STANDARDS AND ASSESSMENT APPROACHES FOR STUDENTS WITH DISABILITIES USING A VALIDITY ARGUMENT

The purpose of this paper is to illustrate the validation process for large-scale assessments using the standards-based assessments of two states while minimizing construct-irrelevant variance or construct underrepresentation. The term “construct-irrelevant variance,” as it applies to standards-based assessments, means that the test measures too many variables, many of which are irrelevant to the content standards. The term “construct underrepresentation” indicates that the tasks that are measured in the assessment fail to include important dimensions or facets of the content standards. This process emphasizes the decision-making process used to design an assessment and the collection of evidence, both procedural and empirical, to evaluate not just an outcome, but also all of the assumptions and decisions made in creating and administering assessments, and scoring students’ performance on specific tasks (items). Procedural evidence focuses on test development, the quality of the items and tasks, the assemblage of “items” into the total test, and the administration and scoring process. Empirical evidence documents content coverage (alignment between the content standards and the assessment), the stability and consistency in sampling behavior over replications, assessment “item” functioning, reliability of judgments and scoring, internal relations among assessment items, response processes, and external relations with other assessments.

In the first section, we describe how to determine the validity of accommodations. Test validity generally refers to the degree to which the inferences about students’ proficiency based on test scores are meaningful, useful, and appropriate. We begin by considering construct-irrelevant variance that might arise from making changes in the general education assessments with the use of accommodations. State standards are the central constructs and they become operationalized through the use of large-scale assessment systems. These assessment systems need to be analyzed carefully to identify the introduction of construct-irrelevant variance into the determination of proficiency. As elaborated by Messick (1989) in his extensive essay on test validity, the validity argument involves systematically collecting and using evidence to evaluate a claim of proficiency based on scores from standards-based assessments.

The second section of the paper compares two states' standards and alternate assessments to highlight the nature of an "item" or task within an assessment system; the assessment approach in the first state is portfolio-based, and in the second state it is task-based. In each of these systems, we consider the kinds of procedural and empirical evidence that need to be collected to evaluate the validity claim that performance on the large-scale assessment is an adequate indicator of proficiency. We focus in particular on construct underrepresentation in this analysis.

The third section of the paper presents seven principles for developing items and tasks to ensure the focus is on grade-level content standards. These principles should be used to guide the process for developing alternate assessments, whether they are judged against grade-level, modified, or alternate achievement standards. We consider the grade-level focus for developing tasks; the breadth, depth and complexity of the items and tasks; the overlap across participation options; development across grade levels; the need for universal design; and finally, what students can do and its relationship to scoring. These seven principles should allow states to develop an assessment system that is completely inclusive and seamlessly integrates all participation options.

The fourth section of the paper operationalizes these principles using an example of reading assessment in which we make changes to accommodate students with disabilities participating in large-scale assessments judged against grade-level content standards, as well as substantial changes that become part of the alternate assessment. We focus on the two types of changes: changes in the supports (assistive technologies, prompts and scaffolds) provided, and changes in breadth, depth and complexity. The three important components of the assessment model are drawn together in this paper.

1. A validation process is articulated using an argument with claims, assumptions, and evidence to evaluate the inferences that are made from the performance of students on assessment tasks representing selected domains of knowledge and skills (i.e., content standards). The process must begin with content standards that need to be operationalized into tasks (items) used to assess student proficiency. The collective tasks represent an approach to assessment.
2. This approach to assessment is then analyzed in its administration for students with disabilities who require appropriate assessment accommodations. If members of the IEP team deem the need for an alternate assessment, other approaches are then

analyzed that rely on indirect teacher judgments (using rating scales and checklists), portfolios, performance events and performance task collections.

3. The assessment system as a whole is finally considered as representing a range of options for the participation of individual students, each of whom can access the general education content standards based on their unique needs. Ideally, the validation process is supportive in both process and outcomes, but the results are tentative and require further attention if any changes are made.

In the fifth section, conclusions and recommendations address these three issues. The recommendations make the validity argument explicit in its assessment approach (i.e., the process of operationalizing content standards into tasks or test items used to assess student proficiency); the options in which students participate (i.e., participate in the regular assessment with, or without, accommodations, or participate in an alternate assessment); and the manner in which changes are made (i.e., changes made to a test item, task, or format).

Construct-Irrelevant Variance and the Need for Accommodations

To understand a validity argument it is essential to have a clear idea of which construct is being tested because it forms the basis of the claim of validity. For example, in the following standards from Massachusetts and Oregon involving a mathematics problem from a fifth-grade state practice test, a word-story problem is presented as a multiple-choice item. It is essential to know whether this test item also has within it other constructs that are irrelevant to the mathematical construct being tested.

To answer the question we must consider (1) the construct being assessed; (2) the knowledge and skills reflected in the specific tasks and the manner in which this knowledge and these skills are sampled, formatted and scored; and (3) the use of test scores to make inferences about the teaching and learning process as well as the accountability system (relative to the construct). The **validity claim** is that the test adequately reflects the **domain of knowledge and skills** of the standards and can be used as the basis for the inference of proficiency.

Table 1

Construct in an Example of a Mathematics Standard and an Assessment Problem[1]

Oregon Standard:	Add and subtract decimals to hundredths, including money amounts.
Massachusetts Standard:	Select and use appropriate operations (addition, subtraction, multiplication and division) to solve problems, including those involving money.
Assessment Problem:	Tommy bought 4 shirts for $18.95 each and 3 pairs of pants for $21.49 each. What was the total Tommy spent?
Assessment Options:	A) $ 11.33 B) $ 135.97 C) $ 139.27 D) $ 140.27

The validity argument considers whether the task presented on the large-scale assessment appropriately measures the domain of achievement or whether it is misrepresented or underrepresented as described in Table 2.

Table 2

Validation Claim and Questions Supported by Evidence

Construct	**Misrepresentation**	**Underrepresentation**
Achievement (domain of tasks)	Does the math story problem include other constructs or rely on access or prerequisite skills that prevent students from displaying their knowledge and skill?	Does the math story problem adequately represent the kind of mathematics operations needed to solve money estimation problems in the presence of suitable distracters (i.e., irrelevant elements of the problem)?

In this simple mathematics problem, reading may be part of construct-irrelevant variance that impedes our efforts to measure the mathematical knowledge and skills as applied in this limited situation (a printed math story problem). However, if we had used a performance task to measure achievement (open-ended problem requiring the student to write his or her answer),

[1] Examples excerpted from (1) the Oregon Department of Education's fifth grade mathematics content standards for computation and estimation, available at: http://www.ode.state.or.us/teachlearn/real/Standards/Default.aspx: (accessed March 25, 2006); and (2) the Massachusetts Department of Education's fourth grade mathematics content standards for number sense and operations, available at: http://www.doe.mass.edu/frameworks/math/2000/num3.html (accessed March 24, 2006).

then writing may have become part of the construct-irrelevant variance. If we had required a demonstration of money estimation in a local store or in the community, however, a host of other factors that are part of the assessment (the type of store in which we shopped, the presence of others at the check out, the bills being used, etc.) would then have become sources of construct-irrelevant variance. Construct-irrelevant variance can arise from several sources, including from the unique needs of students with disabilities or groups of individuals and how they participate in large-scale assessment systems. This source of variance is systematic and either consistently disadvantages or advantages individuals or groups. For example, if students are allowed only 60 minutes to complete a reading test, students with poor reading skills will be consistently disadvantaged. Or if students are given read-aloud assistance and the tester inadvertently prompts the correct choice by inflection, students taking the test from this person are systematically advantaged. In both examples, math performance is confounded with (influenced by) other characteristics of the measurement process that are irrelevant to the construct being measured.

In the math story problem as a measure of achievement, the construct also can be seriously underrepresented, failing to include appropriate operations (addition, subtraction, multiplication or division), steps (making exact change or estimations of change), distracters (elements of the problem that need to be seen as irrelevant), or critical strategies (use of self-guided actions that were used by the student but not documented). In all of these instances, the construct may have been underrepresented.

The validity claim can be threatened by several factors, for example, by insufficient evidence. And in making the claim, serious social consequences are at stake. Misinterpretations could be made (e.g., the student is not proficient in mathematics). Resources could be misdirected (e.g., very complex tasks are used that require intensive manpower to administer and score, for which reliability-related evidence is found lacking). Tasks could be misrepresented as constructs because measurement specialists, content experts, and special educators fundamentally disagree with (or are uninformed by) each other. Knowing the limitations of assessments for making inferences about proficiency in cognitive skills using more complex tasks, it is important to emphasize the need for appropriate and credible assessment approaches.

Accommodations

We introduce accommodations to remove construct-irrelevant variance by making changes in the supports (and not in making changes in the content domains). For example, the mathematics problem could be read aloud to students who cannot read well to eliminate reading as a construct-irrelevant variable. Likewise, we could use a calculator to remove the computational requirements for mathematics problems targeting other constructs. We also could allow more time so the student can finish the item (or test). Tindal and Ketterlin-Geller (2004, p. 8) note the following in their review of mathematics accommodations research on four major classes of accommodations (using calculators, reading mathematics problems to students, employing extended time, and using multiple accommodation packages). Notice, however, that these task (test) features may be problem- and person-specific.

In general, the findings from using calculators and reading mathematics problems to students clearly document the effect of accommodations to be dependent on the type of items and populations. For some items, calculators are facilitative (e.g., solving fractions problems) and for others detractive (e.g., on complex calculations as part of mathematical reasoning). Similarly, item specific findings are beginning to appear in reading mathematics problems: when the problems are wordy (both in count and difficulty) and contain several verb phrases, the accommodations appear effective. Likewise, student characteristic is an important variable. The positive effects of the read-aloud accommodation are more likely with younger students or those with lower reading skills. Finally, the use of extended time appears relatively inert though often it appears as part of other accommodations. For example, calculators and reading mathematics problems often take more time.

Thus, the research on accommodations reflects that changes in the way tests are given or taken (the supports used) indeed can make a difference, sometimes removing construct-irrelevant variance. Furthermore, the effect of an accommodation is dependent on characteristics of the population using the accommodation. At other times, however, accommodations may actually introduce construct-irrelevant variance (e.g. teachers systematically provide extra prompts). So, accommodations cannot be considered a panacea or a simple process. Their usefulness depends on the construct of the standard, the assessment approach or format, and the needs of the student.

At this point in time, most states have both participation and accommodation policies. These policies, however, focus mostly on **who** needs to participate and **how** they should participate,

and less on **why** certain types of participation options should be recommended or applied. This statement is particularly true for the use of accommodations. Very few states have policies that explain the reasoning behind an accommodation in terms of the intended construct to be measured and the evidence needed to support its measurement (see Thurlow & Bolt, 2001). We address that kind of evidence through the consequences of assessment, most of which are seriously underreported (c.f., National Center on Educational Outcomes *Online Accommodations Bibliography).* In the end, states need to have policies on what accommodations to allow **and why;** these policies need to provide IEP teams guidance in determining how the unique needs of students with disabilities require changes in testing.

Table 3

Types of Accommodations

Presentation	**Presentation Equipment**	**Response**	**Setting**	**Scheduling**
Large print	Magnification equipment	Proctor/scribe	Individual	Extended time
Braille	Light/acoustics	Computer or machine	Small group	With breaks
Read-aloud	Calculator	Write in test booklets	Carrel	Multiple sessions
Interpreter for instructions	Amplification equipment	Tape recorder	Separate room	Time beneficial to student
Read/reread/ simplify/clarify	Templates/ graph paper	Communication device	Seat location/ proximity	Over multiple days
Directions	Audio/video cassette	Spell checker/ assistance	Minimize distractions/ quiet/reduced noise	Flexible schedule
Visual cues on test/instructions	Noise buffer	Braille	Student's home	Other
Administration by other	Adaptive or special furniture	Pointing	Special ed. class	
Additional examples	Abacus	Other	Other	
Other	Other			

Alternate Assessments

The general education large-scale assessment (with or without accommodations, or when it involves multiple administrations) is intended to allow educators to make comparable inferences about proficiency on state standards. Yet, at some point, changes are made that are significant enough to constrain the inference, which is when states need to consider them as part of their alternate assessments. In this type of assessment, constraints begin to appear in the inference about proficiency on standards. Because of changes in supports (assistive technologies, prompts or scaffolds) and/or changes in the breadth, depth, and complexity of the material being tested, the **scores** on alternate assessments based on alternate achievement standards cannot be aggregated with the **scores** on regular assessments (and therefore must be reported separately). However, as explained later in this paper, using a validity argument within the context of federal regulations allows for the aggregation of **proficiency levels** based on grade-level, modified, and alternate achievement standards for purposes of reporting Adequate Yearly Progress.

In the sample mathematics problem presented at the beginning of the paper, changes could be made in the assessment approach by observing the student actually making change and using a checklist or rating scale to note the correctness of the response, by assembling into a portfolio materials that document the student making change during an interaction at a local store in the community, or by observing or recording a performance task given to the student in which the student is required to add these amounts of money using real bills and make change accordingly. All of these options could become part of an assessment judged against modified achievement standards or an alternate assessment judged against alternate achievement standards. Remember, however, that these “situated” environments may well introduce other sources of irrelevant variance unrelated to the construct. Therefore, each of these approaches brings with it the need to collect specific kinds of evidence to ensure that the construct is being fully assessed (and not underrepresented), requiring both procedural and empirical evidence.

Validity Argument Using Different Alternate Assessment Approaches

We integrate the validity process, assessment approaches, and populations of students with disabilities by considering two states with considerably different grade-level standards and alternate assessments. For this illustration, we focus on mathematics content standards for grades three to five. Although the selection of states and grade levels was somewhat arbitrary,

some related research has been published previously that aids in making this illustration (see Weiner, 2002 for a description of Massachusetts and Tindal et al., 2003 for a description of Oregon).

Each of the assessment strategies used in an alternate assessment (whether judged against grade-level, modified, or alternate achievement standards) needs to be analyzed using the same validity claim: The test reflects the domain of knowledge and skills for the construct and the tasks that have been sampled. The procedural evidence focuses on test development, the quality of the items and tasks, the assemblage of "items" into the total test, and the administration and scoring process. Empirical evidence documents content coverage (alignment between the content standards and the assessment), the stability and consistency in sampling behavior over replications, "item" or task functioning, reliability of judgments and scoring, internal relations among items and tasks, and response processes, as well as external relations with other measures. Just as for establishing the validity of the general education test (with or without accommodations), attention needs to be given to construct-irrelevant variance and construct underrepresentation in alternate assessments; this latter problem is particularly critical as changes are being made in depth, breadth and/or complexity of the standards.

Massachusetts Content Standards and Alternate Assessments With Portfolios

We confine our analysis to the content standards for grades three and four that focus on number sense (which has seven objectives) and operations (which has three objectives), both critical areas for understanding mathematics. These standards have a certain breadth and depth and, as we will see in comparison to Oregon's standards, represent a very reasonable alignment with a number of mathematics constructs that focus on number sense and operations, fitting well with the standards from the National Council of Teachers of Mathematics (2000).

Table 4

Massachusetts Standards for Grades Three and Four

No.	Number Sense Standards	Essence
4.N.1	Exhibit an understanding of the base ten number system by reading, modeling, writing, and interpreting whole numbers to at least 10,000; demonstrate an understanding of the values of the digits; compare and order the numbers.	♦ Manipulate numbers at a higher level by counting, writing, grouping, sorting, comparing and ordering. ♦ Use a variety of numerical forms/classes. ♦ Recognize and use decimals. ♦ Understand and compare equivalent forms of decimals and fractions.
4.N.2	Represent, order and compare large numbers (to at least 100,000) using various forms, including expanded notation, e.g. 853 = (8 × 100) + (5 × 10) + 3.	
4.N.3	Demonstrate an understanding of fractions as parts of unit wholes, as parts of a collection, and as locations on the number line.	
4.N.4	Select, use and explain models to relate common fractions and mixed numbers (1/2, 1/3, 1/4, 1/5, 1/6, 1/8, 1/10, 1/12, and 1 1/2); find equivalent fractions, mixed numbers, and decimals; and order fractions.	
4.N.5	Identify and generate equivalent forms of common decimals and fractions less than one whole (halves, quarters, fifths, and tenths).	
4.N.6	Exhibit an understanding of the base ten number system by reading, naming, and writing decimals between 0 and 1 up to the hundredths.	
4.N.7	Recognize classes (in particular odds and evens, factors or multiples of a given number, and squares) to which a number may belong, and identify the numbers in those classes. Use these in the solution of problems.	
No.	**Operations Standards**	**Essence**
4.N.8	Select, use and explain various meanings and models of multiplication and division of whole numbers. Understand and use the inverse relationship between the two operations.	♦ Understand the meaning of multiplication and division. ♦ Represent multiplication and division problems
4.N.9	Select, use and explain the commutative, associative and identity properties of operations on whole numbers in problem situations, e.g. 37 × 46 = 46 × 37, (5 × 7) × 2 = 5 × (7 × 2).	

4.N.10	Select and use appropriate operations (addition, subtraction, multiplication and division) to solve problems, including those involving money.	concretely. ♦ Use all operations to solve problem situations related to money. ♦ Understand commutative properties of addition and multiplication (order can be reversed).

To provide a functional equivalence across standards, an "essence" of the standard is distilled in Massachusetts, where standards are translated into some minimal specifications (see Table 4 for examples of standards distilled into essences or minimal specifications) that eventually are used in guiding the development of alternate assessments based on alternate achievement standards. (In contrast, Oregon's content standards are affixed to common curricular goals across grades and are applied directly to the alternate assessments based on alternate achievement standards.) These grade-level standard "essences" are then used in Massachusetts to fully articulate the alternate assessment system to ensure alignment with it and to help structure the assessment approach (in this state, portfolios). For each grade-level standard, the state has illustrations posted on its assessment Web site (retrieved on May 30, 2005); see Table 5 for structuring **activities** related to the standard (left column) and documentation or end product portfolio entry (right column) that eventually is judged as proficient (or not).

Table 5

Application of Massachusetts Learning Standards and Assessment Strategies

How can all students participate in this assessment activity?	
Addressing Learning Standard(s) as Written for This Grade Level	**Possible Assessment Strategies and Portfolio Products**
Ricardo participates in a cooperative group activity with classmates to solve open-ended mathematical problems involving money. They make multiple purchases and compare selections, estimates, total cost and change received.	• Ricardo's flyer/catalog and work samples of items bought, estimations made, amount spent and change received • Chart of Ricardo's grades/scores on quizzes and tests related to mathematical problem solving • One copy of a quiz/test chosen by Ricardo for his portfolio • Journal entry in which Ricardo reflects on his work samples and performance on the quiz/test

Addressing Learning Standard(s) at Lower Levels of Complexity ("Entry Points")	**Possible Assessment Strategies and Portfolio Products**
Dominique participates by making purchases with her classmates. She selects items for purchase and indicates the amount needed by identifying the "next highest dollar" from the price given. A vertical number line provides Dominique with support so she can participate independently in this activity.	• Dominique's vertical number line • Work products in which she selected her purchases and indicated the number of dollars she needs • Dominique's graph, created with teacher assistance, demonstrating her accuracy in identifying the "next highest dollar" for purchases of $10 or less • Photographs of Dominique making purchases in a variety of settings (classroom, cafeteria, school store, drug store, etc.) with her number line using the "next highest dollar" method
Addressing Access Skill(s) (skills embedded in academic instruction)	**Possible Assessment Strategies and Portfolio Products**
Alice participates in this activity by assembling money envelopes paired with pictures. Alice works with a classmate who counts the money needed for each item and helps Alice place the correct amount into its corresponding envelope. Alice exchanges these envelopes when making a purchase.	• Teacher note describing the work accomplished by Alice and her classmate • Data collected on Alice's ability to assemble money envelopes and exchange correct envelopes when making a purchase • Videotape of Alice making a purchase • Alice's choice of money envelopes selected for her portfolio

The assessment activities provide highly connected portfolio products aligned with the "essence" of the standards. The primary issue we address here is the need for making a validity claim and then collecting both procedural and empirical evidence to evaluate the claim.

Procedural evidence arises from the processes used by teachers while they assess the student:

- Was the test developed in a way that is consistent with testing standards and are the scoring procedures credible?
- What is the quality of the portfolio entries and are they formatted in a way that is understandable and accessible?
- How are work samples assembled and organized into a total portfolio?
- How well conducted are the test administration and scoring procedures?
- Do the various assessment activities of the alternate assessment represent the "essence" of the standard?
- Does the actual evidence described in the possible assessment strategy fully reflect this construct?

- Are all representations in the portfolio well displayed so they can eventually be scored?
- Does the student independently complete work that is displayed in the portfolio or is the teacher part of this process and, if so, to what extent does the teacher assist the student?

Score reporting and analyses should address questions such as how standards were established and what kinds of statistical procedures were used to analyze the outcomes. Technical documentation should be available to determine how clear and consistent the results reported to the public are.

Empirical evidence also needs to be established by investigating the dependability and credibility of work samples as reflections of the construct:

- How carefully do teachers collect evidence for the portfolios?
- Are judges who score the portfolio adequately trained?
- Is there agreement among their scores?
- Are there appropriate forms of reliability estimation?

Content coverage needs to be documented:

- What is the alignment of the portfolios with the standards?

Internal structures and item functioning need to be addressed:

- How are different dimensions like generalization, independence, and achievement scored?
- How well do work samples “hang together” to reflect a standard?
- Which work samples in the portfolio are related to each other and to what extent do those samples reflect a consistent structure?

Response processes should be considered as part of the validation evidence:

- Are there patterns in how students respond that reflect systematic variance, for example, whereby all students assessed with real money are judged successful and those solving “artificial” word story problems all fail?

Finally, evidence needs to consider relations with other variables:

- Do students with all disabilities reach proficiencies in equal proportions and what other demographics are related to outcomes?
- How is performance on the portfolio related to any other performances?

All of these are examples of empirical evidence: reliability-related evidence, content-related evidence, internal structures, response processes, and relations with other variables.

We offer this system and these questions to highlight the essential issues that **all** states must eventually face as they develop a fully articulated assessment system for students with disabilities. The assessment system must be sensitive to the needs of students, instructionally relevant to teachers, and related to grade-level content standards.

Oregon Content Standards and Alternate Assessments Using Performance Tasks

We now turn to another state in which comparable standards have been articulated, though with a slightly different grade level (four to five). In Oregon, the standards are affixed to common curriculum goals (CCG) and in this particular example, three CCGs were used to reflect similar state standards: **number sense** standards in Massachusetts are very close to the **calculations and estimations** standards in Oregon; **operations standards** in Massachusetts are similar to the **computations and estimations** AND **operations and properties** in Oregon. The grain size (i.e., specificity of a content standard) is slightly different, which eventually has a bearing on any alignment analysis. In Massachusetts, there are two standards with 10 objectives (seven in one standard and three in another) while in Oregon, there are three standards with 19 objectives or almost double the number in Massachusetts: five objectives address **numbers**, 11 objectives address **computation and estimation**, and three objectives address **operations and properties.**

As noted with the Massachusetts content standards, Oregon's alternate **assessments** based on alternate achievement standards are linked to the grade-level content standards, but differ in complexity as compared to the grade-level achievement standards set for the regular assessment. The specificity varies somewhat from that used in Massachusetts, but obvious overlap is present with a focus on basic mathematics operations (e.g., multiplication and division), fractions, properties and types of numbers (whole and real).

Table 6

Oregon Mathematics Standards: Numbers, Computation and Operations — Grades Four to Five

Common Curriculum Goals	**Oregon Grade-Level Standards** **Grade Four**
Calculations and	

Common Curriculum Goals	Oregon Grade-Level Standards Grade Four
Estimations	
Understand numbers, ways of representing numbers, relationships among numbers, and number systems.	**NUMBERS** Read, write, order, model, and compare whole numbers up to 1,000,000, common fractions, and decimals up to hundredths. Identify the place value and actual value of digits in a number to 1,000,000. Locate common fractions and decimals on a number line. Model, recognize and generate equivalent forms of decimals to hundredths. Determine factors of whole numbers to 100 using models such as arrays.
Understand meanings of operations and how they relate to one another.	**OPERATIONS AND PROPERTIES** Demonstrate the meaning of fractions as part of a unit whole or as parts of a collection or set. Use inverse operations (addition and subtraction, multiplication and division) to solve problems and check solutions involving calculations with whole numbers. Apply the commutative, associative, and identity properties of addition and multiplication and the distributive property to simplify calculations with whole numbers.
Compute fluently and make reasonable estimates.	**COMPUTATION AND ESTIMATION** Develop and evaluate strategies for multiplying and dividing whole numbers and adding and subtracting fractions with like denominators. Apply with fluency efficient strategies for determining multiplication and division facts zero to nine. Multiply a three-digit number by a one-digit number. Divide a three-digit number by a one-digit number with or without remainders. Determine the meaning of whole number remainders in a problem situation. Add and subtract commonly used fractions with like denominators (halves, thirds, fourths, eighths, tenths) and decimals to hundredths. Add and subtract decimals to hundredths, including money amounts. Mentally multiply or divide multiples of 10 (e.g., 40 $\times$ 70 or 2,700 $\div$ 30). Identify the most efficient operation (add, subtract, multiply, or divide)

Common Curriculum Goals	Oregon Grade-Level Standards Grade Four
	for solving a problem. Select and use an appropriate estimation strategy (overestimate, underestimate, range of estimates) based on the problem situation when computing with whole numbers or money amounts. Use place value concepts such as rounding to nearest 10, 100 and 1,000 to estimate and check reasonableness of answers.

In Oregon, these standards have not been “essentialized” but are applied directly to the alternate assessments. Furthermore, in contrast to the portfolio approach used in Massachusetts, a performance assessment is used in Oregon. A sample of items is displayed in Figure 1. For each standard, two items are displayed: (1) a practice test that is used by teachers to get students familiar with the test format; and (2) an actual performance task used in the alternate assessment.

Figure 1

Oregon Mathematics Standards: Numbers, Computation and Operations—Grades Four to Five

Standard: *Supply a missing element in or determine a rule that extends number patterns involving addition or subtraction of decimals.*
Standard*: Describe*, *extend, and make generalizations about patterns and sequences and supply missing elements in chart or table format.*

Practice Item 9: What is the next number in this sequence?
2, 10, 7, 15, 12, 20, 17, 25, __

(A) 22 (B) 30 (C) 31 (D) 32

Alternate Assessment Task 9: Number Line
Present the flashcards face up in a random order in front of the student. Say, "Here are many different numbers." Place the ruler in front of the student and say, "Here is a number line with a missing number. Tell me which number is missing."

7 8 9 10 11 13 14 15

Options: 7, 12, 17, 11, 13

Standard*: Read, write*, *order, model, and compare whole numbers up to 1,000,000, common fractions, and decimals up to hundredths.*

Practice Item 24: Find the missing number in the pattern.
2.6 5.2 ___ 20.8

(A) 7.8 (B) 10.4 (C) 13.0 (D) 15.6

Alternate Assessment Task 11: Order Numbers
Present the number cards in this order: 3, 1, 8, 6. Say: Place these numbers in order from smallest to largest.

Standard*: Identify the place value and actual value of digits in a number to 1,000,000.*

Practice Item 24
What is the expanded notation for 3,056?
(A) 3,000 + 5 + 6 (B) 3,000 + 50 + 6 (C) 3,000 + 500 + 6 (D) 30,000 + 50 + 6

Alternate Assessment Task 18: Place Value
Use three cards for this task: 508, 72, 431
Place the card that has the number 508 on the table in front of the student. Ask*: "Which digit is in the tens place?"*
Place the card that has the number 72 on the table in front of the student. Ask*: "Which digit is in the ones place?"*
Place the card that has the number 431 on the table in front of the student. Ask: "*How many hundreds are there?"*

As in the Massachusetts examples, the point is to highlight issues within a validity argument. In this Oregon example, we focus on the alignment between the general education test and the alternate assessment, addressing the kind of changes that have been made in the tasks to assess students' grade-level expectations using the state's large-scale assessment (a multiple choice test) or proficiency using performance tasks within an alternate assessment judged against alternate achievement standards. While it is easier to define an item in Oregon, other problems may appear in the manner that the items are presented to students and the manner in which students can respond. What is the functional equivalence of these alternate assessments (i.e., to what extent is measured student performance on the alternate assessment equivalent to measured student performance on the regular assessment)? Do these tasks adequately sample the domain of achievement in mathematics and are they adequately represented in content and format?

As stated earlier, **procedural evidence** needs to be collected to evaluate the inference that is being made. Have the tasks been adequately developed and assembled into an alternate assessment? Are the directions for administering the test clear and understandable? Are teachers sufficiently trained in administering and scoring the tests (especially because responses may be scored as partially correct and not just correct or incorrect)? In this particular alternate assessment, teachers are trained on how to make accommodations (vs. modifications): Are the directions for administering these two different types of assessments sufficiently clear? Are student performances accurately recorded and entered into the computer?

Empirical evidence also needs to be collected to evaluate the inferences being made about the construct (the standards) being assessed. Are enough tasks present to reflect the domain (and avoid construct underrepresentation)? What is the alignment of the alternate assessment to the state content standards at this grade level in categorical concurrence, depth of knowledge, range of knowledge and balance of representation? How reliable are teachers in administering tests with accommodations where necessary? How well do items within tasks relate to each other (reflecting internal consistency)? Which tasks "hang together" and are related to each other (reflecting internal structure)? Are there any systematic issues in the response processes (given that they are all constructed responses)? For example, given that they require a constructed response, are there any sensory limitations that preclude students with vision or hearing impairments from responding? What is the relationship between

performance on these tasks and other variables (e.g. type of program, gender of student, disability, or other demographic characteristics)? As noted earlier, these different types of evidence reflect reliability, content, internal structures, response processes, and relation with other variables.

Principles of Task-item Development and the Validity Argument Applied to Populations

In describing the state standards-based assessments in Massachusetts and Oregon earlier in this paper, we looked at the process for collecting both procedural and empirical evidence. In Massachusetts, the evidence came from portfolios as a way to make inferences about proficiency on state standards, while in Oregon, performance tasks were used to make such inferences. These types of evidence, however, can be collected only from existing accountability assessment systems, which begs the question of how to develop effective items and tasks on which to base accountability. In particular, we need to ask how to develop an assessment approach that both reflects grade-level content standards **and** fits the specific, unique needs of individual students.

In this section, we present an overview of a set of universal design principles that can guide test development so that a sizable proportion of students with disabilities can be included appropriately and meaningfully in inferences from testing based on state standards. In addition, for some students, changes in supports as well as breadth, depth and complexity are needed so they can meaningfully participate in a large-scale assessment system. The most important issue is the application of test item development principles that can be applied by states to fit within their particular assessment approach.

Principle 1: Items and tasks should be derived from grade-level content standards in the core academic area.

Grade-level content standards should be the primary resource for developing alternate assessment items and performance tasks. Weak alignment for academic content is likely to occur when items are developed from resources other than the general content standards for the grade level (e.g., from a separate functional curriculum or lower grade-level curriculum) and then back-mapped to the grade level. Poorly aligned indicators[2] are likely to be too broad or

[2] Because the nomenclature to describe standards and assessments is different across the states, we used common language to describe the levels of specificity within the standards. The following levels, from the most general statement to the most detailed description of the standards, were used in this

too narrow, vague, age inappropriate, or not representative of the academic content (Browder et al., 2004). In a 2003 study, Browder et al. found that, for items in state alternate assessments identified as being most closely aligned with national standards in reading and mathematics, more academic tasks **and** contexts were used.

Using grade-level materials and activities in developing items and performance tasks also provides an important strategy for ensuring that the items and performance tasks are grade appropriate. Alternate assessments may sample knowledge forms and skills typically mastered at earlier grades by using grade-appropriate items and performance tasks. For example, a fifth-grade alternate assessment in reading may include emergent literacy skills (e.g., print awareness, picture comprehension) to assess achievement, but the items and performance tasks should use adaptations of the literature from the fifth-grade reading series for the assessment rather than preschool materials. Or, for example, the fifth-grade alternate reading assessment as judged against either modified or alternate achievement standards may include items that focus on early decoding skills but measure listening comprehension of fifth-grade literature using assistive technology to decode the text, listening comprehension as someone reads the story aloud, and/or reading comprehension of the same story simplified with controlled vocabulary based on grade-level content standards (to the extent appropriate) and read independently. Including this range of tasks to reflect skill development and content provides assurance of both appropriate and aligned assessments.

Some students may be able to demonstrate an academic concept for the grade level if given a real-life scenario or activity. For example, a student may be able to demonstrate mathematical comparisons (=, >, <) through comparative shopping. The sufficiency of items directly influences the reliability-related evidence for understanding the repeatability or stability of behavior; a sufficient number of items or performance tasks will help ensure that the student understand the concept rather than simply be able to perform a particular activity. Sufficiency may be achieved in two ways. First, the student may be asked to demonstrate the same concept through multiple activities with different materials. Second, the student may be asked to summarize the task

study: (a) **subject area** (e.g., mathematics); (b) **content standards** (e.g., students develop number sense and use numbers and number relationships in problem-solving situations and communicate the reasoning used in solving these problems); (c) **objectives** (e.g., using numbers to count, to measure, to label and to indicate location); and (d) **performance indicators** (e.g., describe numbers by their characteristic—for example, even, odd, prime, square). In this study, we used the term **assessment item** to represent the performance response that could be a behavioral event or a student work sample (Flowers, Browder, & Ahlgrim-Delzell, 2004).

using the traditional symbols, text, or notations (e.g., select the equation that matches how the task was completed). When using tasks or activities derived from real life, it is important to have the items or performance tasks validated by grade-level content area experts to determine the extent to which the integrity of the concept has been maintained. Also, construct-irrelevant variance needs to be analyzed, to ensure that the tasks (or measurement approaches) do not introduce effects that confound appropriate interpretations of student performance.

Principle 2. The alternate assessment should parallel the breadth, depth and complexity of the general content standard for the grade level.

Alternate assessments should be developed to address sufficient breadth, depth, and complexity of the curriculum, which will vary depending on the grade level being assessed. For example, in the elementary grades, mathematics or science curricula often cover a wide breadth of content with little depth. In mathematics, students may receive some exposure to computation, data analysis, geometry and measurement in a single year. In contrast, high school-level mathematics may provide more depth but only a single mathematics strand (e.g., geometry). Similarly, alternate assessments should be comparable to the regular assessment with respect to breadth, depth, and complexity. That is, if the third-grade science curriculum covers nine strands of science, the alternate assessment judged against grade-level achievement standards should address these nine strands; otherwise, assessment guidelines should contain specific rationale for assessing fewer strands and how the additional strands are sampled in subsequent grades. Assessments judged against modified achievement standards and alternate assessments judged against achievement standards reflect systematic constriction of the breadth, depth and complexity of the grade-level content standards and that needs to be described explicitly.

An important point in this principle is that the depth of content coverage (not just the breadth) also needs to be considered. In most skill areas, a developmental progression is at least implied (if not explicitly provided). Rarely are students merely exposed to content without an analysis of the skills needed to be successful in the content. Often, earlier skills are needed to express later skills (for example, sufficient vocabulary knowledge and word-level reading skills, beyond background knowledge, are needed for success in appreciating various literary or comprehension skills). Therefore, breadth of curriculum coverage needs to be accompanied by considerations of depth of skill development and the complexity of items to reflect appropriate skill progression.

Principle 3. The alternate assessment should consider the complexity of knowledge reflected in the curriculum and the content standards.

In an evaluation of the alignment of three states' alternate assessments with state standards, Flowers, Browder, and Ahlgrim-Delzell (in press) found that, although all three states included items at all levels of knowledge complexity, items were oriented toward simpler levels of knowledge compared to the more complex levels reflected in the state standards. Although an assessment judged against modified achievement standards and an alternate assessment judged against alternate achievement standards may be similarly oriented, it is important to include items and performance tasks that sample multiple levels of knowledge complexity. For example, in assessing comprehension of a story, the student may be asked to answer questions with simple factual recall, to compare the character to him or herself, or to predict what the main character might do next. These multiple levels of knowledge complexity also should create overlap between the assessment options so that students approaching grade-level performance have sufficient opportunity to demonstrate achievement. For example, a student with a significant cognitive disability participating in an alternate assessment based on alternate achievement standards should be presented with some items designed to demonstrate proficiency on an assessment based on modified achievement standards to ensure that the alternate assessment validly measures the proficiency of higher-performing students with the most significant cognitive disabilities.[3]

Principle 4. Items and tasks for alternate assessments need to reflect appropriately the constructs of the standards and include an appropriate skill sequence.

Although alternate assessments can be developed so they are aligned with state content standards, care must be taken not to distort that standard. However, most skills are developed within a progression and it is unlikely that good intentions alone (e.g., getting to grade-level content standards) will result in items and tasks that align with grade-level standards. Learning to read implies a progression of skills that are vertically aligned. For example, graphemes, phonemes, and morphemes comprise the basic linguistic units in an alphabetic writing system; words can be read in ways that allow for segmenting and blending phonemes with graphemes; sentences and passages reflect both local and global linguistic dependence; comprehension is a complex construct with multiple features that build in complexity. Mathematics is similarly

[3] While this paper makes this recommendation, it is important to note that the U.S. Department of Education's Notice of Proposed Rulemaking for modified achievement standards does not make this requirement.

sequential in the basic development of number sense, ordinal properties, operations and computations, complex relationships (place value, fractions and probabilities), and specific applications within geometry and algebra as well as within various contexts (applications). These skill sequences need to be considered in making changes to assessments so that specific items are still aligned with the content standards, perhaps as part of a skill sequence that is vertically aligned, and also reflect both an appropriate assessment for the student and an accurate representation of the construct.

Principle 5. Alternate assessment items and performance tasks should be developed to show sequential achievement across grade levels.

While general education assessments often reflect a spiraling curriculum in which students revisit academic concepts at increasingly difficult levels, the intended outcome is progressively complex levels of knowledge and skill application. Similarly, alternate assessments should be developed so that students can demonstrate progressive levels of achievement. Although a student may revisit the same science concept, for example, in eighth grade as in third grade, some added achievement (often involving more complexity and depth) should be expected in eighth grade. This added achievement might be reflected as additional content (increased depth or breadth), increased levels of complexity, or a wider range of applications in the eighth grade as compared to the third grade alternate assessment.

Principle 6. Alternate assessment items and performance tasks should be universally designed, avoid bias, and ensure access to the content.

Alternate assessment items and performance tasks should be developed so that students with sensory, physical, and behavioral challenges have a means to perform the skills and tasks being assessed. For example, if a performance task requires physically arranging items to show comprehension of a concept, an alternate response should be available to communicate this arrangement if physical challenges preclude manipulating the materials. If the task requires viewing pictures or text, using Braille, reading material aloud, or using small objects may be alternate ways to present the item to students who are visually impaired. The decision to offer support changes that supplement or supplant changes in breadth, depth, and/or complexity needs to be made carefully. Often, a simple change in supports causes little change in the construct (which is much more preferred than changes that reduce the breadth, depth, or complexity). However, changes of breadth, depth, and/or complexity may be very appropriate when considering the skill sequence of the construct and the individual needs of the student

(e.g., accurate and fluent word-level reading is important and it may be negligent to avoid this skill sequence in focusing only on comprehension).

Principle 7. Students may receive partial credit for supported or prompted responses, but the scoring of items and performance tasks should focus on what the student does.

One way to determine if a student has gained competence on a grade-level content standard as judged against modified or alternate achievement standards is to give the student the opportunity to demonstrate the concept with additional support. For example, the assessment may include assistive devices (visual cues), response prompting, or scaffolding. When used, consideration should be given to the extent to which the student could demonstrate the concept with minimal as compared to more extensive support. For example, a student who can locate a correct answer when provided more information has greater understanding of the concept than one who finds it by imitating a modeled answer. No credit should be given for responding that requires no active response or thought on the part of the student (e.g., physically guiding the student to make the response). An important issue to consider in using prompts and models, however, is the consistency with which they are presented and the effect this practice has on ensuring the scores are comparable.

Examples of Reading Assessment Items and Tasks

With these guiding principles, we now present examples of test items and tasks that have been changed (in supports as well as breadth, depth and/or complexity) to reflect accommodations or assessments to be judged against modified or alternate achievement standards. In reading through these examples, it is important to remember that accommodations are intended to allow for comparable inferences to be made in assessment results (in relation to proficiency on state content standards). That is, if the change results in measurement of the same construct, then it would be considered an accommodation and the scores might be aggregated with the general education test. With assessments not based on grade-level achievement standards, however, the inferences that can be made are somewhat constrained (in the case of performance on an assessment judged against modified achievement standards) and quite stipulated (in the case of performance on an alternate assessment judged against alternate achievement standards). That is, if the change results in the measurement of a different construct (e.g., listening comprehension instead of reading comprehension), it would become part of an assessment as judged against modified achievement standards or an alternate assessment judged against alternate achievement standards. Therefore, while the **scores** of alternate assessments based

on alternate achievement standards and the **scores** of assessments based on modified achievement standards cannot be aggregated with the regular assessment **scores** (and, therefore, need to be reported separately), student **proficiency** levels based on grade-level, modified, and alternate achievement standards must be aggregated for AYP purposes. The critical distinction is that scores cannot be aggregated but proficiency levels must be aggregated (based on different achievement standards).

We have described the use of two types of changes to ensure appropriate participation in large-scale testing: (1) use of supports in the manner that tests are administered or taken (e.g., use of assistive technologies, prompts, or scaffolds) and/or (2) changes in the breadth, depth, or complexity of the tasks and items used in the assessment. Of course, it is possible to change both the supports and the breadth, depth, and complexity, as these categories are not meant to be exclusive. We present these examples of test items and tasks only as illustrations. Clearly, these examples need to be embedded within a state assessment and accountability system and then subjected to an analysis using a model that includes a validation argument in which evidence is collected using an approach to assessment on a specified population of students.

It is important to note that accommodations are based only upon changes in supports and not changes in the breadth, depth and/or complexity of the grade-level achievement standards. Similarly, for alternate assessments judged against grade-level achievement standards, the accommodations can include changes in supports but not in the breadth, depth and/or complexity of the grade-level achievement standards. In contrast, alternate assessments based on alternate achievement standards may differ in breadth, depth, and/or complexity from the grade-level achievement standards, and assessments based on modified achievement standards may differ in breadth or depth (but not complexity) from the grade-level achievement standards. However, an important caveat to these changes is that all assessments (i.e., based on grade-level, modified, or alternate achievement standards) must be **aligned with** the grade-level content standards. The reason for this caveat is so that comparable inferences can be made to the content standards. Table 7a describes a fourth-grade-level standard and two kinds of changes that could be made in the assessment.

Table 7a

Decode or Comprehend Meaning of Words in Text

Standard	**Changes using assistive devices, scaffolds and prompts**	**Changes in breadth and depth**
1.1 Distinguish, reproduce and manipulate the sounds in words.	Use recognition device to distinguish; use matching device to reproduce or manipulate.	Control the words (or word types) being presented

Changes that could be considered **accommodations:** The verb “reproduce" in the standard means that the student must produce the word. Nevertheless, the word can be presented to the student in a list, on cards, in text, and juxtaposed with other words. The student can be asked to rehearse the word or prompted with various directions to reproduce the word.

Changes that could be considered part of the **assessment based on modified achievement standards:** In this scenario, two types of changes could be made:

1. **Support changes** (assistive devices, prompts and scaffolds) could be used to make the inference to this standard constrained. For example, the student could be asked to recognize words (e.g., demonstrated through receptive mode communication rather than expressive mode communication). All of the words considered from the general education test would apply otherwise. For example, the words would be read to the student with cards presented; the student who has been presented with several word cards would be asked to point to the word being read.
2. **Breadth and depth changes** would limit the inference to the full range of words being referenced in this grade-four standard. With this change, the full range of words would not be used; a subset of phonetically regular words and the top 100 sight words may be used. The subset of words represents a change constraining the breadth and depth of the curriculum, while the inclusion of phonetically regular words represents a reduction in the complexity of the assessed construct. Or within these domains, the sampling plan would be done with replacement, which means that once sampled to appear on the test, each word would be placed back into the total domain to be potentially sampled again; as a result, the same word could appear more than once on the test. Therefore, the inference would be that the student knows a sample of words but perhaps not the whole range of all words presented in grade-four vocabulary or text.

Changes that could be considered part of the **alternate assessment based on alternate achievement standards:** Again, two types of changes could be made:

1. **Support changes** (assistive devices, prompts, and scaffolds) would limit the inference to that student and those specific changes. For example, the assistive device is a prompt unique to that student and needed with every response (e.g. the student repeats the word after it is read and then sounded out).
2. **Breadth, depth, and/or complexity changes** would limit the inference to only those words being used in the test and not to the domain of words in the fourth grade. For example, only one-syllable words would be used, sampling from both phonetically regular words as well as the 10 most frequent words that account for 24 percent of the English language text: **the, of, and, a, too, in, is, you, that, it.**

Table 7b

Decode or Comprehend Meaning of Words in Text

Standard	**Changes using assistive devices, scaffolds and prompts**	**Changes in breadth, depth or complexity**
2.1.1 Demonstrate knowledge of phonetics, word structure (root words, prefixes, suffixes, abbreviations) and language structure through reading words in text (word order, grammar).	Use assistive reading.	Control the readability of text and increase type-token ratio (unique words to total words).

Table 7b describes another fourth-grade standard and the two kinds of changes that could be made in assessing students with disabilities. **Accommodations** for this standard include giving the student different directions and the opportunity to respond in different settings with different administrators or in multiple sessions. Nevertheless, the skill of actually reading words in text would be tested. As the standard/objective is written to specifically include word order and grammar, word lists would not be allowed.

Changes that could be considered part of the **assessment based on modified achievement standards:** In this scenario, two types of changes could be made:

1. **Support changes** (assistive devices, prompts and scaffolds) could be used to make the inference to this standard more limited by allowing the student to use a recognition

response of “yes” or “no” by nodding his or her head after text is read with a card of the written text placed in front of them. “Yes” would signify the text being read and the text on the card are the same, while “no” would signify they are different. A second example might be to present the student with two passages, one of which has numerous syntactic and semantic errors. The student would identify the passage that is most correct.

2. **Breadth, depth, and/or complexity changes** would limit the inference to the full range of text being referenced in this grade-four standard. In this change, the text being read might be reduced in its complexity using a readability formula; a type-token ratio may be changed so that more words are repeated and fewer unique words (those appearing only once) would be present in the passage.

Changes that could be considered part of the **alternate assessment based on alternate achievement standards:** In this scenario, two types of changes could be made:

1. **Support changes** (assistive devices, prompts and scaffolds) would limit the inference to that student and those specific changes. The student may have a fixed vocabulary of words written so that they appear in three- to five-word sentences that are displayed one sentence per line (not wrapped as in normal text).
2. **Breadth, depth, and/or complexity changes** would limit the inference to only those words being used in the text and not to the domain of text in the fourth grade; in this example, the passage may be highly constricted to those words that are known to be within the student’s reading vocabulary.

Summary and Recommendations

We began this paper by considering construct-irrelevant variance and construct underrepresentation. In both, the starting point is to be clear about the construct and two kinds of distortions that can arise in measuring it. Clear inference about achievement can be prevented by the manner in which items and tasks are given or taken; often, other knowledge and skills are being assessed in addition to those viewed as central to the construct. Distortion comes about when the assessment contains very few opportunities for assessment of the construct.

Recommendation 1 — For each item or task, be clear about the central construct AND consider the access skills that are needed to complete the item or task (and what other related knowledge or skills are becoming entwined with this central construct). A number

of changes can be made to the way that a test item, task, or test is given or taken. We listed several of them in Table 3, grouped into five major types (presentation, presentation equipment, response, scheduling, and setting), considering them as accommodations. Using these accommodations results in a score that can be aggregated with the general education scores and should yield comparable inference about achievement of the standard.

Recommendation 2 — When changes are allowed in testing and deemed to NOT change the inferences that can be made about the central construct, accommodations should be accompanied by an explanation that differentiates them from changes causing a different inference (which then becomes part of the alternate assessment). In reflecting on the validity argument, we considered two states, presenting both the state standards and example items and tasks. The first state (Massachusetts) used a portfolio approach as its alternate assessment; the second state (Oregon) used performance tasks. After describing the approaches to assessment, we considered the kinds of procedural and empirical evidence that need to be collected for evaluating any validity argument.

Recommendation 3 — The assessment system needs to be described in its entirety with both grade-level content standards and associated assessments, providing both procedural and empirical evidence that supports any inferences about proficiency. We presented seven universal design principles for developing items and tasks that should be considered in developing any alternate assessments, whether based on grade-level, modified, or alternate achievement standards. These seven principles provide primarily procedural evidence that supports the development of items and tasks and assures a strong relationship between them and grade-level content standards.

Recommendation 4 — Principles of item and task development need to be clear in explicating their connection to grade-level content standards to ensure appropriate breadth, depth, and complexity and to design alternate assessments with appropriate inferences. We illustrated different changes that could be made in (1) the supports used as part of an assessment, as well as (2) the breadth, depth and complexity of grade-level content standards. These changes need to be considered on a continuum in which accommodations use changes in support only and alternate assessments use either or both types of these changes.

Recommendation 5 — The inferences made from the general assessment with accommodations should be designed to be comparable to the general assessment without accommodations, but even with the greatest design efforts to align with the state's grade-level academic content standards, inferences are expected to be (a) somewhat constrained for an assessment judged against modified achievement standards and (b) quite stipulated for alternate assessments judged against alternate achievement standards. In summary, the validation argument should provide states the necessary framework for ensuring that any claims about proficiency on grade-level content standards have meaning and are supported by noting the explicit assumptions and evidence that are part of the large-scale assessment system. This argument allows states to use a variety of content assessment approaches (with various task formats) and to systematically provide a variety of participation options for students with disabilities.

References

Browder, D., Flowers, C., Ahlgrim-Delzell, L, Karvonen, M., Spooner, F., & Algozzine, R. (2004). The alignment of alternate assessment content to academic and functional curricula. *Journal of Special Education, 37,* 211–224.

Browder, D. M., Spooner, F., Ahlgrim-Delzell, L., Flowers, C., Algozzine, R., & Karvonen, M. (2003). A content analysis of the curricular philosophies reflected in states' alternate assessments. *Research and Practice for Persons with Severe Disabilities*, *2,* 165–181.

Flowers, C., Browder, D. M., & Ahlgrim-Delzell, L. (2006). An analysis of three states alignment between language arts and mathematics standards and alternate assessment. *Exceptional Children, 72,* 201-215*.*

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: American Council on Education.

National Center on Educational Outcomes (2005). *Online Accommodations Bibliography.* Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved Dec. 8, 2005 from http://education.umn.edu/NCEO/AccomStudies.htm.

National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics.* Reston, VA: Author.

Thurlow, M., & Bolt, S. (2001). *Empirical support for accommodations most often allowed in state policy* (Synthesis Report 41). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved Dec. 8, 2005 from http://education.umn.edu/NCEO/OnlinePubs/Synthesis41.html.

Tindal, G., McDonald, Tedesco, M., Glasgow, A., & Almond, P., Crawford, L., Hollenbeck, K. (2003). Alternate assessments in reading and math: Development and validation for students with significant disabilities. *Exceptional Children, 69*(4), 481–494.

Tindal, T., & Ketterlin-Geller, L. R. (2004). *Research on mathematics test accommodations relevant to NAEP Testing.* Washington, DC: National Assessment Governing Board. Available at http://www.nagb.org/pubs/conferences/tindal.pdf (accessed March 21, 2006).

Wiener, D. (2002). *Massachusetts: One state's approach to setting performance levels on the alternate assessment* (Synthesis Report 48). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved Dec. 8, 2005 from http://education.umn.edu/NCEO/OnlinePubs/Synthesis48.html.

This document was produced in December 2005 under U.S. Department of Education Contract No. ED4CO0025/0002 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.