability.
- Color alone is not relied on to convey important information or distinctions.
- Font size and color scheme can be easily modified (through browser settings, style sheets, or on-screen options).
- Stimulus and response options are viewable on one screen when possible.
- Page layout is consistent throughout the test.
- Computer interfaces follow Section 508 guidelines (www.section508.gov).

Navigation

- Students have received adequate training on use of test delivery system.
- Navigation is clear and intuitive; it makes sense and is easy to figure out.
- Navigation and response selection is possible by mouse click or keyboard.
- Option to return to items and return to place in test after breaks.

Screen reader considerations

- Item is intelligible when read by a text/screen reader.
- Links make sense when read out of visual context (“go to the next question” rather than “click here”).
- Non-text elements have a text equivalent or description.
- Tables are only used to contain data, and make sense when read by screen reader.

Test specific options

- Access to other functions is restricted (e.g., e-mail, Internet, instant messaging).
- Pop up translations and definitions of key words/phrases are available if appropriate to the test.
- Students writing online can get feedback on length of writing on-demand in cases where there is a restriction on number of words.
- Students are able to record their responses and read them back as an alternative to a human scribe.
- Students are allowed to create persistent marks to the extent that they are already allowed on paper-based booklets (e.g., marking items for review; eliminating multiple choice items, etc.).

Computer capabilities

- Adjustable volume.
- Speech recognition available (to convert user’s speech to text).
- Test is compatible with current screen reader software.
- Computer-based option to mask items or text (e.g., split screen).
- Computer software for test delivery is designed to be amenable to assistive technology.

classical test theory to complex methods based on contemporary item response theories (IRT). Four statistical approaches currently used in the field by researchers have practical usefulness for identifying items that potentially violate universal design principles for groups of students.

Table 4 shows the four statistical analysis methods that are useful for flagging items to identify those that are potentially more challenging for particular students. Each of these has been used to identify items that may not be universally designed (see *Technical Report 41* listed in the Resources). Statistical methods are based on various assumptions that determine when items should be flagged as producing different performance from what would be expected. The flagging of an item is taken as an indication that the item may not be as accessible as possible, and may be creating barriers that do not allow the student to demonstrate his or her knowledge and skills.

Statistical analyses are useful for understanding which items may be biased and need revision, but they do not provide information on why particular items function the way they do. Understanding why is often critical to knowing what to do in making revisions to items.

Think Aloud Methods

Think aloud methods provide a way to answer questions about why particular items may be problematic for students. For state assessments, think aloud methods tap into the short-term memory of students who complete assessment items while they verbalize. 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 long-term memory, they may be tainted by personal interpretations. Therefore, an excellent way to determine whether design issues really do exist for students is to have students try out items themselves.

Think aloud methods have been used in research to identify problematic items for students with disabilities (see *Technical Report 44* in Resources). When students verbalize everything they think while completing an item, it becomes easy to see how the design of the item affects the understanding of the item. If a student does have difficulty with the item, it will also be easy to determine whether the difficulty is a result of design features or a lack of curricular knowledge. The depth of understanding that results from think aloud techniques is this method's strength. Follow-up questions can supplement any unclear data derived from think aloud techniques.

▶ Recommendations

Attaining universal design in statewide assessments is a goal for states as they continually refine and improve their assessments. With the understanding that this goal means that states are retaining the constructs and content that their tests are intended to measure, the question becomes how to identify items that may violate the principles of universal design.

It is important to have a systematic approach to reach universal design in assessments. Research is paving the way to identifying techniques that are workable. Any one technique by itself, however, may be insufficient. The methods identified here will reduce the possibility of erroneously flagging and eliminating items that reflect poor performance due to students' lack of opportunity to learn.

Specifically, using sets of considerations for expert review can make the test development process more transparent, informed, and focused on the needs of the entire population of students and ensure that the assessment results are more meaningful for the widest range of students. Statistical analysis methods can help pin-point test items that are potentially problematic and that may have universal design issues. Think aloud methods can be used with students themselves who can provide information that will help illuminate whether there are design issues that need to be addressed. These are all aspects of striving to reach universal design, which holds the promise of improved student performance. This goal can be reached without

Table 4. Four Statistical Analysis Methods for Reviewing Test Item Results

Method	Procedure
Item Rankings	Items are ranked from most to least difficult for total population and for particular groups. It is assumed that ranks should be similar for groups and total.
Item Total Correlation (ITC)	Within-group investigation of how individual items correlate with the total score on the same test; poor correlations may signal a problem. Different ITCs for the same item across different groups of test takers may indicate that the item behaves differently across those groups.
Differential Item Functioning (DIF) – Contingency Table Methods	Performance on items is compared for large groups that perform at similar levels on the total test. In contingency table methods DIF statistics are calculated by comparing the proportion of students answering an item correctly in target and comparison groups with the same total score range; statistically significant differences may point to an item's problematic nature.
Differential Item Functioning (DIF) – Item Response Theory Approaches	In IRT, characteristics of each item are represented by an item response curve, which is a function of individual test takers' characteristics called "latent traits." Items are compared based on their item response curves between target and comparison groups.

compromising the validity of the assessment.

Resources

Using the Think Aloud Method (Cognitive Labs) to Evaluate Test Design for Students with Disabilities and English Language Learners (Technical Report 44). Johnstone, C.J., Bottsford-Miller, N.A., & Thompson, S.J. (2006). Available from the National Center on Educational Outcomes at http://education.umn.edu/nceo/OnlinePubs/Tech44/.

Analyzing Results of Large-Scale Assessments to Ensure Universal Design (Technical Report 41). Johnstone, C.J., Thompson, S.J., Moen, R.E., Bolt, S., & Kato, K. (2005). Available from the National Center on Educational Outcomes at http://education.umn.edu/nceo/OnlinePubs/Technical41.htm.

Consideration for the Development and Review of Universally Designed Assessments (Technical Report 42). Thompson, S.J., Johnstone, C.J., Anderson, M.E., & Miller, N.A. (2005). Available from the National Center on Educational Outcomes at http://education.umn.edu/nceo/OnlinePubs/Technical42.htm.

Universal Design Applied to Large-Scale Assessments (Synthesis Report 44). Thompson, S. J., Johnstone, C. J., & Thurlow, M. L. (2002). Available from the National Center on Educational Outcomes at http://education.umn.edu/nceo/OnlinePubs/Synthesis44.html.

Universally designed assessments: Better tests for everyone! (Policy Directions No. 14). Thompson, S., & Thurlow, M. (2002). Available from the National Center on Educational Outcomes at http://education.umn.edu/nceo/OnlinePubs/Policy14.htm. ▲

About NCEO

The National Center on Educational Outcomes (NCEO) was established in 1990 to provide national leadership in the identification of outcomes and indicators to monitor educational results for all students, including students with disabilities. NCEO addresses the participation of students with disabilities in national and state assessments, standards-setting efforts, and graduation requirements.

The Center represents a collaborative effort of the University of Minnesota, the Council of Chief State School Officers (CCSSO), and the National Association of State Directors of Special Education (NASDSE).

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.

NCEO Staff

NCEO *Policy Directions* is a series of reports that address national policy issues related to students with disabilities. This report was prepared by Christopher Johnstone, Martha Thurlow, Michael Moore, and Jason Altman. It is available in alternative formats upon request.

NCEO
350 Elliott Hall
75 East River Road
University of Minnesota
Minneapolis, MN 55455
Phone: 612-626-1530
Fax: 612-624-0879

Visit our Web site at:
http://www.nceo.info