INCLUDING STUDENTS WITH DISABILITIES IN LARGE-SCALE ASSESSMENT: EXECUTIVE SUMMARY

The Technical Work Group
on
Including Students With Disabilities in Large-Scale Assessment

April 2006

Prepared by
Behavioral Research and Teaching, University of Oregon
and
American Institutes for Research, Washington, D.C.

Introduction

Both the *No Child Left Behind Act* (*NCLB*) and the *Individuals with Disabilities Education Act* (*IDEA*) require states to provide students with disabilities access to the general education curriculum and to hold schools accountable for the academic achievement of all students. This executive summary highlights the core findings and recommendations of *Including Students With Disabilities in Large-scale Assessment* (Technical Work Group, 2006), a set of papers commissioned by the Office of Special Education Programs of the U.S. Department of Education. These papers are written for educators who are responsible for administering large-scale assessment and accountability systems and address several topics related to the inclusion of students with disabilities in large-scale assessments. The first paper, titled "Validating Assessments for Students with Disabilities," discusses different types of assessment approaches that can be used to validly assess students with disabilities. The second paper, titled "Reliability Issues and Evidence," focuses on the reliability of assessments and the evidence needed to establish reliability. The third paper, titled "Validity Evidence," focuses on documenting assessment validity evidence. The fourth paper, titled "Standards and Assessment Approaches for Students with Disabilities Using a Validity Argument," illustrates the validation process using actual state standards and assessments. The fifth paper, titled "A Decision Framework for IEP Teams Related to Methods for Individual Student Participation in State Accountability Assessments," describes a systematic framework for IEP teams to determine the most suitable way for students with disabilities to participate in the annual statewide

assessments. The final paper, titled "Professional Development on Assessment Systems," discusses the need for on-going professional development for educators.

There are four assessment options available for the participation of students with disabilities in large-scale assessments that are used to judge academic achievement in schools and districts: test accommodations, alternate assessments, and modified and alternate achievement standards[1]. The papers identify the critical elements of an assessment system that requires careful stewardship to maintain validity when students with disabilities are fully included in the system. These papers (1) present a **model** for statewide assessment systems that encompass the four options, and (2) provide **criteria** for states to use in ascertaining the technical quality of their state assessment systems.

States differ in the content standards they have adopted and the assessments they use to measure proficiency. Therefore each state must approach student participation in a manner that is consistent with its standards and assessments. For students with disabilities who cannot participate meaningfully in general education assessments, states must provide both appropriate accommodations and alternate assessments as part of the statewide approach to assessment. The IEP team must determine how a student with a disability can meaningfully participate in the statewide assessment (e.g., whether the student needs testing accommodations or should take an alternate assessment). The outcome from this participation can be used to meet *NCLB*'s accountability requirement, that states report annually the academic achievement of all students in their schools and districts. This entire accountability process is based on grade-level academic content standards, assessments aligned to the standards, and performance judged against academic achievement standards (either those developed for the general education assessment or those developed as part of the alternate assessment). Adequate yearly progress (AYP) is then based on these assessment results.

Regulations published in the *Federal Register* (Dec. 9, 2003) announced options for evaluating proficiency of students with the most significant cognitive disabilities based on **alternate**

[1] Achievement (also known as performance) standards describe "how good is good enough." According to the *Standards and Assessments Peer Review Guidance* (U.S. Department of Education, 2004), ". . . [they] include at least two levels of achievement (proficient and advanced) that reflect mastery." Most states have three or more performance levels that represent "proficient," "below proficient," and "above proficient."

achievement standards, where proficient scores can be used in determining AYP (subject to a one percent cap). On Dec. 15, 2005, the U.S. Department of Education published a Notice of Proposed Rulemaking (NPRM) in the *Federal Register* that would allow states to develop modified achievement standards and use assessments aligned with those modified standards for a group of students with disabilities who can make progress toward, but may not reach, grade-level achievement standards in the same timeframe as other students.[2] Regardless of which achievement standards are used to evaluate performance (modified or alternate), they must be aligned with a state's grade-level content standards.

Testing Methods Used in Statewide Assessments

Methods of Assessing Academic Achievement

States currently can use any of four testing methods to measure the achievement of students with disabilities for the purpose of determining whether they and their schools and local education agencies (LEAs) have made AYP. Three of the four testing methods — regular assessment, regular assessment with accommodations, and alternate assessment judged against grade-level achievement standards — entail judging achievement test scores against the grade-level achievement standards in place for all students. The other testing method — alternate assessment judged against **alternate** achievement standards — allows states to judge performance against different achievement standards. In addition, a recently proposed rule by the U.S. Department of Education allowing states to develop **modified** achievement standards would provide a fifth testing method to assess the academic achievement of students with disabilities. Both of these latter two methods are optional; states are not required to develop either modified or alternate achievement standards. Regardless of the testing method, however, all achievement standards must be either aligned with or linked to (in the case of alternate achievement standards) the grade-level content standards that are in place for all students. Therefore, adaptations to the regular large-scale assessment must be carefully planned and be appropriate for students with disabilities, with a rationale provided for any changes that could alter the interpretations of proficiency of grade-level content. This entire process of adaptation becomes part of the validation framework of a statewide assessment system.

[2] Retrieved from the World Wide Web on Feb. 8, 2006 at http://www.ed.gov/legislation/FedRegister/proprule/2005-4/121505a.html

Two Types of Adaptations

Two types of adaptations to the statewide assessment can be used to create alternate assessments aligned with grade-level content standards: (1) modifying the types of supports used when the assessment is given or taken; and/or (2) limiting the breadth or depth of the assessment "content" (i.e., the standards, objectives, skills and tasks covered by the assessment). "Supports" refers to the types of materials, techniques, scaffolds, prompts, and assistive technologies used in the administration of the assessment. "Breadth" refers to the number of standards being addressed in the assessment; "depth" of standards refers to the number of objectives as well as the requisite skills and range of exemplary tasks considered appropriate for the standards and objectives. For some students with disabilities, the regular assessment is appropriate but accommodations need to be made in the manner in which the test is given or taken, in which case various supports are used as an accommodation (e.g., Braille, large print, reading math problems, separate settings, etc.). For some students, however, these adaptations are insufficient and an alternate assessment is needed, in which case there are three types of achievement standards that can be used to judge proficiency. If grade- level achievement standards are being used to judge proficiency, then the adaptations are only in the types of supports being provided, with no change in the breadth or depth of the assessment content. These adaptations are likely to exceed those allowed as accommodations and therefore performance needs to be part of an alternate assessment option as judged against grade-level achievement standards. Adaptations comprising alternate assessments based on alternate achievement standards and assessments based on modified achievement standards imply a reduction in the breadth and/or depth of the achievement standards being assessed.

Seven Principles for Developing Test Questions and Tasks

To guide states in developing these assessments, seven principles are presented for developing test questions and tasks that are based on grade-level content standards, whether they are assessments judged against **modified** or **alternate** achievement standards. These principles, explained in these papers in detail with descriptions and examples, are: (1) derive test content based on grade-level content that is grade specific; (2) parallel the breadth and depth of grade-level curricula; (3) include items and performance tasks that sample multiple levels of skill and knowledge complexity; (4) reflect a developmental progression of skills that provides a fair and appropriate representation of the content standard; (5) show progressive levels of achievement across grade levels; (6) reflect universal design and thereby reduce bias

while ensuring student access to content; and (7) represent the student's own work even when partial credit is given.

These principles ensure testing methods that are based on state grade-level academic content standards and preclude the development or administering of tests that are below grade level. The five testing methods and their defining characteristics are displayed in Table 1. Performance on the first three types of tests is judged against the **same** grade-level achievement standards adopted for all students. Performance on the last two types of tests is judged against different achievement standards. These latter two testing options are available only for students with disabilities designated as eligible for assessments based on modified or alternate achievement standards by IEP teams and only in those states that choose to establish modified and alternate achievement standards. The table also indicates the various "caps" on the use of proficient scores in AYP calculations.

Table 1

Type and Characteristics of Assessment Methods Based on U.S. Department of Education Policy for Inclusion of Students With Disabilities in Standards-based Assessment Used in Determining Adequate Yearly Progress (as of February 2006)

	Assessment Methods	Foundation for Content Assessed	How Performance Is Evaluated	Who Can Participate	Caps on Using Proficient Scores for AYP
Tests Based on Grade -level Achievement Standards	1. Regular assessment based on grade-level achievement standards	State's academic grade-level content standards	Grade-level achievement standards	Open to all students, including any student with a disability	None
	2. Regular assessment with accommodations based on grade-level achievement standards	State's academic grade-level content standards	Grade-level achievement standards	Any student with a disability. Some states make this option available to other students as well.	None

	3. Alternate assessment based on grade-level achievement standards	State's academic grade-level content standards	Grade-level achievement standards	Any student with a disability. Some states make this option available to other students as well.	None
Tests Based on other Achievement Standards	4. Assessment based on modified achievement standards*	State's academic grade-level content standards	**Modified** achievement standards	Student with a disability who can make progress toward, but may not reach, grade-level achievement standards in the same time-frame as other students and who may need changes in the breadth or depth of the assessment to appropriately reflect his or her proficiency†	Proficient scores may be counted for AYP subject to a cap of 2.0 percent of all students assessed at the state and district levels; no limit on number who can participate in this option†
	5. Alternate assessment based on alternate achievement standards‡	State's academic grade-level content standards	**Alternate** achievement standards that promote access to the general curriculum based on professional judgment of high expectations	Student with the most significant cognitive disabilities	Proficient scores may be counted for AYP subject to a cap of 1.0 percent of all students assessed at the district or state level; no limit on number who can participate in this option

*Some states may choose not to use modified achievement standards.

† No final regulations had been established at the time this paper was released.

‡ Some states may choose not to use alternate achievement standards.

The Decision Framework

The paper titled “A Decision Framework for IEP Teams Related to Methods for Individual Student Participation in State Accountability Assessments” describes a systematic framework for IEP teams to determine the most suitable way for students with disabilities to participate in the annual statewide assessments consistent with the *IDEA* statute and regulations. A critical point is made in the paper that IEP teams must determine **how** students with disabilities participate in statewide assessments for accountability, not **whether** they participate. Decisions are to be made for each student individually and not linked to a disability category, classroom placement, or the student’s involvement in instruction related to functional or daily living skills. Furthermore, the participation decision is to be made for each academic subject separately (e.g., reading, mathematics).

Although four possible testing options are currently available and a fifth testing method has been proposed, individual states may adopt and present the methods differently; therefore, the IEP teams need to be familiar with the testing methods available for students with disabilities in their state. Their deliberations about testing methods must be based on a systematic decision-making process that takes into account the need for accommodations, alternate assessments, and use of alternate assessments based on alternate achievement standards or assessments based on modified achievement standards if available in the state, relative to the testing method used. The framework draws attention to the requirement that students with disabilities have access to the general curriculum (*IDEA*, 1997 and 2004) and re-emphasizes an important condition of assessment: Students need the opportunity to learn the material on which they will be tested. By adhering to seven principles that are described in the paper referenced above, IEP teams can ensure that students receive instruction based on grade-level academic content, and they can promote instructional practices supported by research.

In their decision-making about how students should participate and their selection of assessment methods, IEP teams are directed to consider the educational needs of each student by answering five questions. The excerpt from Table 2, below, displays how the framework links the five questions to the choice of testing methods. Ultimately, IEP teams need to base their assessment recommendations on students’ responses to special education, interaction with text, instructional supports, and accommodations and assistive technologies used in the administration of an assessment.

Table 2

*Decision Framework for Individualized Education Program Teams in Choosing an Assessment Method**

	Foundation for Content Assessed				
	Based on Grade-level Achievement Standards Testing Methods			Based on Other Achievement Standards	
Testing Methods **Questions**	Regular assessment	Regular assessment with accommodations	Alternate assessment	Assessment with **modified** achievement standards	Alternate assessment with **alternate** achievement standards
Question 1: In what way does the student access the general curriculum?	Student shows progress in the full scope and complexity of the grade-level curriculum, although the student may not yet be on grade level.			The student can make progress toward, but may not reach, grade-level achievement standards in the same timeframe as other students, and changes in breadth and depth of the materials taught would facilitate his or her access to the general education curriculum and the grade-level content standards.	Due to significant cognitive disabilities (e.g., in memory, transfer of learning), student needs extensive prioritization.

Table continues with remaining questions.

* See Table 2 in "A Decision Framework for IEP Teams Related to Methods for Individual Student Participation in State Accountability Assessments," for the complete table. The excerpt here is provided as an example of the questions that should guide the choice of assessment.

A Model for Including Students With Disabilities in Large-scale Assessment Systems

As discussed in the paper titled "Validity Evidence," the underlying premise of testing the academic achievement of students with disabilities is that such testing can be valid if certain

conditions are met satisfactorily. And it is important to always frame validity with the following two questions:

1. How valid is the interpretation of a set of test scores?
2. How valid is it to use this set of test scores in an accountability system?

The model for inclusion developed in the series of papers has three main components that a state education agency, local school district, and school would implement: (1) a systematic decision-making framework for determining the population of students appropriate for each of the assessment methods; (2) an approach to assessment; and (3) a validity argument that includes specific types of evidence to be collected when making changes in the approach to assessment to ensure full participation of all student populations (see Table 3).

Table 3

A Model for Including Students With Disabilities in Large-scale Assessment Systems

(1) Decision-Making for Participation	**(2) Assessment Approaches**	**(3) Collection of Evidence to Support Claims and Inferences**
Five methods of assessment for students with disabilities: • Regular assessment • Regular assessment with accommodations • Alternate assessment based on grade-level achievement standards • Assessment based on modified achievement standards • Alternate assessment based on alternate achievement standards	Testing approaches within a statewide assessment system • Multiple choice • Short constructed response • Rating scales and checklists • Portfolios • Performance tasks and events	Technical evidence used in validity argument **Procedural evidence** (how assessment decisions and processes are implemented) • Test Development and Administration • Alignment • Standard Setting **Statistical evidence** (empirical outcomes that result from implementation) • Reliability evidence (e.g., internal consistency, inter-rater agreement) • Item statistics (e.g., difficulty and differential item functioning) • Validity evidence (e.g., internal structures, response processes, and relationship with other variables) • Construct validity evidence (e.g., construct under-representation and construct irrelevant variance)

The key elements of the model include defining the population of students with disabilities who need to be included (in each of the five methods) in a large-scale assessment system, identifying the testing approach or approaches that have been adopted statewide, and collecting technical evidence supporting the validity argument in relation to any changes made in the testing approach. The model (1) focuses on a total assessment system in which students participate in any number of ways, and (2) is based on an iterative validation process of making claims about assessment approaches and then collecting evidence to support the claims. In assembling the evidence, a number of specific guidelines are based on the latest educational standards (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999) that address reliability evidence (such as internal consistency and inter-judge agreements) and validity evidence (content-related evidence, response processes, internal structures, and relations to other variables). The specific types of evidence depend on the decision-making framework used to include students with disabilities (column 1 in Table 3) and the approach to assessment (column 2 in Table 3). Students with disabilities reflect a diverse group; each student needs to be considered individually by his or her IEP team as it recommends appropriate participation in the large-scale assessment program. Likewise, the assessment approach used to measure student performance on grade-level content standards also needs to be considered, as it directs the kind of evidence that can and should be collected, given that each approach makes certain assumptions and relies on certain strategies to measure achievement. The validity evidence collected depends upon how students with disabilities participate and how the state enacts its large-scale assessment.

Although providing appropriate accommodations to students with disabilities on the regular large-scale assessment still allows educators to make inferences about proficiency on state content standards that are comparable to the inferences made about students' proficiency when accommodations are not provided, at some point changes are made that are significant enough to alter the breadth and/or depth of how grade-level content is measured. Significant changes present a shift in the inferences that are warranted, and the changes become a part of alternate assessments judged against different achievement standards. The series of papers take up two major issues when changes are made: (1) distinguishing between test accommodations that allow comparable inferences from the assessment and changes that result in different

inferences, and (2) changes in breadth and/or depth that maintain links to grade-level standards.

Changes in Testing

Often changes can be made that involve using supports when administering tests (e.g., use of assistive technologies, prompts, or scaffolds) to remove construct irrelevant variance and maintain the meaning of the construct being measured. When such changes are made, they can be considered accommodations and allow educators to make inferences that are comparable to those for assessments administered without accommodations. A list of such accommodations should be made available in administration manuals to provide test users an explanation about the accommodation and conditions under which the accommodation can or should be applied. In addition, technical documentation should be provided on the empirical evidence supporting the effects of using the accommodations. Both types of evidence need to provide support for making the same inference of proficiency when no accommodations are present.

When changes to the way tests are administered or taken modify the breadth and/or depth of items, the content of the test is being changed. In these kinds of changes, an alternate assessment is being considered and the critical issue is simply to determine what achievement standard is being applied. Whether the **modified** or **alternate** achievement standard is being used to judge proficiency, the inferences about proficiency are not the same as when the test is provided without or with accommodations. Although grade-level content standards are being used, their breadth and/or depth has been changed to warrant constraints to the inferences that can be made. The difference in the inferences between these two achievement standards lies in the procedural and empirical evidence collected. This evidence needs to be provided in both the technical documentation and the reporting systems.

Skill Development in Achievement Testing

Skills and knowledge from content standards typically evolve gradually across grades; as a consequence, it is difficult to develop items or tasks for a given grade (“on-grade items”) that are unique and not relevant for adjacent grades (“cross-grade items”). To reflect this progression of skills, different regular assessment test forms can be created specifically for each grade to tap grade-level content standards and then be statistically linked through a vertical scaling or linking process. A scaling process is generally one in which raw test scores (usually the total number of

correct responses) are transformed into standardized scores, with a particular mean and standard deviation. With vertical scaling, common (anchor) assessment items across grades are used so the score in each grade can be compared to scores from previous and subsequent grades. As a consequence, the assessment score across the grades can be placed on the same scale, and changes in value can be considered equal intervals. This linking is done to provide a common scale for showing growth across grades and to reflect the idea that skills develop in a sequence (e.g., a difficult item in an earlier grade becomes an easy item in a subsequent grade). This scale is constructed through a statistical process called “vertical scaling” in which anchor items are used in more than one grade-level test (e.g., an item appears in both the third- and fifth- grade item). The items within the scale measure the same construct, and scores are typically used to track yearly progress between adjacent grades.

Under certain conditions, these cross-grade items might be acceptable for alternate assessments based on grade-level achievement standards or assessments judged against modified achievement standards. The papers offer a series of questions and criteria that can be used to help gauge the degree to which cross-grade items are suitable. They may be appropriate where assessments are aligned with grade-level content standards, have been linked to cover a common cross-grade core of the curricula, and do not constitute a major breach of the construct being assessed (thus providing procedural evidence). Furthermore, statistical evidence needs to be collected to reflect a vertical scale, comparing the performance of students who take these items on grade level and others who take them as cross-grade items.

This same logic of vertical scaling or linking also may be important for assessments judged against modified achievement standards to ensure progressive levels of achievement across grade levels. Because most skills in reading and mathematics reflect a progression or sequence in which proficiency of subsequent skills is based on proficiency of earlier requisite skills, this sequence may be articulated as part of the validity evidence collected. Both types of evidence would nevertheless need to explicitly relate to the grade-level content standards through changes in the breadth and/or depth.

Validity of Inferences Made From Test Scores

Students with disabilities can and should demonstrate achievement even though some cannot do so in the regular large-scale assessment even after intensive, evidence-based interventions

and appropriate, allowable assessment accommodations. For alternate assessments, the inferences about proficiency on state content standards must take into consideration the individual needs of students with disabilities. With both **modified** and **alternate** achievement standards, an inference is made that the breadth and/or depth of content have been reduced to make the assessment content accessible to a subgroup of students with disabilities. To assist states in implementing participation methods, the papers provide further definitions of the inferences to be made for the different types of achievement standards (see Table 4).

The regular assessment with accommodations and the alternate assessment based on grade-level achievement standards permit the same inferences as the regular assessment. These three methods are considered comparable because they represent changes in the types of supports or assessment formats but **not** changes to the breadth and/or depth of the assessment content. Alternate assessments based on different achievement standards do not permit the same inferences because they are **not** comparable to those assessments. Modified and alternate achievement standards **do** represent changes to the breadth and/or depth of assessment skills and knowledge.

Table 4

Making Explicit the Inference for Each of the Achievement Standards

Assessments judged based on:

- **Grade-level achievement standards** are designed to enable inferences to the breadth and/or depth of standards as specified in the test specifications for the general education large-scale assessment without or with accommodations. Both the assessment with accommodations and an alternate assessment based on grade-level achievement standards allow comparable inferences. Inferences about comparability and meaning of proficiency are **not constrained by the assessment methodology.**
- **Modified achievement standards** are designed to enable inferences to grade-level expectations with specified levels of breadth and/or depth. Inferences about comparability and meaning of proficiency are **constrained by the assessment methodology.**
- **Alternate achievement standards** are designed to enable inferences to grade-level expectations that have been extensively prioritized but maintain high expectations for progress in the general curriculum and assume student performance is contingent on having the supports specified for the assessment. Inferences are **stipulated because of the assessment methodology**.

Conclusion

Including students with disabilities in assessment and accountability systems can involve five testing methods. How students with disabilities participate is determined by the IEP team and must be driven by student need, not disability category or placement. Given this need, changes can be made in the types of support provided (prompts, scaffolds, or assistive technologies) and/or in the breadth and/or depth of the assessment to allow students with disabilities to participate in the statewide assessment system. The decision to make any changes, however, is very important because the test scores from every testing method are used to calculate AYP and this use warrants validation and the collection of evidence (both procedural and empirical). In the end, improving the quality of an assessment system that fully includes students with disabilities should be ongoing, guided by a periodic review of technical quality that considers proficiency as a function of a validity argument.

References

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

Title I—Improving the Academic Achievement of the Disadvantaged: Final Rule, 68 Fed. Reg. 68,697-68,708 (Dec. 9, 2003) (to be codified at 34 C.F.R. pt. 200).

Individuals with Disabilities Education Act of 1997, 120 U.S.C. §1400 et seq.

Individuals with Disabilities Education Improvement Act of 2004, 20 U.S.C. §1400, H.R. 1350

No Child Left Behind Act of 2001, Pub. L. No. 107-110, 115 Stat.1425 (2002).

Technical Work Group (2005, August). *Including students with disabilities in large-scale assessment.* Paper presented at the U.S. Department of Education, Office of Special Education Programs 2005 Project Directors' Conference, Washington, D.C.

U.S. Department of Education (2004, April). *Standards and assessment peer review guidance: Information and examples for meeting requirements of the No Child Left Behind Act of 2001.* Retrieved December 8, 2005 from http://www.ed.gov/policy/elsec/guid/saaprguidance.doc.

The U.S. Department of Education is reviewing public comments received on the notice of proposed rulemaking regarding modified achievement standards. As this analysis is not completed, the content of this document may not necessarily reflect the final views or policies of the Department concerning modified achievement standards.

This document was produced under U.S. Department of Education Contract No. EDO4CO0025/0002 with the American Institutes for Research. Renee Bradley served as the contracting officer's representative. No official endorsement by the U.S. Department of Education of any product, commodity, service or enterprise mentioned in this report or on Web sites referred to in this report is intended or should be inferred.