Technical Report # 0508

Including Students with Disabilities in Large-Scale Testing: A White Paper for Establishing Federal Policy

Gerald Tindal – University of Oregon

Pat Almond – Consultant

Diane Browder – University of North Carolina

Lindy Crawford – University of Colorado, Colorado Springs

Steve Ferrara – American Institutes for Research

Tom Haladyna – Arizona State University

Huynh Huynh – University of South Carolina

Naomi Zigmond – University of Pittsburgh

Published by

Behavioral Research and Teaching
University of Oregon • 175 Education
5262 University of Oregon • Eugene, OR 97403-5262
Phone: 541-346-3535 • Fax: 541-346-5689

Note: This paper was written with funds from the U.S. Department of Education. Originally, the paper was posted on a website that was terminated, so the paper has been published as a Technical Report.

Tindal, G., Almond, P., Browder, D., Crawford, L., Ferrara, S., Haladyna, T., Huynh, H., & Zigmond, N. (2005). *Including Students with Disabilities in Large-Scale Assessments: A White Paper for Establishing Federal Policy (Technical Report # 0508).* Eugene, OR: Behavioral Research and Teaching, University of Oregon.

INCLUDING STUDENTS WITH DISABILITIES IN LARGE-SCALE ASSESSMENTS

EXECUTIVE SUMMARY WITH PRINCIPLES

SECTION I – A MODEL FOR PARTICIPATION OF SWD IN LSA

Chapter 1 – Overview of the Model, Types of Evidence, and Implementation Issues

Chapter 2 – Validating Assessments for Students with Disabilities

Chapter 3 – Approaches to Assessing Students in Statewide Assessment Programs

Chapter 4 – A Decision Framework for IEP Teams Related to Individual Student Participation in State Accountability Assessments

Chapter 5 – Standards and Assessment Approaches using A Validity Argument

SECTION II – ISSUES IN GATHERING VALIDITY EVIDENCE

Chapter 6 – Test Design, Development, Scoring, and Analysis

Chapter 7 – Evidence of Item Quality

Chapter 8 – Reliability Issues and Evidence

Chapter 9 – Validity Evidence

SECTION III – ISSUES IN IMPLEMENTATION

Chapter 10 – Score Reporting

Chapter 11 – Documenting the Validity of Using Test Scores of Students with Disabilities for Accountability

Chapter 12 – Consequences of Assessments

Chapter 13 – Implications for Professional Development

Glossary – Terminology

Abstract

This White Paper offers a framework for including students with disabilities in large-scale assessment and accountability systems. Responding to IDEA and No Child Left Behind requirements, it distinguishes participation through general assessments, accommodations, and alternate assessments judged against grade-level, modified, or alternate achievement standards. The authors frame validity as an evidence-based argument linking test scores to interpretations and decisions, emphasizing alignment with grade-level content standards and careful attention to breadth, depth, complexity, knowledge, and skills. Guidance addresses assessment design, item quality, scoring, reliability, validity evidence, reporting, accountability uses, consequences, and professional development. The paper stresses that IEP teams should make individualized, annually reconsidered participation decisions based on student needs rather than disability labels or placement. It calls for transparent technical documentation, ongoing evaluation of intended and unintended consequences, and collaboration among measurement specialists, educators, administrators, families, and policymakers to ensure equitable access, defensible score interpretations, and educationally beneficial assessment practices.

CHAPTER 1: EXECUTIVE SUMMARY

The mandate to include students with disabilities in large-scale assessments began with the legislative requirements of IDEA 1997 of NCLB, and finally of IDEA 2004. As a result of these legislative acts, research and development have increasingly focused on accommodations and universal design (UD) applied to general assessments. This legislation also served as the stimulus for research and development on alternate assessments, including those judged against alternate achievement standards and more recently those judged against modified achievement standards.

In 1997, research on general education testing with accommodations was just beginning to develop as a national research agenda. In the first major review of accommodations published by Tindal and Fuchs (1999), 106 studies from 20 years of research were described in terms of author and date, nature of the accommodation, populations addressed, methodology employed, measures used, and findings obtained. The studies were grouped into several major categories of accommodation that had been explicated earlier by Thurlow, Ysseldyke, and Silverstein (1993): presentation, presentation equipment, response, scheduling, and setting.[1] Research on accommodations has continued with funding from several federal competitions and through the initiatives by state educational agencies (SEAs). Indeed, the research also has changed considerably to become even more encompassing with the development of *universal design for learning* as described by Rose (2000). The National Center on Educational Outcomes has more recently applied this concept to *assessment* in two key publications by Thompson, Johnstone, and Thurlow (2002) and Johnstone (2003). Whether oriented to *learning* or to *assessment*,

[1] This initial compilation and review of research on accommodations have been continued by the National Center on Educational Outcomes with the development of an electronic database: http://education.umn.edu/NCEO/AccomStudies.htm.

universal design has three key attributes: (a) multiple venues are provided for access, (b) a wide range of standard features are considered initially to minimize the need for post-hoc changes and accommodations, and (c) access skills are scaffolded (Tindal & Crawford, 2005).

For a period of five years (from 1997 to 2001), states also diligently developed various forms of alternate assessments. Two approaches were most popular: "(a) simplify the regular standard until we can find something (anything!) that the student can do, or (b) redefine the regular curriculum standard so that it represents some type of functional skill" (Ford, Davern, & Schnorr (2001, p. 213-214).

Then, in 2001, President Bush signed into law the most sweeping educational reform in the country: No Child Left Behind (NCLB). This law stated that schools would be held accountable for achievement – that is, for making Adequate Yearly Progress (AYP) through annual testing. Specifically, students in grades three through eight would be tested each year and the results would be disaggregated by economic background, race and ethnicity, English proficiency, and disability. This phrase of 'no child left behind' quickly created a decisive rallying call for reform of assessments and a change in the road map being used to create opportunities for students with disabilities to participate in large-scale testing programs, whether using accommodations or through alternate assessments.

On December 9, 2003, further changes were instituted in the determination of Adequate Yearly Progress (AYP), allowing up to 1% of the population, students with 'the most significant cognitive disabilities,' to be counted as proficient based on participation in alternate assessments judged against alternate achievement standards; these alternate achievement standards were defined as "an expectation of performance that differs in complexity from a grade-level achievement standard" (Federal Register, December 9, 2003, p. 68699). States were allowed to

define multiple achievement standards but nevertheless had to "employ commonly accepted professional practices to define the standards" (p. 68699). The key issues in that regulation was that "an alternate assessment must be aligned with the state's content standards, must yield results separately in both reading/language arts and mathematics, and must be designed and implemented in a manner that supports use of the results as an indicator of AYP" (Federal Register, 2003, p. 68699). Further language noted that the alternate assessments had to be of high technical quality, including validity, reliability, accessibility, objectivity, and consistency with nationally recognized professional and technical standards. This list of words was not defined but providing meaning to them and describing procedures for succeeding in their attainment are an important focus of this volume.

In 2004, IDEA was reauthorized with several significant changes[2], but what did not change was the requirement to include all students with disabilities in statewide assessment and accountability systems. Then, on May 10, 2005, Education Secretary Margaret Spellings announced that states would be allowed to develop alternate assessments judged against modified academic achievement standards for students with persistent academic disabilities and

[2] For example, teachers would no longer be required to write short-term objectives (STOs) as part the student's Individualized Educational Program (IEP). Three-year evaluations would no longer be required unless requested by a teacher or parent; otherwise, "a reevaluation might be warranted if the child's performance in school significantly improves, suggesting that he or she no longer requires special education and related services, or if the child is not making progress toward the goals set out in his or her IEP, indicating that changes are needed in the education or related services the LEA is providing" (Apling & Jones, 2005, p. CRS-20). And the concept of 'resistance to intervention' (RTI) was introduced as a way of stemming the burgeoning identification rates for special education, particularly of children with learning disabilities. This concept has been in the professional literature for some time (see Fuchs and Fuchs, 1998 for a discussion of 'treatment validity'). Most recently, examples of the logic undergirding 'resistance to intervention' as an identification tool can be seen in the work of Fuchs, Speece, Torgesen, and Vaughn (c.f., McMaster, Fuchs, Fuchs, & Compton, 2005; Speece, Case, & Malloy, 2003; Torgesen, 2000; Vaughn, Linan-Thompson, & Hickman, 2003).

served under the IDEA. States also would be allowed to count as proficient scores from such assessments in making AYP decisions, but the number of such scores would be capped at 2.0% of the total tested population. At the same time, states would be allowed to continue to use alternate assessments based on alternate achievement standards for students with the most significant cognitive disabilities (which would remain capped at 1% for making AYP decisions) (see http://www.ed.gov/print/policy/elsec/guid/raising/alt-assess-long.html#notes)[3]

Of course, the basic requirements of NCLB still held, even with this new population of students (1% and then 2%) being allowed to meet proficiency by taking alternate assessments as judged against modified or alternate achievement standards, respectively. For example, statewide participation still had to be at 95% with valid alternate assessments in place by 2006-2007; appropriate accommodations still had to be made available to students with disabilities to ensure valid interpretations; policies still had to be in place and training made available for IEP teams;

[3] The following research was used to support the announcement that, for an additional for 2% of the population, alternate assessments judged against *modified* achievement standards was needed. First, the 1% cap was deemed too low if the Department of Education followed the (then current) definitions of students in one of the 13 disability categories who cannot reach grade-level standards. Second, several studies supported increasing the percentage of students who needed an alternate assessment as judged against modified achievement standards. For example, the best designed instructional interventions were successful for only 50% to 90% of participating students in reaching grade level standards, resulting in anywhere from 0.5% to 3.0% of the total population being eventually labeled as learning disabled (Lyon, Fletcher, Fuchs, Chhabra, in press). And, in the research by Torgeson Wagner, Rashotte, Rose, Lindemood, Conway, and Garvan (1999), almost one-quarter of the students did not reach grade level standards even though they had received explicit reading instruction; this amounted to about 2.4% of the total population unable to reach grade level expectations. Finally, in Fletcher's summary of remedial research from Mathes, Denton, Fletcher, Anthony, Francis, & Schatschneider (in press) and McMaster, Fuchs, Fuchs, and Compton (in press), 10% of the students did not reach reading benchmarks indicative of average proficiency, which was equal to 2% of the total population. In a summary of four studies by Fletcher (2005), lack of response to intervention (RTI) was found: Klingner, Vaughn, Hughes, Schumm, & Elbaum (1998). Foorman, Francis, Winikates, Mehta, Schatschneider, & Fletcher, (1997), Torgesen, Alexander, Wagner, Rashotte, Voeller, & Conway (2001), and Fletcher, Denton, Fuchs, & Vaughn (2004).

and all alternate assessments still had to be reviewed as part of the Department's peer review process (http://www.ed.gov/print/policy/elsec/guid/raising/alt-assess-long.html).

Making Decisions by Conceptualizing Validity as an Argument

As a result of these legislative and regulatory actions, a large group of students with disabilities currently participates in our state large-scale assessment systems. Yet it is unknown how or if the assessments are appropriate as so many questions remain. What are the inferences that can be made from current state assessment options? How are accommodations deemed appropriate or inappropriate? How are alternate assessments developed and linked to grade-level content standards? How are states to distinguish between students participating in the alternate assessment judged against modified achievement standards versus alternate achievement standards? What kind of information is needed to answer these questions or what kind of evidence is needed to justify state initiatives?

To address these questions, we propose a validity argument that focuses not only on the descriptive interpretations of the test scores but also on the decision-based inferences that can be made from the outcomes (Kane 2001). In either case, it is an "interpretive argument, or network of inferences and supporting assumptions leading from scores to conclusions and decisions" (p. 5). Like any argument, several assumptions and inferences typically are associated with it. In essence, the validity argument is about the plausibility of the interpretations: Are the assumptions reasonable, do the conclusions follow from them, and are the conclusions supported by evidence?

An interpretive argument helps explicate the assumptions and the evidence to justify a decision. What is critical to understand, then, is that "it is the interpretation of test scores that is validated" (Kane, 2001, p. 7). For example, the additional allowance of 2% can be questioned in

terms of the appropriate population (students with high incidence disabilities), but it performance in reading was used to justify the additional allowance, it is unlikely that success in grade-level content standards can occur without basic knowledge and skill in reading to learn. This argument needs to be made in an open forum, with evidence collected to support it.

The Focus of Change to Relate Options to Decisions

In developing a validity argument, we make some assumptions about *context, content* and *knowledge forms (concepts, principles, procedures) and skills*. These terms are quite specific and should not be confused. For example, *context* typically is placement-oriented (e.g., the general education classroom) while *content* refers to the focus and type of information that is being presented or received in the curriculum. Content is typically topical: it is about 'what' is being taught and learned. We assume that within each state, teachers are using state content standards to organize the content, though considerable latitude is still possible in organizing this content. *Knowledge forms* are the basis for organizing concepts, principles, and procedures. *Concepts* reflect a knowledge form that provides the substance of our standards; they include labels, attributes, and examples and non-examples; they also can be concrete or abstract. Likewise, *principles* are a knowledge form expressed as "if-then" relationships. *Procedures* are applications of algorithms within problem solving activities. Finally, *skills* are cognitive routines expressed in applying specific symbol systems (alphabetic or numeric). Whether knowledge forms or skills, the changes made need to be part of the validity argument. The integration and application of knowledge and skills comprise a newer concept in student learning that focuses on cognitive abilities, such as reading comprehension, writing, or mathematical problem solving.

In using these terms, we make the general argument that it is the *knowledge (forms) and skills* that are the basis of manipulation in determining whether a construct (of a standard)

remains unchanged in meaning (e.g. reflects an accommodation) or is changed (in the alignment with the full breadth and depth of standards or in its complexity) and therefore becomes part of an alternate assessment. When it is changed, we only need to further ascertain whether it is to be generalized to some grade level content standards or stipulated to only the grade-level standard to which it is linked. Like all arguments, we make certain assumptions so that we can focus exclusively on the knowledge forms and skills.

First, we assume that students with disabilities are provided access to the general education curriculum as stipulated in IDEA 97 and reaffirmed in 2004. We have remained neutral about the context for this access given the least restrictive environment options required by IDEA and the expectation that students with disabilities in every type of setting (general class, self contained class, separate school) should have access to the general curriculum. In contrast, we realize that providing access to the general curriculum in contexts outside of the general education classroom is creating some of the current challenges to the field, such special education teachers becoming highly qualified in the content area they teach and finding ways to teach grade level content in separate settings. We assume that we can discuss content without discussing context (which may be a false assumption and therefore addressed in the policy implications as well as consequences). Nevertheless, this assumption does allow us to focus primarily on the knowledge forms (concepts, principles, and procedures) and skills that underlie most state grade level expectations.

Second, we assume that content (materials) can be manipulated in the knowledge forms and skills that are to be emphasized in teaching or learning concepts, principles, and procedures. The three ways that content can be manipulated involve changes in the breadth, depth, and complexity of knowledge forms or skills.

Breadth: How many objectives in the standards are addressed at each grade level?

Depth: How many sub-objectives in the standards are addressed at each grade level (which may vary in complexity)?

Cognitive complexity: What level of cognitive demand is required in the standards at each grade level?

Third, we assume that the process of manipulating knowledge forms and skills must be planned (logical and intentional) as well as appropriate to define the construct for the target audience. There must be documented answers to the questions: Why and how are decisions made? How is the content organized from grade to grade? What is the evidence to document the process and what kind of construct representation is possible or can be avoided? Has the essential knowledge form and skill of the grade level expectation been preserved?

Accommodations modify delivery of assessment to reduce impediments to access to the assessment (due to language and disabilities) but do not alter the construct: the inference about what students know and can do is equivalent to the inference made to those without accommodations. Changes in breadth, depth, and/or cognitive complexity reduce the inferences made (at the descriptive level) as well as have an impact on decisions. The difference between 2% and 1% is primarily in this interpretative distinction: when an alternate assessment judged against modified achievement standards is used, there is less manipulation of breadth, depth, and cognitive complexity than when extensive prioritization has been made in the breadth, depth, complexity, and/or level of support for 1% of the population. In the former manipulation (modified), inferences can be made to grade-level standards; in the latter manipulation (alternate), the inference is stipulated.

Over-Arching Issues that Became Apparent

In this analysis of assessment options, decisions, and change, we have come across various issues that we had to address to allow us to proceed with this volume. What follows is a summary of various complexities within each issue.

Assumption 1 – This volume references large-scale assessments as required by No Child Left Behind

The language of NCLB uses the term '*assessment'* throughout the government web pages (http://www.ed.gov/admins/lead/account/saa.html#ayp), even though the measurement field probably reserves this term for something different than is specified in the language of the law or the regulations that followed. NCLB requires annual *assessments in reading and mathematics* for grades three through eight beginning in 2005-2006, with the addition of science *assessments* in 2007-2008 (but only in same three grade spans as the 1994 law). It requires *reading assessments* using tests written in English for any student who has attended school in the US (excluding Puerto Rico) for 3 or more consecutive years, with local educational agency discretion to use tests in another language for up to 2 additional years. States also must annually assess English proficiency for all limited English proficient students beginning with the 2002-03 school year. Finally, it requires, beginning in school year 2002-03, biennial State participation in *NAEP reading and math assessments* for 4th and 8th graders so long as the Department pays the costs of administering those *assessments*.

In traditional measurement parlance, "a test is a measuring instrument intended to numerically describe a characteristic under uniform, standardized conditions … Responses to tests are scorable. The use of scoring rules helps to create a test score based on a test taker's responses to the test items" (Haladyna, 1999, p. 4). In contrast, assessment is "a process of

collecting data for the purpose of making decisions about individuals and groups" (Salvia & Ysseldyke, 2003, p. 5). This distinction is important because measurement experts would likely argue that no single score should ever be used to make a high stakes decision, yet the language of NCLB is to award proficiency (for accountability use at the school and state level) based on standards that typically are established for single tests. We emphasize this issue because the primary focus of this volume is on those single tests used as part of the large-scale assessment system.

Assumption 2 – Inferences depend on the achievement standards used to make judgments, which are considered the same as performance standards.

The term achievement standard does not appear in the 1999 *Standards for Educational and Psychological Testing*. Rather, the term *performance standards* are used with two definitions (p. 179):

1. An objective definition of a certain level of performance in some domain in terms of a cut score or a range of scores on the score scale of a test measuring proficiency in that domain.

2. A statement or description of a set of operational tasks exemplifying a level of performance associated with a more general content standard; the statement may be used to guide judgments about the location of a cut score on a score scale. The term often implies a desired level of performance.

We have adopted the second definition, above, to define *achievement* standards. As a group, we wrestled over the seeming confluence of content standards (what knowledge and skills are being required as grade level expectations) with performance/achievement standards (the level of proficiency in performance on these content standards). In the end, we maintained the integrity of the content standards as noted in the December 9, 2005, regulations: "To serve the

purposes of assessment under title I, an alternate assessment must be aligned with the State's content standards…" (Federal Register, December 9, 2003, p. 68699).

Assumption 3 – We applied the same grade level content expectations or standards for all students to set high and appropriate expectations for students with disabilities to learn reading, mathematics, and science.

We use the validation arguments of Kane (2001) to highlight the logic of evidence needed to push for policy and reform of current practice. This issue reflects the classic chicken-and-egg dilemma: When many students with disabilities fail to achieve proficient status on large-scale assessments, is it because of the inappropriateness of their preparation or the (in)appropriateness of test they took?

The field of large-scale assessments is beginning to acquire increasing information on the use of accommodations for students with high incidence disabilities; we still know very little about how students with the most significant disabilities can or will perform on these alternate assessments. For example, in 1993, 28 states had *written policies on the participation* of students with disabilities in their tests that nevertheless were quite varied; 21 states had *written policies on accommodations* (Thurlow, Ysseldyke, & Silverstein et. al., 1993). By 1997, this number was 40 states with active policies on accommodations and by 2000, 48 of 50 states reported active policies on accommodations (Thurlow, House, Boys, Scott, & Ysseldyke, 2000).

We know very little, however, about how students with the most significant disabilities can or will perform on alternate assessments. Many states are still developing appropriate alternate assessments. Some of the confusion may have come about from distinctions between alternate assessments to be judged against grade-level content standards while also permitting states to have "multiple alternate achievement standards" (Federal Register, December 9, 2003,

p. 686999). No doubt, further confusion is likely to result from the difference between alternate assessments judged against modified as opposed to alternate achievement standards.

Assumption 4 – Systematic decision-making must be used for students with disabilities to participate in large-scale assessment programs and, with annual testing, assessment options must be reconsidered and changed in subsequent years for individual students.

The Technical Work Group struggled over ensuring that students with disabilities did not become labeled in the process of finding an appropriate assessment. For example, we chose not to define a population of students to participate in any option, whether the option was viewed as an accommodation or an alternate assessment judged against modified or alternate achievement standards. Rather, we defined specific behaviors that might warrant taking the test in any of these three ways.

Continually, we emphasize in this volume that evidence needs to be collected to support the way students participate in large-scale assessments and therefore, the interpretation that can be made may need to be reconsidered. Individualized Educational Program (IEP) teams need to go beyond a student's disability to find the most appropriate way for them to participate. We provide considerable information for IEP teams to use in helping determine the best fit between the assessment approach, the individual student and option for participation.

Assumption 5 – We have emphasized that test developers need to begin with the grade level content expectations, determine the breadth, depth, and complexity of the standards and measurement tasks, and then address the knowledge forms (concepts, principles, and procedures) and skills that underlie these expectations and tasks to develop appropriate changes in testing program.

The construct underlying the standard is the beginning point in making changes in the way large-scale assessments are given or taken; test developers need to be clear of this construct. For example, a math money problem may be presented as an item on a multiple-choice test and require extensive reading skills, even though the construct is 'money estimation with arithmetic operations situated in a context.' Acceptable changes in the test administration may incorporate reading the mathematics problem or providing a calculator, as well as extending testing time. An appropriate change to develop an alternate assessment may be to modify the complexity of the problem in a manner that supports a more limited inference to the grade-level expectations (perhaps in the number or type of calculations). The degree to which this latter option is stipulated (and supported) would then need to be considered in determining what inferences can be made.

The key to the process, nevertheless, is the grade level expectations as the starting point. In this volume, the first five chapters provide the conceptual framework for taking the validity argument and applying it to various assessment approaches (from the traditional multiple-choice test to various forms of alternate assessments), and finally to specific options for participation by various groups of students (with a range of access skills). The next four chapters provide specific issues to be addressed in providing appropriate technical evidence. The final four chapters focus on implementation issues in managing an entire assessment system.

Assumption 6 – We have emphasized that all assessment options must meet professional standards and go through a validation process that includes a report of technical quality. Both procedural and empirical evidence must be documented in print form, archived, and made easily and publicly accessible.

Both procedural and empirical evidence needs to be collected to support each state's large-scale assessment program. We included several chapters in the Section II of this volume to allow test developers and assessment offices in state education agencies to focus on very specific technical features of their assessment programs. Information needs to be collected on how the tests are designed, developed, and implemented; how items and tasks are developed in line with test specifications; how reliability related evidence needs to support the dependability of our measures, and how the internal structures, response processes, and relations with other variable scan used to understand what our measures mean.

Increased use of the Internet certainly provides unique opportunities to increase the information currently available on the technical characteristics of our large-scale assessment, to better understand how they can or should be interpreted and used, and to more quickly ascertain the meaning of the constructs and the social consequences from their use. This information needs to be not only systematically collected and stored but also organized to create value-added knowledge that can inform future iterations tests and measures.

Assumption 7 – The consequences (both positive and negative) of the large-scale assessment program need to be evaluated in relation to the purpose and outcomes of participation.

Procedural and empirical evidence are not sufficient; consequences also need to be considered, particularly in terms of the interpretations we make (how we define our constructs) and their use (the social implications), whether defending current practices (Phelps, 2005) or coping with collateral damage (Thomas, 2004). Inclusion of students with disabilities in large-scale assessments is a recent phenomenon so it is difficult to make any definitive judgments about effects or impact. Clearly, the kind of position taken by Phelps has bearing: The positive effects of testing are far more supported by empirical evidence than the negative effects, though

the press is far more oriented to publishing the problems than the benefits. Nevertheless, the incidents that have been systematically collected and reported by Thomas provide a potent reminder that large-scale assessments reflect a narrow-gauge track for making critical decisions and eventually need to be supplemented by other information (not the least of which may be growth rather than static models for measurement).

Assumption 8 – Both state and local educational agencies need to provide on-going commitment to improving the quality of curriculum, instruction, and assessment systems and provide the training of professionals as they use assessment tools.

As members of the Technical Work Group, we approached the problem of including students with disabilities in large-scale testing in terms of the long haul, not as a quick fix that could be resolved and checked off so that other issues could be addressed. The interaction between curriculum, instruction, and assessment is complex and whether we take a psychometric perspective for understanding how items work by using differential item functioning, or a cognitive perspective for organizing cognition, observations, and interpretations (Pellegrino, Chudowsky, & Glaser, 2001), the process for developing credible and useful assessment information that leverages instruction requires time and experience.

Teachers and students need to participate in standards-based reforms before we can be safe in making any conclusions about our schools. The concept of "opportunity to learn" has a very limited presence in the professional literature (see Porter, 1991, 1993, 1995; Ysseldyke, Thurlow, & Shin, 1995), probably because of the difficulty in documenting it; this concept is very relevant, nevertheless. What gets tested gets taught. And in a standards-based assessment, states need to be clear about alignment, teachers need to be explicit on opportunities, and IEP teams need to be clear on decision-making.

Assumption 9 – Although a state may adopt an approach to assessment, it is the IEP team that makes the decisions about the appropriate assessment for a particular individual from among the options made available by the state educational agency.

The volume provides a complex line of logic and evidence in which state and local educational agencies must set the stage for best practice and then assist in training all educational professionals. Departments of education at the state level can no longer allow the participation of students with disabilities to be the purview of the measurement technicians alone; nor can special educators afford to manage their participation without reference to the principles of measurement. Rather, it is a shared responsibility in which measurement professionals are sensitive to the unique needs of individuals with disabilities and special educators work within known models of measurement for validating their assessments.

IEP teams must be multi-disciplinary to capture the 'wisdom of crowds' (Surowiecki, 2003), and include parents, general education teachers, special education teachers, related services, school psychologists, and administrators. In addition, university training programs must become active participants at the pre-service level of training, so that all educational professionals work in a collaborative environment and are prepared to become professionals in the field.

A Parting After Thought

In *Guns, Germs, and Steel,* Jared Diamond notes that, of the 148 big, wild, terrestrial, herbivorous mammals that could serve as candidates for domestication, only 14 passed the test. He explains this phenomenon as due to the Ana Karenina Principle: "We tend to seek easy, single-factor explanations of success. For most important things, though, success requires avoiding many separate possible causes of failure" (p. 1999, 15). We hope the information

contained in this volume allows states to avoid the many separate causes of failure in serving students with disabilities in their large-scale assessment programs.

Process and Acknowledgments

This white paper was begun on May 16, 2005, and completed on August 1, 2005, a mere 10 weeks.[4] In this time the Technical Work Group had two conference calls and three face-to-face meetings in Denver, San Antonio, and Charlotte. Four drafts were completed for review on June 1, June 18, July 5, and July 21. In developing a work plan, individual members took primary responsibility for specific chapters; however, significant portions of chapters were written by several individuals and at the very least input and feedback was solicited amongst members of the team from draft to draft.

Members of the Technical Work Group in alphabetical order:

Almond, Pat – Consultant

Browder, Diane – University of North Carolina

Crawford, Lindy – University of Colorado, Colorado Springs

Ferrara, Steve – American Institutes for Research

Haladyna, Tom – Arizona State University

Huynh Huynh – University of South Carolina

Tindal, Gerald – University of Oregon

Zigmond, Naomi – University of Pittsburgh

Specialist Writers and Internal Reviewers

Barton, Karen – CTB-McGraw

Flowers, Claudia – University of North Carolina

[4] The text of the Commission is provided in an Appendix.

Karvonen, Meagan – Western Carolina University

Wakeman, Shawnee – University of North Carolina

A review process also was instituted with the following individuals offering extremely helpful and insightful comments and recommendations.

External Chapter Reviewers in alphabetical order:

Jamal Abedi, Lizanne DeStefano**,** Russell Gersten, Carolyn Haug, Henry Kleinert, and Michael Rodriguez

White Paper Reviewers in alphabetical order: Steve Elliott and Susan Phillips

State Reviewers

Assessing Special Education Students SCASS in CCSSO, Project Leadership Team – Aran Felix (AK), Dan Farley (NM, Sharon Hall (MD), Beth Cipoletti (WV), Sandra Warren (CCSSO), Brian Touchette (DE) – as well Sue Bechard (Measured Progress) and Kelley Burling (NC)

Center Reviews

References

Apling, R. N., & Jones, N. L. (2005). *The Individuals with Disabilities Education Act (IDEA): Analysis of Changes made by P. L. 108-446.* Washington, DC: Congressional Research Service – The Library of Congress. Web site retrieved on 7/18/05:http://www.cec.sped.org/pp/docs/CRSAnalysisofNewIDEAPL108-446.pdf

Diamond, J. (1999). *Guns, Germs, and Steel*. New York: W. W. Norton and Company.

Fletcher, J. M., Denton, C. A., Fuchs, L. & Vaughn, S. R. (2004). *Multi-tiered Reading Instruction: Linking General Education and Special Education*.

Foorman, B.R., Francis, D.J., Winikates, D., Mehta, P., Schatschneider, C., & Fletcher, J. (1997). Early interventions for children with reading disabilities. *Scientific Studies of Reading*, 1, 255-276.

Ford, A., Davern, L., Schnorr, R. (July/August 2001). Learners with significant disabilities. *Remedial and Special Education*, *22*(4), 214-222.

Fuchs, L. S., & Fuchs, D. (1998). Treatment validity: A unifying concept for reconceptualizing the identification of learning disabilities. Learning Disability: Research and Practice, 13(4), 204-219.

Haladyna, T. M. (2000). *Developing and validating multiple-choice test items*. Mahwah, NJ: Lawrence Erlbaum.

Individuals with Disabilities Education Act Amendments of 1997. (1997). 20 U.S.C. 1400, et seq.

Johnstone, C. J. (2003). *Improving validity of large-scale tests: Universal design and student performance* (Technical Report 37). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved [today's date], from the World Wide Web: http://education.umn.edu/NCEO/OnlinePubs/Technical37.htm

Kane, M. (2001). *The role of policy assumptions in validating high-stakes testing programs*. Paper presented at the annual meeting of the American Educational Research Association in Seattle.

Klingner, J.K., Vaughn, S., Hughes, M.T., Schumm, J.S., & Elbaum, B. (1998). Outcomes for students with and without learning disabilities in inclusion classrooms. *Learning Disability Research and Practice*, 13, 153- 161.

Lyon, G.R., Fletcher, J.M., Fuchs, L.S., & Chhabra, V. (In press). Learning Disabilities, in E. Mash and R. Barkley (Eds.) *Treatment of childhood disorders* (2nd Ed.). New York: Guilford.

Mathes, P.G., Denton, C.A., Fletcher, J.M., Anthony, J.L., Francis, D.J., & Schatschneider, C. (in press). An evaluation of two reading interventions derived from diverse models. *Reading Research Quarterly*.

McMaster, K. L., Fuchs, D., Fuchs, L. S., & Compton, D. L. (2005). Responding to nonresponders: An experimental field trial of identification and intervention methods. *Exceptional Children*, 71(4), 445-463.

McMaster, K.L., Fuchs, D., Fuchs, L.S., & Compton, D.L. (in press). Responding to nonresponders: An experimental field trial of identification and intervention methods. *Exceptional Children*.

Messick, S. (1989). Validity. In R. Linn (Ed), *Educational Measurement*, pp. 11-102. New York: Macmillan.

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (2001). *Knowing what students know*. Washington, D.C.: National Academy Press.

Phelps, R. (2004). *Defending standardized tests*. Mahwah, NJ: Lawrence Erlbaum.

Phelps, R. (2005). *Defending standardized tests*. Mahwah, NJ: Lawrence Erlbaum.

Porter, A.C. (1991). Creating a system of school process indicators. *Educational Evaluation and Policy Analysis, 13,* 13-29.

Porter, A.C. (1993). School delivery systems. *Educational Researcher, 22*(5), 24-30.

Porter, A.C. (1995). The uses and misuses of opportunity-to-learn standards. *Educational Researcher, 24*(1), 21-27.

Rose, D. (2000). Universal design for learning. *Journal of Special Education Technology*, *15*(2), 56-60.

Salvia, J. & Ysseldyke, J. (2002). *Assessment*. Boston: Houghton Mifflin.

Sfard, A. (1998). On two metaphors for learning and the dangers of choosing just one. *Educational Researcher*, *27*, 4-13.

Speece, D. L., Case, L. P., & Malloy, D. E. (1003). Responsiveness to general education instruction as the first gate to learning disabilities identification. *Learning Disabilities Research and Practice, 18*, 147-156.

Surowiecki, J. (2004). *The wisdom of crowds*. New York: Double Day.

Thomas, M. (2004). *High stakes testing*: Collateral damage. Mahwah, NJ: Lawrence Erlbaum.

Thomas, R. M., (2005). *High stakes testing: coping with collateral damage*. Mahwah, NJ: Lawrence Erlbaum.

Thompson, S. J., Johnstone, C. J., & Thurlow, M. L. (2002). *Universal design applied to large scale assessments* (Synthesis Report 44). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved [today's date], from the World Wide Web: http://education.umn.edu/NCEO/OnlinePubs/Synthesis44.html

Thurlow, M. L., Ysseldyke, J. E., & Silverstein, B. (1993). *Testing accommodations for students with disabilities: A review of the literature* (Synthesis Report 4). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Thurlow, M., House, A., Boys, C., Scott, D., & Ysseldyke, J. (2000). *State participation and accommodation policies for students with disabilities: 1999 update* (Synthesis Report No. 33). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved [today's date], from the World Wide Web: http://education.umn.edu/NCEO/OnlinePubs/Synthesis33.html

Tindal, G., & Crawford, L. (2005). Technology applications for students with disabilities. D. Edyburn, K. Higgins, R. Boone (Eds.). *Handbook of Special Education Technology and Practice*.

Tindal, G., & Fuchs, L.S. (1999*). A summary of research on test changes: An empirical basis for defining accommodations.* Lexington, KY: University of Kentucky, Mid-South Regional Center.

Torgesen, J. K. (2000). Individual differences in response to early interventions in reading: The lingering problem of treatment resisters. *Learning Disabilities Research and Practice, 15*, 55-64.

Torgesen, J.K., Alexander, A.W., Wagner, R.K., Rashotte, C.A., Voeller, K.K.S., & Conway, T. (2001). Intensive remedial instruction for children with severe reading disabilities:

Immediate and long-term outcomes from two instructional approaches. *Journal of Learning Disabilities*, 34, 33-58.

Torgeson, J.K., Wagner, R.K., Rashotte, C.A., Rose, E., Lindamood, P., Conway, J., & Garvan, C. (1999). Preventing reading failure in young children with phonological processing disabilities: Group and individual responses to instruction. Journal of Educational Psychology, 91, 579-594.

U. S. Department of Education (2003). *24th Annual Report to Congress*. Washington, DC: Author.

Vaughn, S., Linan-Thompson, S., & Hickman, P. (2003). Response to instruction as a means of identifying students with reading/learning disabilities. *Journal of Learning Disabilities, 28,* 264-270, 290.

Warlick, K, & Olsen, K. (1998). *Who takes the alternate assessment? State criteria*.

Ysseldyke, J., Thurlow, M., & Shin, H. (1995). *Opportunity to learn standards (Policy Directions 4)*. Minneapolis, MN: National Center on Educational Outcomes.

CHAPTER 2: VALIDATING ASSESSMENTS FOR STUDENTS WITH DISABILITIES

The information in this chapter is intended to help states and school districts more validly assess the achievement of students with disabilities for the purpose of accountability. Creating valid assessments for students with disabilities can be very challenging. If test scores derived from those assessments are used for state accountability purposes, threats to validity must be identified and actions taken to remove or reduce those threats.

For example, a student with a vision impairment may have trouble reading a complex mathematics problem presented in small print on a single page. This student's performance on the math task is compromised by his/her disability. By changing the mode of presentation (e.g., large print), we remove the effect of the vision impairment and thus provide a fairer opportunity for the student to perform. Accommodations, and sometimes modifications, in the design or administration of a test may be necessary for students with disabilities because the result will be more validly interpreted and used.

The study of validity is greatly aided by following the *Standards for Educational and Psychological Testing* (American Educational Research Association--AERA, American Psychological Association—APA, and the National Council on Measurement in Education — NCME, 1999). The *Standards* book was developed by experts in testing to help test sponsors provide the highest quality testing programs possible. Chapter 10, Testing Individuals with Disabilities, is devoted exclusively to the issues faced when testing students with disabilities, provides definitions of key terms, and espouses principles and procedures to be used in testing. These testing standards are recommended for adoption by all states, school districts, and other jurisdictions that conduct achievement testing of students with disabilities.

Validity

States and school districts have a responsibility to educate students with disabilities. Part of this responsibility is to validate the interpretations and uses of test scores of these students. The process of validation is described here from an ideal perspective. However, all states and school districts have limited resources for validation. At the end of the chapter, recommendations are made for how states and school districts can and should do as part of this responsibility for validation.

What is Validity?

Validity is a way of reasoning about a desired interpretation of test scores and their subsequent use. The process followed to make valid interpretations and uses of test scores is *validation*. The process is used to answer several important questions.

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

The questions are the same as for all students although the focus of this volume is on validation of assessments designed and used with students with disabilities in a state or school district.

In the process of validation, we engage in four activities: (a) defining what students learn, (b) stating the validity argument, (c) making the claim for validity, and (d) gathering the validity evidence to support the argument and claim. At the final stage of the validation process, a qualified evaluator takes into consideration the logic of the argument and its plausibility, the claim, and the evidence in support of the claim. A summative judgment is made and, usually, recommendations are made for how the testing program could be improved to strengthen supporting evidence and eliminate threats to validity uncovered in the evidence gathering.

Chapter 10 in this volume discusses how the evidence might be presented to the public and how a technical report is used as a repository for this evidence and for conclusions about validity.

1. Defining What Students Learn

The most important and fundamental step in any testing program that permits the assessment of student learning is to define the content (AERA, et al., 1999, chapter 1). Knowing more about what students learn in school is fundamental to designing tests that help us assess student learning.

Two types of student achievement are described here. The first is very traditional and familiar to educators. The second is an outgrowth of the public's and educators' desire for more effective thinking among all citizens (Messick, 1984; Lohman, 1993; Sternberg, 1998). The first type is closely linked to the behavioral learning theory that reflects much of our teaching and learning in the twentieth century. The second type is closely linked to cognitive learning theory and social (constructivist) learning theory both of which are currently receiving more attention from educators. *N.B*. It is likely that the assessments reflected in the current standards-based environments are generally oriented to the first type of achievement. It also may well be the case that for students with disabilities, learning in the first type of achievement cannot be ignored but must form the core focus of teaching and learning in our classrooms and accountability systems.

Type One: The Domain of Knowledge and Skills. A basic education typically includes the learning of knowledge and skills. A convenient way to think about this first type of achievement is to imagine a *domain* of tasks that a student learns to perform. In mathematics, part of this domain includes adding, subtracting, multiplying, and dividing whole numbers, fractions, and decimals. Other parts of this domain are linked to student learning objectives that comprise states' content standards and are well identified by the National Council of Teachers of

Mathematics (NCTM) standards (NCTM, 2000). A test score represents a level of achievement in that domain. How many tasks can that student perform? How well can he/she perform them? We imagine levels of learning: beginning, intermediate, and advanced, that imply that students go through natural stages of learning. For purposes of accountability, we tend to use terms like the ones shown below.

Table 2.1. Academic Achievement Standards

Levels[1]
Advanced – Well above the minimal acceptable level of mastery of the material in the State's academic content standards.
Proficient – Mastery of the material in the State's academic content standards.
Basic – Progress of lower achieving students toward mastering the proficient and advanced levels of achievement.

These levels are conceived by a panel of subject-matter experts, whose experience with students enables them to make valid determinations of the cut scores that separate test scores into these levels.

Traditional criterion-referenced or domain-referenced testing that was so popular in the past featured tests designed to measure a student's status about a large domain of knowledge and skills for which every bit of knowledge and every skill had a reference to a student learning objective. We could isolate the knowledge and skills very well, and we could teach this knowledge and these skills very effectively. Each achievement test was supposed to be a representative sample of student learning from that domain. A test score would signify the level of achievement in this domain. A student with a score of 75% might lead us to conclude that the student could perform proficiently on 75% of all items in that domain.

Most state content standards contain objectives that identify knowledge and skills that fit nicely into this view of achievement. Table 2.2 shows student learning objectives for three

[1] These terms are used in the Standards and Assessment Peer Review Guidance (April 28, 2004, p. 2).

subject matters that reflect knowledge and skills. All students need this kind of foundational knowledge and skills.

Table 2.2. Objectives for Knowledge and Skills

Subject Matter	Example of a Student Learning Objective
Reading	Identify main characters in a story. Identify a cause-and-effect relationship in a story. Identify root words. Distinguish fact from opinion.
Writing	Proofread a text. Place commas correctly. Use active voice as appropriate to purpose. Spell correctly.
Mathematics	Identify examples of mathematics terms (e.g., logic, manipulative, pi, integer, scatter plot). Construct a bar graph given data. Put numbers in rank order from low to high.

For many reasons, the most desirable format for measuring knowledge and skills is multiple-choice (selected response) (Haladyna, 2004). The main reason is that the multiple-choice format provides the best chance for very good sampling from this domain. However, simply possessing knowledge and skills is not enough to function in society.

Type Two: The Domain of Complex Tasks. Students also need to learn strategies for using knowledge and skills in complex ways to solve the more complex problems they face in life. Some call this complex type of achievement *problem solving ability;* others use terms like *fluid abilities* (Lohman, 1993), *developing abilities* (Messick, 1984), and *learned abilities* (Sternberg, 1998). Regardless of the term we choose, the implication is that this kind of achievement requires having the knowledge and skills identified in the other type of achievement and ***using*** this knowledge and these skills in some complex problem-solving activity. This kind of achievement develops very slowly during the educational career of each student and is very important to functioning in life. It rarely is the target of large-scale assessments.

Generative writing is a good example of this kind of achievement developed in the elementary and secondary school years. Writing requires much knowledge and many skills. It also has a creative aspect. There are several modes of writing (e.g., persuasive, narrative, expository) and at least six analytic writing traits (e.g., fluency, organization, conventions) that are scored using rubrics—rating scales. In some scoring of writing assessments, we embed these traits into a single holistic rating scale. Writing starts early and improves throughout life. One's writing ability grows slowly and unevenly, with times where improvement is rapid and times where there is little growth. The most authentic test of writing ability is a performance test, not a multiple-choice test, even though the correlation between the two types of tests can be very high. Acquiring the knowledge and skills of writing does not make a good writer. Learning how to apply this knowledge and these skills in writing an effective narrative piece or persuasive essay is more like what we know to be writing ability.

Other examples of this second type of achievement include scientific problem solving, critical thinking, and creative thinking. Indeed, there are many examples of objectives from state content standards that identify and promote this kind of achievement. (See Table 2.3)

Table 2.3. Examples of Objectives Representing Complex Tasks

Subject	**Example**
Reading	Evaluate an instructional manual (e.g., programming a TV or DVD player). Evaluate an author's persuasive technique in a written selection. Compare characters, plots, or settings across reading selections.
Writing	Write a personal experience narrative Write a communication (memorandum, letter, invitation). Write a summary of an author's work.
Mathematics	Collect, analyze, and interpret data. Predict the likelihood of an event from theory. Solve real-world problems.

Recently, educators have been emphasizing that the domain of complex problem-solving tasks should resemble the tasks encountered in our daily living, such as budgeting and money

management, planning, problem solving, and carrying out complex projects like planning a vacation or a family reunion.

Why Is It Important to Know About and Distinguish Between These Two Types of Achievement?

Knowing about these two types of student achievement helps to ensure that all students, including those with disabilities, have opportunities to develop both the domain of knowledge and skills and the domain of tasks that reflect these important problem-solving abilities that they will use throughout their lives. Instruction for all students should not be limited to knowledge and skills at the expense of applying that knowledge and those skills to situations commonly encountered in life. Take this math problem, for example: *A teacher recently rented a car for two days. The gas tank was 3/8's full. After driving for several days, the car was disabled, but not before the gas tank was refilled at the teacher's cost. How much does the car rental agency owe the teacher, after it replaced the car?* Solving problems like this one is typical of the thousands of incidents in life that require problem-solving or critical thinking.

In requiring that all students learn grade level content, the intent is that *all* students develop these problem-solving abilities to the extent possible so that they can deal with life's exigencies. This includes students with disabilities. This intent often justifies accommodations in testing or developing alternate assessments that are better suited to a student with a severe degree of disability.

N.B. Later in this chapter, we present concepts about construct irrelevant variance and construct misrepresentation in which the tasks presented to students and subsequent scoring of performance contains extraneous factors that interfere with appropriate inferences. In many of these more complex assessments, little is known about these threats to validity (see Haladyna & Dowing, 2004). Nevertheless, to the degree that these various threats can be eliminated, Figure

2.1 shows that when students are asked to perform a complex task, inferences may be made about their ability in this domain of complex tasks, but inferences can also be made about the knowledge and skills needed to perform that task.

Thus, it is important to see that the domain of knowledge and skills supports the learning of any cognitive ability. A student will have difficulty performing a complex task without the necessary knowledge and skills afforded by prior learning. Recent research of Bob Mislevy and his colleagues' points to new methods for identifying knowledge and skills that support more complex thinking (Mislevy and Riconscentes, in press). Because current content standards include objectives that promote problem-solving abilities, assessment of this second type of achievement may be important.

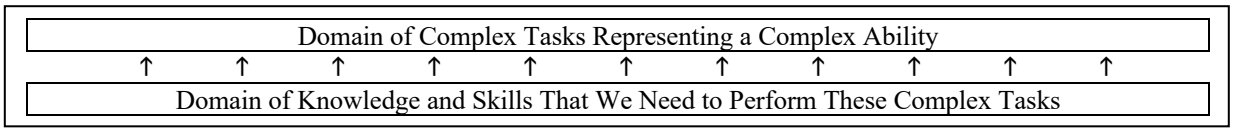

Figure 2.1. The Relationship between the Two Types of Achievement

The Validity Argument

The validity "argument" states declaratively that some tests will produce test scores that can be validly interpreted as measures of student achievement and can be validly used in a manner that is stated publicly (Kane, 1992). In making this argument, we make assumptions about the causal connections to student learning of out-of-school factors, such as intelligence and family background and in-school factors such as quality and quantity of instruction, learning environment, opportunity to learn, and school leadership to student learning. We also determine if the student's disabilities are in some way interfering with the validity of any assessment of that student's learning.

Part of the validity argument will say that testing students with disabilities can be valid if certain conditions can be satisfactorily met. Any student's disability ought not to interfere with the assessment of that student's learning. For instance, a student with mild mental retardation may have difficulty reading a mathematics test that features items from the domain of complex mathematics problem solving tasks. To administer an unaltered mathematics problem solving test to this student with low reading comprehension prevents that student from performing even if he/she can solve the mathematics problem. To address mathematical problem solving is a manner that is more suitable for a student with this type of cognitive functioning might require that we simplify the problem or its cognitive demand. In other words, an accommodation or modified assessment is developed to remove factors that blur a valid assessment of each student's learning.

Validity Claim

When the argument is made about the validity of a test for measuring student achievement, the argument is subject to analysis, evaluation, and approval by state policymakers. This may be a public encounter in which all constituencies participate. The fact of this participation is one contributing form of validity evidence (Kane, 2002). The sponsors of the tests, the state, will want to make the claim that the use of the test scores of these students with disabilities is highly valid. At the end of this validation process, we want the evidence to support the claim. Even though our validation is intended to support the argument and claim, we have a duty to seek out evidence that may not support the argument and claim. By doing this we show that validation has integrity. Also, by identifying evidence that threatens validity, we can make recommendations for changes in the testing program that will eliminate or reduce these threats. Validity research is a key to uncovering these threats to validity (Haladyna, in press).

Validity Evidence

The evidence we collect in the validation process takes several forms: procedural and statistical/empirical, and each form is discussed in this section of the chapter. Assembling a complete body of evidence can take many years. In fact, we can consider this process as unfolding in an evolutionary way over a long time. The strength of the evidence supporting the claim for validity will increase each year, unless threats to validity are uncovered and ignored. In any effective testing program, the process of validation reviews the validity argument, considers the claim for validity, and evaluates the validity evidence to make a summative judgment about validity. After making this summative judgment, a formative process leads to recommendations for improving the testing program, and, by that, increasing the validity evidence and the validity of the desired interpretation and use.

Figure 2.2 shows how the two kinds of validity evidence work to support the argument and claim but also how some evidence works in the opposite direction. Most evidence supports the argument claim, but some evidence threatens validity and detracts from the argument and claim. The goal in developing a valid assessment is to strengthen one type of evidence and minimize or reduce the other, undesirable type of evidence.

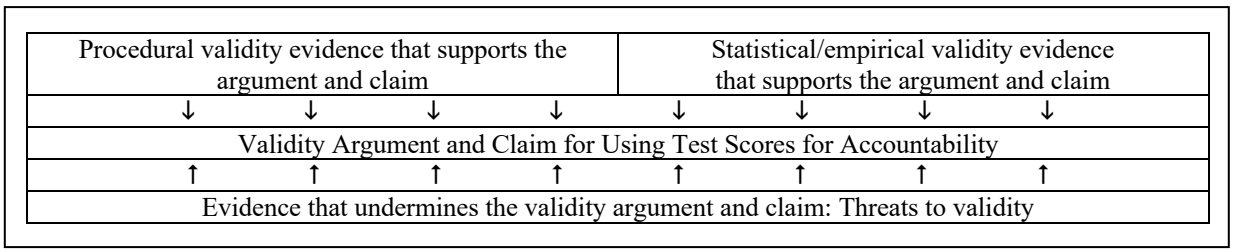

Figure 2.2. How Evidence Works For and Against the Validity Argument and Claim

We have some important distinctions to make about validity evidence that can help states, school districts, and testing companies make better decisions about what kind of validity evidence to collect and why. Decisions about validity evidence are best made by first considering the use of

the test. In this instance, the use is state, district, and school accountability. We want to infer degrees of improvement in learning for groups of students.

Procedural validity evidence. Procedural validity evidence provides documentation that the assessment test was developed in a way that is consistent with testing standards (AERA, et al., 1999). At the same time, it provides evidence that the content of the test and the administration of the test were appropriate for each student with a disability. The procedural evidence arising from following recommended testing industry procedures is a valuable source of validity evidence. Additionally, the *Handbook on Test Development* (Downing & Haladyna, in press) contains much advice about procedures to follow in test development that provide validity evidence. And Chapter 6 of this volume gives advice on validity evidence with respect to students with disabilities. Chapter 7 discusses content considerations, Chapter 8 presents ideas about item quality, and Chapter 9 covers reliability evidence.

Test scoring procedures also are important to document. Post-test activities are part of the evidence needed to support the argument and the claim for validity. The *Standards* (AERA, et al., 1999) are very clear about the need for documentation of procedures and other actions as a type of validity evidence that supports a specific test score interpretation or use. (Chapter 6 in the *Standards* is exclusively dedicated to this type of evidence.) Haladyna (2002a) discussed this need with respect to developing a technical report, which is one major source of validity evidence. But what are some of these procedures and the documentation evidence we seek?

Table 2.4 summarizes the bodies of evidence that can be used to characterize procedural evidence. Procedures and their documentation contribute greatly to the body of validity evidence supporting a state's policy on testing students with disability and the validity of interpreting and using these scores.

Table 2.4. Types of Procedural Validity Evidence

Type of Procedural Evidence	***Description and Reference***
Content-related evidence	The content of the test is well defined, using a set of content standards. The structure of the content is known, unidimensional or multidimensional. The basis for identifying content is systematic (Messick, 1989; Messick, 1995a, 1995b; Kane, in press; Webb, in press)
Item quality	Items are developed following well-established principles and procedures, and documents provide evidence of this. Items can be field tested using non-quantitative techniques (e.g. think-aloud) (Downing, in press; Downing & Haladyna, 1997; Haladyna, 2004; Welch, in press).
Reliability	Reliability of scores is estimated. If group statistics are used in an accountability formula, then the appropriate reliability index is not the same as it is for an individual score. Standard errors of measurement are important.
Scaling for comparability	When tests are modified, the resulting score should be on a scale that is validly interpretable. Procedures for achieving valid accommodations/modifications or altering the domain of tasks and the measure should be spelled out in documents.
Test design	Any test can be assembled using a variety of strategies that are based on differing statistical theories and principles. If tests are altered, the rationale for any alteration should be stated. These choices and actions should be well documented.
Test administration	Test administrator's guide How accommodations are determined Procedures for test administration Reports of irregularities
Test scoring	Scoring protocols Quality control
Standard setting	Report of a standard setting study
Reporting results	Report of the development of scores reports (see Ryan, in press)
Security	Security policies and procedures

Procedural validity evidence typically appears in written form, as a report, memorandum, letter, newsletter, brief, or file document[2]. It exists in an archive, and many states are now posting such documents on their web pages as a means of documenting their validity evidence. Keeping a well-organized archive of these documents and a record of procedures is crucial to providing an adequate body of validity evidence for supporting the claim for validity.

[2] Chapter 10 provides more detail for this process of documentation.

Empirical/statistical validity evidence. This type of validity evidence is a complementary category to procedural validity evidence. Not only must we have procedures that contribute to the body of validity evidence, but we also must provide data that support the inferences we are making about what a test score means for a group of students with disabilities. Studies of reliability of scores, the structure of test data, the quality of items, and the equivalence or comparability of different tests (including those with accommodations and modifications and alternative assessments) are crucial to forming this body of empirical validity evidence. Haladyna (in press) discusses the kinds of validity studies that can be done and their bearing on the body of validity evidence that is assembled. Chapters 6, 7, 8, and 9 in this volume provide much information about the kind of validity evidence used for analyses of data. Policy makers and others often identify problems that they think threaten their testing program's validity. Validity studies can be used to reveal the seriousness of the threat and to recommend a remedy to reduce or eliminate each threat to validity. Every state must engage in empirical studies to the greatest extent possible to assemble empirical validity evidence that complements the procedural evidence.

Table 2.5 shows categories of empirical validity evidence much like Table 2.4 shows categories of procedural evidence. The entries in this table are only suggestive and not exhaustive. Many types of empirical validity evidence exist, and most of these come naturally from standards that we find in the *Standards* (AERA, et al., 1999).

Table 2.5. Types of Empirical Validity Evidence

Types of Empirical Validity Evidence	Descriptions
Content-related evidence	Analysis of item responses to identify dimensions Correlations to like and unlike variables (e.g., formerly known as predictive validity and concurrent validity)
Item quality	Item analysis Ratings of test items based on subject-matter experts Differential item functioning
Reliability	Estimates of reliability for individuals or groups Estimates of standard errors around cut scores
Scaling for comparability	If new scales are produced for low functioning students, the validity of the extended scale should be reported. Equating results
Test Design	Procedures for test design should produce test information curves
Test Scoring	Descriptive statistics
Standard Setting	Results of a standard-setting study Impact study of a recommended cut score

As part of any validation, both procedural and empirical validity evidence should be assembled to support the argument and claim. No single source of evidence is sufficient. The "mix of evidence" is important to the evaluator called upon to make a summary judgment about validity. The next section calls for another type of validity evidence that is seldom considered in validation.

The Search for Negative Validity Evidence

As noted in previous discussion and summarized in Figure 2.3, personnel for any testing program will try to identify procedural and empirical/statistical validity evidence that support the validity argument and claim for validity. That is the main purpose of validation. However, as Cronbach (1987) observed, the claim for validity is stronger when challenges to validity have been investigated and dismissed. Kane, Crooks, and Cohen (1999) state that the chain of reasoning in validation is only as strong as its weakest link. Thus, all testing programs should have in place policies and procedures that seek negative validity evidence with the objective of either not finding it or when finding it concluding that it is immaterial. In instances, where negative validity evidence is material, the threat to validity needs to be considered and resolved.

The alternative is to invalidate the interpretation or use what we have intended, which defeats the purpose of the testing program. We know from information presented in this volume and attested to in the many references provided, that negative validity evidence is a serious problem when testing students with disabilities so added scrutiny is needed.

We have two major types of negative validity evidence: construct misrepresentation or under-representation and construct irrelevant variance.

Construct misrepresentation or under-representation. For students with disabilities, test accommodations are made to ensure students' appropriate access to the assessment. An accommodation is "a general term for any action taken in response to a determination that an individual's disability or level of English language development requires a departure from established testing protocol" (Koenig & Bachman, 2004, p. 1). When a test is accommodated or altered for a student with a disability, the most serious question is whether the result has led to a misrepresentation or under-representation of the content that was intended. If a student with a disability who requires longer test administration time is NOT given sufficient time to complete the test, then that student's score will likely be underestimated. A student with a hearing impairment may need an accommodation to the verbal instructions typically given before or during the test so that the instructions are appropriately transmitted and received. If the accommodation is not given or is inadequate, that student's performance and achievement might be underestimated. By accommodating a test or by altering the construct that the test measures, we run a risk of misrepresenting or under-representing the construct. A test that is altered to better suit a low-functioning student may alter the achievement domain being assessed.

There is no surefire statistical procedure for determining the seriousness of this threat other than professional judgment by an appropriate content specialist (e.g. reading consultant).

Both research and the sensitivity and understanding of the teacher and the rest of the IEP team are needed to guide us in the valid assessment of a student.

Systematic error (bias) or construct irrelevant variance. Any factor that is independent of the achievement domain that raises or lowers a test score unfairly is an instance of systematic error (bias). Measurement specialists use the term *construct-irrelevant variance (CIV)*. The keyword "irrelevant" indicates that a factor unrelated to the content of the test is distorting a student's score in an upward or downward direction. There are many sources of CIV (see Haladyna & Downing, 2004). These include (a) inappropriate test preparation, (b) inappropriate instruction or lack of instruction, (c) student factors, mainly of an emotional nature such as low motivation, negative attitude, or test anxiety, and (d) cheating. There are numerous reports of teachers and others tampering with test results to provide a more positive image of an educational program[3].

There are also many other factors that can contribute to CIV. In one state, a scoring error led to lower scores for a group of students who were subsequently identified as at-risk. They were assigned to a summer remedial program and only after the program was completed, was the error discovered, much to the embarrassment of the offending test company (Haladyna, 2002b).

CIV can be represented symbolically to identify the general problem and how we should deal with it for both students with disabilities and other students. The following equation suggests that every student score consists of three unique components:

[3] See the web site caveon.com for current updates on this problem.

$$y = t + er + e_{civ}$$

y is the student's actual test score.
t is a true score, if error did not exist.
e_r is random error that is associated with reliability.
e_{civ} is systematic error that is associated with an individual or a group of examinees.

Random error can be large or small, positive or negative. It cannot be calculated accurately but reliability gives us a way to estimate the error term. This estimate is called the *standard error of measurement (SEM)* and in item response theory, the SEM variable is according to the test score scale. By knowing characteristics of the test, we can estimate the approximate amount of random error (SEM) in a test score and by that have a certain degree of confidence that the score we obtain is representative to that student's status in a domain. If we are using cut scores to categorize students, this margin of error is crucial to understanding in which category a student might belong.

Systematic error is systematically large or small, positive or negative. In instances of cheating, the systematic error results in higher-than-deserved scores. With a scoring error, the source of CIV usually results in lower-than-deserved scores. The goal in the design and development of any testing program is to make that third expression in the equation attributed to CIV equal to zero. In other words, in the context of assessing the learning of students with disabilities CIV should be equal to zero or so small as to be judged immaterial (hence the term *construct-irrelevant)*.

Table 2.6 below lists some sources of CIV for the population of students with disabilities. This table provides only a few observations and is by no means exhaustive of the full range of threats to validity that exist due to CIV.

Table 2.6. Sources of Construct-Irrelevant Variance

Source	Description
Sampling	Exclusion rates vary by states; lack of uniformity misrepresents students with disabilities.
Accommodations	Accommodations offered state-by-state are not standardized, leading to differential treatment of students as a function of the state where each student resides (Haladyna & Downing, 2004).
Identification	Students with disabilities are identified in different ways and at different states giving evidence to a lack of standardization in identification of those requiring special education.
Data	Lack of standardized data makes comparisons of how students with disabilities are learning in different states.
Research on accommodation	Although research has been done, it is by no means exhaustive or conclusive (Koenig & Bachman, 2004, p. 103). The consequence of this situation is that IEP teams and policy makers cannot formulate policies and procedures that improve accommodations as well as possible with principles derived from research.
Affect of disability	A student inconsistently tracks visually from the test booklet to the scan sheet so that the student marks what is the correct answer on the wrong line. A student's handwriting is nearly indecipherable, and the student prints some words to accommodate a tendency towards reversals. Voice, composition, and conventions are excellent, but most rates cannot read the student's writing product.

CIV represents a very under-represented field of study. We have a natural inclination to not look for evidence that undermines our claim for validity. On the other hand, by not looking for CIV, we run the risk of having others expose frailties in our argument, claim, and validity evidence that undermine the entire effort to better assess students with disabilities. With the population of students with disabilities more instances of CIV are encountered than with other populations of other students. Thus, more vigilance for CIV is recommended.

Consequences

AERA (1999) issued *Guidelines for High Stakes Testing.* One of these guidelines states: *Where credible scientific evidence suggests that a given type of testing program is likely to have negative side effects, test developers and users should make a serious effort to explain these possible effects to policy makers.* Elsewhere in this volume, a chapter is

devoted to the consequences of any test score interpretation or use[4]. These consequences are subject to evaluation, and if negative should be reported, action should be taken to remove such negative consequences. For instance, under-estimating a student's reading, writing, or mathematical problem-solving ability may lead to an educational program that fails to let the student achieve what is possible. Over-estimating achievement may lead to unfair expectations and disappointment later in the instructional life of the student.

According to Koenig and Bachman (2004), giving accommodations to students with disabilities increases participation in testing programs, such as the National Assessment of Educational Progress. If this consequence is true, then the differential policies of states and school districts should have a telling effect on participation in testing programs and results of these programs as they benefit students with disabilities.

Documenting the Validation Process

In Chapter 10, the issue of documentation is presented and discussed. Several additional sources provide guidance on how this can be done (AERA, et al., 1999, chapter 6; Becker & Pomplun, in press; Haladyna, 2002a). Using the technical report, web pages to host dissemination of information, conferences, newsletters, and press releases are all effective means for documenting and disseminating evidence. All states, school districts, and other testing program sponsors are encouraged to document validity evidence and make this information available to the public and all concerned, interested parties.

[4] See Chapter 12

Establishing Assessment Validity: Summary and Recommendations

This chapter has provided an overview of how validity works to improve the assessment of all students. With the population of students with disabilities, the task of validly assessing student learning is daunting. Whether that student learning is in the domain of knowledge or skills or in the domain of complex tasks in which we prepare all students to function, validation of test score interpretations or uses is a continuing responsibility of all testing programs that purport to help students learn. Toward that end, we must build a validity argument, make a claim, and collect evidence to support the claim for validity. This ongoing process never ends, because validation is a continuing process that works to improve the testing program and the inferences that the testing program is purported to support.

It is the responsibility of the sponsor, not a testing contractor who provides test development or scoring services, to engage in validation procedures to support the use of their test scores in an accountability system for their state or school district. Although contractors may carry out many test development and scoring duties, the sponsor must assure its public that it is doing what is best for these students with respect to assessing their achievement.

Toward that end, the following recommendations are offered:

1. Create a research agenda for validity studies that seek to solve problems or uncover threats to validity for students with disabilities.
2. Document all validity evidence. The best tool for this documentation is a comprehensive technical report, but establishing and stocking an archive with documents is also important.
3. Annually, conduct a validity evaluation. Its purpose is to determine the strength of the validity evidence supporting the claim. This evaluation should recommend what actions will be needed to reduce threats to validity. This evaluation will also endeavor to strengthen the overall

testing program via recommendations for improvement. In their extensive discussion of why a testing program's tests should be evaluated, Buckendahl and Plake (in press) write: "There are public and professional interests that are served through continuous evaluation and improvement. The independent auditing of an industry that is so critical to society offers needed protection to the public. Ensuring the credibility of tests and testing programs through external verification enhances the reputation of our profession."

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.

Becker, D. F., & Pomplun, M. R. (In press). Technical reporting and documentation. In S.M. Downing and T.M. Haladyna (Eds). *Handbook of test development*. Mahwah, N.J.: Lawrence Erlbaum Associates.

Buckendahl, C., & Plake, B. (In press). Evaluating tests. In S.M. Downing & T.M. Haladyna (Eds). *Handbook of Test Development*. Mahwah, N.J.: Lawrence Erlbaum Associates.

Cronbach, L. J. (1987). Five perspectives of the validity argument. In H. Wainer & H. I. Braun (Eds.), *Test validity*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Crooks. T. J., Kane, M. T., & Cohen, A. S. (1996). Threats to valid use of assessments. *Assessment in Education,* 3(3), 265-285.

Downing and T.M. Haladyna (Eds). *Handbook of Test Development*. Mahwah, N.J.: Lawrence Erlbaum Associates.

Downing and T.M. Haladyna (Eds). *Handbook of Test Development*. Mahwah, N.J.: Lawrence Erlbaum Associates.

Haladyna, T. M. (2002a). *Essential of standardized achievement testing: Validity and accountability.* Needham Heights, MA: Allyn & Bacon.

Haladyna, T. M. (2002b). Supporting documentation: Assuring more valid test score interpretations and uses. In G. Tindal & T. M. Haladyna (Eds.) *Large-scale assessment for all students: Validity, technical adequacy, and implementation* (pp. 89-108). Mahwah, NJ: Lawrence Erlbaum Associates.

Haladyna, T. M. (2004). *Developing and validating multiple-choice test items* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice*. 23(1): 17-27.

Haladyna, T.M. (In press). Roles and importance of validity studies in test development. In S.M. Downing and T.M. Haladyna (Eds). *Handbook of Test Development*, pp. Mahwah, N.J.: Lawrence Erlbaum Associates.

Kane, M. T. (1982). A sampling model for validity. *Applied Psychological Measurement, 6*, 125-160.

Kane, M. (1992). An argument-based approach to validation. *Psychological Bulletin, 112,* 527-535.

Kane, M. T. (2002). Validating high-stakes testing programs. *Educational Measurement: Issues and Practice, 21*(1), 31-41.

Kane, M. T. (In press). Content-related validity evidence in test development. In S. M. Downing and T. M. Haladyna (Eds.) *Handbook on test developmen*t. Mahwah, NJ: Lawrence Erlbaum Associates.

Kane, M. T., Crooks T.J., & Cohen, A.S., (1999). Validating measures of performance. *Educational Measurement: Issues and Practice, 18,* 2, 5-17.

Koenig, J. A., & Bachman, L. F. (Eds.). (2004). *Keeping score for all: The effects of inclusion and accommodations policies of large-scale educational assessments*. Washington: The National Academies Press.

Linn, R. L. (In press.) The *Standards for Educational and Psychological Testing*: Guidance in test development. In S. M. Downing & T. M. Haladyna (Eds.) *Handbook on test development*. Mahwah, NJ: Lawrence Erlbaum Associates.

Lohman, D. F. (1993). Teaching and testing to develop fluid abilities. *Educational Researcher, 22*, 12-23.

Mehrens, W. (1998). Consequences of assessment: What is evidence? *Education Policy Analysis Archives, 6*(13). http://epaa.asu.edu/epaa/v6n13.html.

Messick, S. (1984). The psychology of educational measurement. *Journal of Educational Measurement*, *21*, 215B237.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13B104). New York: American Council on Education and Macmillan.

Messick, S. (1995a). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*, 741-749.

Messick, S. (1995b). Standards of validity and the validity of standards in performance assessment. *Educational Measurement: Issues and Practice, 14*(4), 5-8.

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

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (Eds.) (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press.

Sternberg, R. J. (1998). Abilities are forms of developing expertise. *Educational Researcher, 27*(3), 11-20

U. S. Department of Education (2004). *Standards and assessment peer review guidance: Information and examples for meeting the requirements of the No Child Left Behind Act of 2001*. Washington, DC: Author.

Welch, C. (In press). Item/prompt development in performance testing. In S.M. Downing and T.M. Haladyna (Eds). *Handbook of test development*. Mahwah, N.J.: Lawrence Erlbaum Associates.

CHAPTER 3: APPROACHES TO ASSESSING STUDENTS IN STATEWIDE ASSESSMENT PROGRAMS

In the previous chapter we articulated a model for designing a comprehensive assessment system built on a validity argument. This chapter has two goals: to consider assessment design and technical quality issues and to describe and evaluate a range of approaches to standards-based assessment in terms of their appropriateness for students with disabilities. We consider both issues in the context of the different options for participation: (a) general assessment, (b) general assessment with accommodations, (c) alternate assessments judged against grade level standards, (d) alternate assessments judged against modified achievement standards, and (d) alternate assessments judged against alternate achievement standards.

We begin by describing the relationship between these options for participation and six general approaches to assessment that support a comprehensive assessment system and provide all students with access to the general education curriculum. Then we provide details on the six assessment approaches. Details include administration, scoring procedures, interpretation of results, and examples of each approach. Finally, we evaluate the six assessment approaches in terms of their practical feasibility, technical adequacy, and the inferences supported by each approach in relation to what students know and can do.

Participation Options and Assessment Approaches

The most prevalent assessment approaches used in large-scale testing systems include the typical pencil/paper multiple choice and constructed response assessments for the general assessment judged against grade level achievement standards and teacher judgments using rating scales and checklists, portfolios, performance events, or collections of performance tasks for the alternate assessments judged against grade level, modified, or alternate achievement standards.

Any assessment approach possibly could be applied at any level of participation, but some relationships are more common to statewide assessment systems than others. Students who participate in assessments that focus on grade level achievement standards —because their instructional program focuses on grade level content and achievement standards —must be able to participate in typical statewide assessments under standard administration conditions or with test administration accommodations or in an alternate assessment that is comparable. Other alternate assessments are for students whose instructional program still focuses on grade level content, but whose performance must be judged against modified or alternate achievement standards. These assessments support different inferences about what students know and can do because the content and achievement standards are modified and because the assessment conditions are modified.

"Typical" or General Statewide Assessment

It often is risky and sometimes foolhardy to define what is typical. However, based on sources such as the Annual Survey of State Student Assessment Programs (Council of Chief State School Officers, 2003), "typical" state assessment programs in grades 3-8 and at the high school level contain multiple choice items and some number of short and extended constructed response items[1].

Multiple-choice items are the most widely used selected-response (SR) approach to assessment (Popham, 2000) but have been less frequently implemented in alternate assessment options. Constructed response (CR) items also are found in statewide general assessments. They require students to produce an answer as opposed to merely identify a correct option. Constructed responses also appear as items in many alternate assessment approaches.

[1] See *Standards for Educational and Psychological Testing* for definitions

Students with disabilities may participate in the general statewide assessment with or without accommodations. Research on test administration accommodations that is beginning to emerge addresses crucial questions about the (a) effectiveness of specific accommodations in reducing the impediment to optimal performance posed by various disabilities, and (b) impact of the accommodation on the comparability of inferences about student performance and achievement under accommodated and standard test administration conditions for students with and without disabilities. However, results from much of the research thus far are mixed or inconclusive. For example, Sireci and colleagues concluded from their comprehensive review of accommodations studies that "the benefits of oral accommodation remain unclear. This finding could be because only specific subgroups of SWD need this type of accommodation and administering it to a larger group of SWD obscures its effects" (Sireci, Li, & Scarpati, 2003, p. 64).

Teacher Judgment Using Rating Scales and Checklists[2]

Rarely used, if at all, as a general assessment participation option, many states include this approach in their alternate assessment systems. Teachers or members of IEP teams complete checklists or rating scales evaluating a student's processes (e.g., behaviors, required levels of assistance, levels of self-determination) or products (e.g., work samples, completed performances, test scores). Checklists are dichotomous, indicating whether a student has exhibited a behavior or completed a product. Rating scales consist of multiple levels and require teachers to evaluate the level of accuracy and/or independence a student typically exhibits.

Rating scales and checklists completed *in* "real-time" rely on direct observations of a student as he or she performs a particular task at any point in time and in this way are in direct

[2] The NCEO website is a useful resource on alternate assessment approaches being used in various states. See http://education.umn.edu/nceo/TopicAreas/AlternateAssessments/StatesAltAssess.htm.

contrast to checklists or rating scales completed post-hoc. Or, in other words, "real-time" completions of checklists and rating skills most often evaluate process whereas post-hoc scales most often evaluate products. Finally, "real-time" observations may occur in natural settings or be completed within a structured setting at a set point in time. Natural observations may result in more valid information, but structured observations usually take less time and provide more specific information related to the content standard in question.

Rating scales also are used in most portfolio and performance event assessments in which rubrics are developed to evaluate the achievement of students. In this application, the rating scale has two components: (a) a definition of a dimension (e.g., independence, generalization, or quality) and (b) various levels of proficiency, each with their own descriptors arranged in an ordinal scale (usually ranging from 1 to 5 or 7). These rating scales are then used to evaluate products (work samples or performance).

Assessment Portfolios

Portfolios contain a collection of student work but may also include descriptions of student performances, observational data, and other information required by the state. Some states provide explicit detail and direction to IEP teams and teachers related to portfolio requirements or the content standards the portfolio needs to address. In other states, IEP teams and teachers are given more flexibility as to the contents of the portfolio, depending on each child's unique educational goals. In these instances, teachers or team members explicate the relationship between portfolio contents and grade level academic content standards.

Performance Events

A *performance event* implies a comprehensive process or product from which inferences can be made about specific sub-components of knowledge and skills; various rubrics are then

used to evaluate the event's sub-components. A performance task differs from an event in the discreteness or specificity of the task; therefore, a collection of multiple performance-based tasks is needed to make accurate inferences about a student's knowledge or skills in a particular domain. Assessing student achievement using performance events represents direct assessments of student skills, as opposed to indirect assessments such as observational checklists.

When used with students with significant disabilities, performance events often are scored across three behaviors: (a) the level of support the student needs to respond, (b) the student's ability to generalize the skills, and (c) the accuracy of responses. Furthermore, these performances are not limited to paper and pencil tasks; for example, students might use a storyboard to sequence story events to demonstrate text comprehension.

Performance Tasks with Summative Rating Scales

A performance task differs from a performance event in its more limited coverage of content standards and the smaller number of prompts or items in the task. A collection of multiple performance tasks is needed to create a score scale and to enable accurate inferences about a student's knowledge or skills in a particular domain. Performance tasks with summative rating scales represent direct assessments of students' skills. One example of this type of system is the Colorado Student Assessment Program Alternate (CSAPA: http://www.cde.state.co.us/cdesped/StuDis-Sub2.asp.). In this system, students are assessed on a collection of performance indicators aligned with the state's grade level content standards.

Assessing Progress on IEP Goals

Assessing progress on IEP goals is not an acceptable approach to alternate assessment.[3] Nevertheless, it has been reported in use. This approach is like other portfolio assessment approaches in that collections of evidence are scored using rubrics. It is distinguishable in that the collections contain evidence of products, performances, activities, and interactions that are linked to content standards but are direct illustrations of IEP goals. As with teacher-directed portfolio assessments, measuring a student's performance on state standards within the context of *individualized* goals may not fully sample the content domain.

In the seven figures that follow we provide further details about each assessment approach including administration formats, scoring procedures, interpretation of results, and examples (adapted from Ferrara & Anderson, 2004). At the bottom of each figure is a reference to a state or states using this approach.

[3] Discussion in the Federal Register of 12/9/03, 34 CFR Part 200, Title I--Improving the Academic Achievement of the Disadvantaged, appears to rule out assessing progress on IEP goals as an acceptable approach to alternate assessment.

"Typical" GENERAL EDUCATION Grade Level Assessments

Typical state content area assessments include multiple choice items and often include short construct response items and extended constructed response items. Items are linked to specific state content standards and indicators.

Administration

Tests are administered to groups of students, except when small-group and individual accommodations are provided. Testing sessions typically are scheduled on specific days and times by schools or school systems, except where accommodations are provided. Students respond on scannable answer documents or directly in test booklets. Administration sessions typically are completed under flexible time allotments or untimed.

Scoring Procedures

Student responses to multiple choice items are scanned by machine and scored by computer. Responses to constructed response items are scored by trained readers.

Interpretation of Test Performance

Total test scores and subscale scores may support norm-referenced interpretations (e.g., percentiles) or criterion-referenced interpretations (e.g., student's performance in relation to an achievement level such as "Proficient"). The latter interpretation is required for NCLB to document Adequate Yearly Progress. Inferences about what students know and can do are generalized to the content standards that are targeted by the test.

Examples

Maryland School Assessment, which are augmented norm-referenced tests; Ohio diagnostic and achievement assessments for grades K-8; Ohio (High School) Graduation Tests.

Figure 3.1 Approaches in Statewide Assessment Systems: General Assessment

Alternate Assessments Judged Against Grade Level Achievement Standards

The intention for such assessments appears to be that content standards, achievement standards, and psychometric rigor are sufficiently similar to support inferences about what students know and can do that are comparable to inferences from the general assessment. Current comparable assessment procedures typically are judgmental procedures.

Administration

School staff submit, on behalf of a student, evidence to support the contention that the student has achieved state content and achievement standards comparable to those of the state content area assessment.

Review Procedures

Evidence of student achievement is reviewed by an authorized panel. Procedural rigor of panel review and decision process may vary.

Interpretation of Test Performance

It is very difficult to achieve comparability when substantial changes have been made in the types of evidence for judging proficiency. Panel judgments about submitted evidence are not likely to support comparable interpretations about what students know and can do. Thus, inferences about what students know and can do in relation to grade level content and achievement standards are not likely to be comparable unless a functional perspective is used to view the value of performance in context.

Examples

Oregon's juried assessment; New Jersey's Special Review Assessment.

Figure 3.2 Approaches in Statewide Assessment Systems: Alternate Assessment Judged Against Grade Level Achievement Standards

ALTERNATE ASSESSMENTS –
Rating Scales and Observational Checklists

Rating scales include a list of behaviors, work samples, and so forth and one or more dimensions on which to rate the quality and/or accuracy of students' performance. Checklists include lists of observable behaviors (e.g., skills), types of work, quality of work, or appropriateness of behaviors which are checked off as present or not present. To be consistent with NCLB, rating scales and checklists must focus on state academic content standards but may also include life skills, or other instructional objectives.

Administration and Scoring Procedures

Special education or other teacher rates student performance in relation to state content standards or instructional goals as completed/successful vs. not completed/unsuccessful or on a multi-level rating scale (e.g., poor/adequate/successful). Evidence that is rated may include teacher observations of students while performing tasks, student work samples, and other sources. Rating scales or rating notes may also indicate extent of teacher cuing and support (e.g., hand-over-hand assistance, verbal encouragement) and nature of student responses (e.g., pointing, spoken responses).

Checklists may include lists of observable behaviors (e.g., skills), types of work, quality of work, or appropriateness of behaviors. Behaviors and work must be linked to grade level content standards and may also focus on life skills. Observer checks option to indicate presence or absence of the behavior or work. Items in the checklist may include extent of teacher cuing and support (e.g., hand-over-hand assistance, verbal encouragement) and nature of student responses (e.g., pointing, spoken responses).

Inferences About What Students Know and Can Do

Student performance and progress is evaluated in relation to the identified observable behaviors and work in situational contexts; information is provided about both the situation and the performance, allowing inferences to be made about several dimensions of behavior, stability, frequency, and generalizability.

Examples

Arizona, Indiana, Louisiana, Mississippi, New Mexico, and Wisconsin have alternate assessments using rating scales (based on state responses to a pre-release version of a 2005 survey by the National Center for Education Outcomes).

Figure 3.3 Approaches in Statewide Assessment Systems: Alternate Assessments – Rating Scales and Observational Checklists

Alternate Assessments – Portfolios

Assessment portfolios are "a systematic collection of educational or work products that have been compiled…according to a specific set of principles." (From the *Standards for Educational and Psychological Testing*, 1999, p. 179).

Administration

Over a period, teachers gather evidence of student work and performance of tasks. Evidence may include data charts, performance checklists, work samples, videotapes, audiotapes, observations, and interviews. Teachers follow guidelines for selecting evidence for inclusion in the portfolio that is directly relevant to grade level content standards and other instructional goals. The evidence in a portfolio is unique to that student and different to varying degrees from the evidence in portfolios of other students.

Scoring Procedures

Usually, trained raters score the entries in a portfolio on several dimensions using scoring rubrics and anchor papers. Scoring dimensions may include student progress, connection to the standards, level of independence, generalization of skills, multiple settings, opportunity for choice-making, and self-evaluation. All evidence may be scored once per dimension (i.e., holistically) or multiple pieces of evidence may be scored separately on each dimension.

Interpretation of Evidence of Student Achievement in the Portfolio

Typically, guidelines for portfolio contents ensure that evidence is aligned with specified content standards. Often, dimension scores are calculated by summing and weighting scores from several scoring dimensions. Scores are intended to indicate achievement of the student but may also reflect the skill of the teacher in selecting appropriate evidence of student achievement. Comparability of inferences from portfolio scores for different students is influenced by the quality and comprehensiveness of sampling of content standards, similarities of pieces of evidence, and number of pieces of evidence in the portfolio.

Examples

Ohio Alternate Assessment for Students with Disabilities; South Carolina's PACT-Alt assessment portfolio.

Figure 3.4 Approaches in Statewide Assessment Systems: Alternate Assessments – Portfolios

Alternate Assessments – Performance Events

Performance events represent a single performance task that may contain several prompts (e.g., as many as 15 questions or directions to the student that elicit a response, like a test item) that may be arranged in increasing order of difficulty or cognitive complexity.

Administration

Teachers administer. Administration may occur over several sessions. These larger performance tasks are generally similar to extended constructed-response tasks (e.g., writing prompts). Students respond to as many of the prompts in the task as they are able.

Scoring Procedures

Using a rubric, teachers may score responses to each prompt as the students produce them or they may be scored in a separate scoring project by trained raters. Scoring rubrics usually focus on the accuracy and/or appropriateness of the response.

Interpretation of Test Performance

These larger performance tasks may assess several content standards or indicators (e.g., as many as 15). The total score across all prompts in the task may reflect the quality and accuracy of the responses to the prompts, level of independence in responding, or other response features. Since all students respond to the same task, inferences about what students know and can do are comparable. Inferences to the broad domain of grade level content standards may be limited. Locally dependent responses across prompts may limit score reliability.

Examples

South Carolina's HSAP-Alt assessment.

Figure 3.5 Approaches in Statewide Assessment Systems: Alternate Assessments – Performance Events

Alternate Assessments – Performance Tasks with Summative Score Scales

Assessments with performance tasks require students to construct a spoken or written response, create a product, or perform a demonstration. Often, they contain several tasks, where each task is a coherent collection of items or prompts that elicit responses from the student. The tasks generally do not elicit a single correct answer or solution but allow for a wider range of responses. Evaluation of student responses, products, and performances are based on judgments guided by scoring criteria. (Adapted from McTighe & Ferrara, 1998, p. 34.) Total test scores are based on a sum of score on items and tasks.

Tests like the Brigance IED II and Woodcock-Johnson Psycho-Educational Battery (Rev.) illustrate some of these features.

Administration

Teachers administer a series of performance tasks to an individual student following scripted directions.

Scoring Procedures

Teachers score student responses to each item in a task or responses to an entire task as they complete administration of each item or task. Student responses may also be scored on level of independence exhibited when completing each individual task.

Interpretation of Test Performance

Since several tasks can be administered to students, several grade level content standards can be assessed. Student performance, as represented by a performance level that is based on the summed total score, can be viewed as an estimate of the student's achievement in relation to the full range of content standards targeted by the test.

Examples

Oregon's Extended Reading, Writing, and Mathematics Assessments; Colorado's Student Assessment Program Alternate (CSAPA); the new South Carolina alternate assessment.

Figure 3.6 Approaches in Statewide Assessment Systems: Alternate Assessments- Performance Tasks with Summative Score Scales

Alternate Assessment of Progress Toward IEP Goals
Student progress toward achieving IEP goals is assessed using collections of evidence based on products, performances, activities, and interactions that directly illustrate specified IEP goals. **Administration** Teachers or other IEP team members record information about learning and performance goals from the IEP and student progress toward the IEP goals. Additional information may include details such as content area or skills domain, level of complexity of the goal, and amount and types of support to be provided to the student. **Scoring Procedures** Collections of evidence are scored by trained raters or by teachers and other IEP members. Rubrics or rating scales are generally used for scoring. **Interpretation of Performance** IEP goals included for assessment must be related to grade level content standards, but may also include life skills, and other areas. Student performance and progress is evaluated in relation to IEP goals, some of which should be relevant to content standards. **Examples of IEP Approach** In their analysis of alternate assessments based on performance in relation to IEP goals, NCEO (2005, to be released) found that four states (Alabama, Georgia, Texas, and Virginia) used this approach in the 2004-05 school year. However, discussion in the Federal Register of 12/9/03, 34 CFR Part 200, Title I--Improving the Academic Achievement of the Disadvantaged, appears to rule out assessing progress on IEP goals as an acceptable approach to alternate assessment.

Figure 3.6 Approaches in Statewide Assessment Systems: Alternate Assessments – Progress Toward IEP Goals

Evaluating Assessment Approaches

As discussed in earlier chapters, validity evidence for an assessment system should indicate the degree to which intended inferences from test scores are supportable. If the primary goal of an assessment system is to make valid inferences about what students know and can do in relation to grade level content standards, then a foundation must be established to support these inferences. This foundation should consist of procedural as well as empirical evidence to include: (a) practical feasibility of the assessment approach, (b) technical adequacy of the assessment instruments and/or procedures, and (c) plausibility of inferences made when interpreting and using test results. In the remainder of this chapter, we define these three foundational features and use them to evaluate the strengths and weaknesses of the six different assessment approaches.

Evidence in Three Areas

In evaluating ***practical feasibility*** of an assessment approach we consider fiscal costs, time burden on students and teachers, and training requirements. We also discuss public acceptability.

The requirements for ***technical adequacy*** of an assessment approach and the relative importance of those requirements are different for the various assessment approaches. In this section we also consider evidence for the soundness of scoring procedures (e.g., rater agreement, rater accuracy) and indicators of psychometric soundness (e.g., score reliability, correlations with other measures, appropriateness of scaling and equating).[4]

[4] We emphasize the important differences among ***rater agreement*** (i.e., percentage of times two raters assign exact or adjacent scores to student responses), ***rater accuracy*** (e.g., percentage of times that all raters assign scores that are in exact or adjacent agreement with the "true scores" assigned to student responses), and ***score reliability*** (e.g., internal consistency, test-retest, and score dependability G and D coefficients).

Evidence for the ***plausibility of inferences*** from test scores underlines the intended interpretations and uses of an assessment approach rather than the technical adequacy of the approach. Here we consider alignment of the demands of the items, prompts, and procedures that comprise an approach with the content standards targeted by the approach; evidence that inferences about what students know and can do are related to the achievement standards and descriptions of those standards; and evidence of the generalizability of those inferences from the assessment situation to classroom learning situations and other situations.

Practical Feasibility

Cost. The costs for an assessment approach may be incurred as part of development, which may happen only once or on an on-going basis; training of teachers to administer, score, and interpret results from the assessment; scoring student responses; and score reporting. Although costs per student for state content area assessments may be considerable, the per-student costs for portfolios, performance events, and performance task collections tend to be higher unless classroom teachers score student responses as part of the administration and data collection process. Rating scales and observational checklists and assessing progress on IEP goals are the least expensive approaches to assessing students with disabilities once initial development costs are covered. These observations about costs do not account for the costs of staff time. Schools and school systems may find it necessary to pay for additional staff to complete work that teachers would have completed if they were not required to prepare for and administer assessments.

Time burden. Time burdens are a function of test design, administration, scoring, and score dissemination requirements. Administration time burdens on teachers and students may have the most impact on the public acceptability of any assessment approach. Time plays an

important role in most assessment approaches used with students with disabilities because they require individual rather than group administrations. Rating scales and checklists may require less time to administer, and teachers can respond quickly to items; consequently, it may be possible to assemble many more items than is possible with other approaches. However, if these scales are extensive or need to be completed during direct observation of student behavior, they can create a scheduling/timing burden.

Both performance events and performance task collections usually require a one-to-one administration, often with a focus on the accuracy of a response and the level of independence displayed by each student during each response. Many of these approaches require multiple prompts, further increasing the amount of administration time required.

The time required to assemble and present a portfolio collection varies across states and is largely dependent on both the number and intensity of state requirements for collecting evidence for inclusion in the portfolios. Time commitments also vary in response to the level of professional development activities and support states can provide to those responsible for collecting portfolio evidence. In unstructured interviews conducted in Maryland, Ohio, South Carolina, and Virginia, teachers reported that they spend 10 hours or more per student to complete portfolio data collection and assembly.

Training. All approaches to assessment require a great deal of training; some alternate assessment approaches, however, are critically dependent on thorough training. Some even require teachers to rehearse the administration and scoring process (e.g., performance events and performance tasks). For example, an extensive amount of time is needed to train someone on how to select, administer, score, collate, and submit a collection of products for a portfolio assessment. Furthermore, the depth of training needed across different assessment approaches

varies considerably, often according to the degree of professional judgment required by the approach. Rating scales and checklists require accuracy of administration, but they have definite parameters around what to do and not to do. Portfolio assessments, on the other hand, often are highly dependent on the evaluator's (teacher's) skills and knowledge across a range of activities. Regardless of assessment approach, quality of training is, at the least, indirectly linked to the quality and accuracy of test score interpretations and resulting inferences and therefore critically important when considering practical feasibility and acceptability of approaches.

Recommendations for Improving Practical Feasibility

Administration time is reduced when the assessment is completed with a group of students at the same time as opposed to a one-on-one approach. Providing accommodations during a "typical" grade-level assessment for a group of students is one way to minimize administration time. For example, extended time or oral directions can be simultaneously provided to a small group of students. Portfolio collections also increase efficiency of "administration time" as multiple students can be working on portfolio products simultaneously.

Other recommendations for decreasing the burden of administration time on teachers and students include: (a) providing teachers with standardized directions clearly written with specific examples, (b) including items that strategically sample the domain of interest, being careful not to over-sample nor under-represent the targeted content standards, (c) employing methodologies that require observation of structured activities reducing the amount of time teachers have to observe any one student, and (d) considering the number and intensity of state requirements for the collection of evidence in portfolio-based approaches.

Financial costs may be lessened across numerous approaches if the following recommendations are followed: (a) be forward-thinking in the development and design of the

initial assessment so that major structural changes are not needed across the years and costs are reduced to creation of alternate forms, and/or generating a test bank of items or a set of comparable prompts, and (b) train teachers to administer and score assessments instead of outsourcing the work. In this final example, initial investment in training will be high but can be lessened through use of a "trainer of trainer" model and by retaining trained teachers within the district from year to year.

Training on administration and scoring takes time and may be expensive, regardless of approach. Recommendations for improving practical feasibility of training include: (a) providing thorough training for "new" teachers or assessment administrators and brief annual review sessions for experienced teachers, (b) requiring teachers to videotape an administration and submit for review (Colorado has asked teachers to participate in a similar program for two years, allowing an opportunity to monitor treatment fidelity and collect inter-rater reliability data), (c) pair an inexperienced teacher with a local peer who can provide ongoing training and support as well as conduct direct observations of the administration, and (d) provide numerous practice opportunities during training so that teachers become fluent with administration procedures.

Technical Adequacy of Instruments and Procedures

As a starting point, we must evaluate the trustworthiness of the evidence provided by different approaches to assessment. Trustworthy evidence from an assessment speaks to the reliability and validity of inferences made about those scores. In this section we discuss general features related to technical adequacy and evaluate various assessment approaches according to these features. Important to this section is an understanding that technical adequacy is evaluated in degrees–it is not an all-or-nothing phenomenon–and in relation to the inferences that are intended.

Rater consistency and accuracy. Assessment approaches must enable consistent scores, which are indicators of student performance and achievement, regardless of differences across students, settings, and time at which an approach is administered. Rater dependability and accuracy apply to professionally trained scorers who may score portfolios in a scoring project or classroom teachers who may score student responses during administration of a performance event or collection of performance tasks. Rater accuracy refers to the frequency with which scorers assign scores that are consistent with "true scores." But consistency is not enough. Scorers may score consistently but inaccurately (e.g., think of an archer consistently hitting the lower right section of a target.) Scorers also must score accurately, that is, assign scores to student responses that are reliably within the expectations and features described in scoring rubrics. (Think of an archer hitting the bull's eye of a target.) Both consistency and accuracy (i.e., consistently hitting the bull's eye) are required to produce dependable scores.

Some challenges to consistency and accuracy are more readily apparent in the context of alternate assessments than others. One can see how alternate assessments relying on direct observations of student behavior in natural settings may not provide the rater with enough opportunities to observe students in multiple settings or enough times to provide a valid estimation of students' true skills and abilities. Furthermore, raters are most often special education teachers, and their ratings may be affected by a lack of opportunities to work with students in general education. These special education teachers may have a limited understanding of an age-appropriate reference group that impedes the accuracy of scores assigned (resulting most often in an inflation of scores). Portfolio approaches provide additional examples of possible threats to the dependability of scores and score interpretation because they are heavily influenced by human judgment and the likelihood of scorer bias.

Internal consistency score reliability. All assessment approaches must produce reliable results if they are to support valid inferences about what students know and can do. The reliability of assessment results is indicated in a few ways, including widely known and used reliability coefficients and more complex generalizability analyses.[5]

Traditionally, the internal consistency of an assessment has been evaluated through use of internal consistency reliability coefficients, which are easy to calculate when an assessment approach consists of a fixed set of test items, prompts, or tasks. The interpretation is that scores from the assessment are reliable or dependable because they are based on items, prompts, and tasks that produced similar observations about student performance and achievement. The overall design of performance events and collections of tasks resemble typical content area assessments and can demonstrate adequately high internal consistency reliability. Rating scales and observational checklists are likely to be so diverse in terms of the content standards to which they align that internal consistency reliability would have to be evaluated in individual content focus strata (if adequate numbers of items per stratum are available). Because internal consistency coefficients are based on the relationship among a collection of items, prompts, or tasks, collections of relatively unstandardized evidence in assessment portfolios cannot yield internal consistency reliability coefficients (as discussed in Chapter 6).

Other evidence of score reliability. Some assessments enable and even require evidence of the reliability or replicability of scores. For example, assessments using performance events and tasks may exist in multiple forms that are administered at different times of year or in alternating years. The correlation of scores between these alternate forms—referred to as

[5] References to the generalizability of assessment scores over samples of items, persons, and observational conditions in the context of score reliability should not be confused with the generalizability of inferences about what students know and can do from assessment situations to other contexts such as classroom learning situations and other situations.

alternate forms reliability – provides evidence of the replicability of scores over parallel versions (i.e., different sets of item prompts, and tasks) of the same assessment. In fact, the replicability of scores from different collections of evidence on the same students – a way of thinking about score reliability for portfolios—is conceptually similar to the notion of alternate forms reliability.

The importance of examining test-retest score reliability is commonly understood yet test-retest reliability is only rarely implemented. Correlations between scores for the same assessments administered on different occasions provide this form of evidence of the replicability of scores over occasions. Unlike alternate forms reliability evidence, test-retest evidence could be collected for any approach to assessing students with disabilities (i.e., disregarding the limitations to this approach posed by practice and other confounding effects).

Score generalizability analyses. Generalizability analyses allow examination of the separate effects of different sources of measurement error on the dependability of scores from an assessment. Typically, the sources of error that are investigated are samples of items (which is conceptually similar to test-retest reliability), persons (which is conceptually similar to investigating rater agreement and accuracy), and observational conditions (e.g., different settings and occasions, which is partially similar conceptually to test-retest reliability). Generalizability analyses, and decision studies which focus on specific sources of measurement error, can be designed for all approaches to assessing students with disabilities.

Convergent and discriminant validity evidence. One way to strengthen the reliability of an assessment approach is to ensure that items or tasks within the assessment are related to each other when they are meant to represent a similar construct (convergent evidence) and not related to each other when they are not meant to assess the same construct (discriminant evidence). Furthermore, evidence supporting the trustworthiness of the approach is strengthened when a

collection (domain) of items or tasks or performances contribute independent information toward a total test score.

Evidence that is convergent (i.e., strong correlations of items intended to measure similar constructs) or discriminant (i.e., weak correlations of items intended to measure different constructs) partially indicates the construct validity of an assessment – that is, the degree to which scores from an assessment can be taken as evidence of what students know and can do across various domains. It is crucial to seek this type of evidence and similarly, crucial to seek disconfirming evidence as well. Content experts can provide this type of evidence when evaluating possible overlap across individual domains in a test before it is distributed, and statisticians can provide correlation evidence once test score data are collected.

Equivalence of parallel forms of an assessment. Assessments that are developed as multiple forms are intended to yield equivalent inferences about what students know and can do. Multiple (i.e., alternate) forms of typical content area assessments are made equivalent through the development process and subsequent scaling and equating processes. Scaling and equating processes also can be used for multiple forms of performance events (e.g., the alternate assessment for the South Carolina high school exit exam, HSAP-Alt) and collections of performance tasks. Though it probably is not widely practiced, multiple forms of a rating scale or observational checklist should be equated through development and statistical processes.

Recommendations for Improving Technical Adequacy

One recommendation for improving the consistency and accuracy of scoring assessments requiring human judgment is to provide adequate training to ensure intra- and inter- rater reliability thus minimizing the effects of rater bias (e.g., the tendency to assign high scores, even when unwarranted; the tendency to assign higher or lower scores to particular types of evidence)

and rater drift (i.e., shifts over time and occasions in consistency, accuracy and severity). Specifically, raters can practice scoring multiple products containing similar strengths and weaknesses (or corrects and errors) and compare the consistency of their own scoring (*intra-rater*) across these products. Similar methods would be used for achieving *inter-rater* reliability with plenty of opportunity for raters to discuss their reasoning for score assignment including their interpretation of the meaning of rubric descriptors, scale indicators, etc.

Accuracy of scoring may be improved by requiring that all teachers and assessment administrators provide documentation of ratings assigned to different performance events or portfolio collections and randomly sampling this documentation to conduct a large-scale evaluation of scoring accuracy. Another recommendation is to require teachers to provide a justification of a percentage of their scores to evaluate whether scoring procedures are clearly understood and followed.

Technical adequacy, in terms of content-related validity evidence, of approaches relying on teachers to collect a sample of student products or performances (i.e., portfolio approaches and some performance events) can be strengthened by requiring documentation of how each product, collection of evidence, or performance event links to the academic content standard it is intended to sample. Once content alignment is assured, content experts can provide evidence that is convergent or discriminant through their evaluation of the amount of overlap across unique domains assessed before the assessment is distributed. Further evidence can then be analyzed for convergent and discriminant relationships through statistical correlations and factor analyses once data are collected.

Plausibility of Inferences

A critical question is whether our inferences are warranted or plausible. More specifically, the question is, *what kind of inferences about student knowledge can we defend with the data provided by a specific assessment approach?* Some approaches allow us to make broader inferences about a students' grade-level content knowledge than other approaches. Inferences also gain plausibility when threats to validity are systematically acknowledged and addressed. Some threats can be addressed during the design phase while others appear later in the process. Some assessment approaches are more prone to certain threats to validity than other approaches. As discussed in Chapter 2, however, a threat in and of itself does not deem an approach unacceptable but instead adds to the evidence (or in this case, lack of evidence) for the validity argument. In this section we discuss what it means to increase the plausibility of inferences from assessment scores as well as provide examples that demonstrate how threats to validity can undermine the plausibility of inferences.

Reliability, dependability, and replicability. In the previous section score reliability was discussed in terms of the design, administration, and scoring of various assessment approaches. In this section reliability concepts are discussed in terms of their necessity in making accurate inferences about a particular student's knowledge and skills. One might ask, *would we get similar scores and draw similar conclusions about what students know and can do had we chosen a different collection of products in our portfolio assessment?* Similar questions can be posed for all approaches to assessing students with disabilities and, in fact, for all students.

Alignment with content standards. All assessment approaches in a comprehensive assessment system must align with grade-level content standards. And it is the strength of this alignment in terms of the depth, breadth, and complexity of an assessment that provides a critical

piece of validity evidence. In the following paragraphs we provide examples and non-examples of assessment approaches having varying degrees of alignment with the general education content standards in terms of depth, breadth, and complexity.

Typical statewide content area assessments are designed to capture the depth, breadth and complexity of students' knowledge and skills related to grade level academic content standards. Use of multiple-choice items provides test users with information about the breadth of students' understanding. Their strength lies in the number of items that can be sampled and the amount of content coverage they entail. And, although multiple choice items have traditionally been constructed in a manner that has not yielded information about the depth of a student's knowledge, current methodologies are being developed to provide this type of information, as well (see, for example, the work of Barton & Huynh, 2003, on error analysis and design of test item distracters).

Depth and complexity of understanding are usually assessed through constructed response items. As previously mentioned, typical statewide content area assessments generally include both multiple choice and constructed response types of items. Also, because students completing typical content area assessments are generally able to work with abstract representations and/or hypothetical situations (usually presented within complicated text materials), items can be designed to measure more than one content standard. Therefore, the information provided by these "typical" assessments allows test users to make a few warranted inferences about a student's knowledge and skills in the broader grade level content domain.

An assessment approach at the other end of the continuum, for example, a portfolio approach, may be successful in capturing the depth and complexity of student knowledge and performance but this strength comes at a cost in terms of limited breadth of content coverage. In

other words, when using grade-level academic content standards as the construct of interest, portfolios suffer from construct under-representation; they are limited by number of products (e.g., data points). More products may lead to broader inferences about the content domain but in doing so may decrease the practical feasibility of the approach.

Examining alignment of assessment items, prompts, and of tasks to content standards is crucial to supporting inferences from assessment scores. Alignment studies provide some of the evidence necessary for judging the relevance of an assessment to the content standards that the assessment targets. Similarly, assessment information must be relevant to the achievement standards (i.e., cut scores) established for an assessment and the performance level descriptions (PLDs) used to establish and define the achievement standards. Judgmental alignment procedures that focus on content standards could focus on the achievement standards and PLDs, as well.

Generalization of inferences to other situations. We use the term generalization to describe how well information garnered from a student's performance on a large-scale assessment generalizes to the larger set of standards and to classroom instructional situations and other situations. The generalization of inferences for any academic achievement assessment is limited to some degree. And gathering relevant evidence is challenging and costly, so generalizability studies for even typical content area assessments are not often undertaken. That does not diminish the importance of realizing that the types of conclusions we draw about student performance and achievement are limited primarily to a unique assessment situation (e.g., tasks and supports that might be provided as evidence in a portfolio), to structured and supported instructional situations (e.g., as with rating scales and observational checklists based

solely on classroom situations), and to other standardized assessment situations (e.g., content areas assessments, performance events, and collections of performance tasks).

Recommendations for Increasing Plausibility of Inferences

Dependability, replicability, and alignment pose unique challenges for all alternate assessments, but particularly portfolio-based approaches. Recommendations for addressing these challenges include thorough teacher training (as previously discussed); invalidating portfolio entries that do not align with the appropriate content standards; specifying the specific standards to assess; giving examples of tasks that sample these standards; and providing examples of how to sample multiple standards with one product or performance.

Recommendations for improving the dependability, replicability, and alignment of other approaches (including multiple choice tests, performance task collections, and rating scales), include statistical analysis of form equivalence within and across years, evaluation of content match between items and academic content standards through creation of test specifications and other procedures, and computing reliability coefficients in order to demonstrate internal consistency of assessments.

Conclusions

In this chapter we reviewed current approaches to assessing students with disabilities in large-scale assessment programs. We also evaluated the approaches in terms of their ability to support valid inferences about a student's knowledge and skills. Considering this discussion, we draw one primary conclusion: At this point in their development, no single assessment approach can capture the breadth, depth, and complexity of knowledge of a curriculum in its entirety while maintaining a high level of technical adequacy and retaining a reasonable level of practical feasibility.

We offer the following general conclusions and recommendations:

1. All assessment approaches, including all alternate assessment approaches, intended for students with disabilities must meet the same standards for feasibility, technical adequacy, and plausibility of inferences, as do other statewide content area assessments.

2. Most students with disabilities should participate appropriately in the general ("typical") statewide content area assessments with or without test administration accommodations.

3. Providing the various options for participation of students in statewide assessment systems – that is, both the general assessment with or without accommodations and alternate assessments – must be supported by evidence of technical adequacy, and plausible inferences about what students know and can do.

4. Each of the approaches to alternate assessment has advantages and drawbacks in meeting standards for feasibility, technical adequacy and plausibility of inferences.

These conclusions, thus, highlight the importance of framing a validity argument supporting the assessment approach(es) used. The validity argument would begin with the delineation of: (a) a clear assessment purpose for the assessment, (b) stakeholders' intended score use, (c) desired inferences to be made with collected data, and (d) possible unintended consequences for students, teachers, and others. The argument will then be challenged, on multiple fronts, once the approach is implemented. Issues concerning practical feasibility will be raised by teachers and state personnel, data collected may or may not support the assessment's technical adequacy, and results may or may not support intended inferences and may or may not contribute to positive consequences.

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.

Barton, K. & Huynh, H. (2003). Patterns of errors made by students with disabilities on a reading test with oral reading administration. *Educational and Psychological Measurement, 63(*4), 602-614.

Colorado Department of Education (2003). *What you should know about the Colorado Student Assessment Program Alternate (CSAPA).* Retrieved March 1, 2005, from http://www.cde.state.co.us/cdesped/StuDis-Sub2.asp

Council of Chief State School Officers. (2003). *Annual survey of state student assessment programs: 2001-2002.* Washington D.C.: Author.

Ferrara, S., & Anderson, C. (2004 August). *Evolving taxonomy of approaches to alternate assessment. Internal working document*. Unpublished manuscript, American Institutes for Research, Washington, DC.

Johnson, E., & Arnold, N. (2004). Validating an alternate assessment. *Remedial and Special Education, 25,* 266-275.

La Marca, P. M., Redfield, D., Winter, P. C., Bailey, A., & Hansche, L. (2000). *State Standards and State Assessment Systems: A Guide to Alignment.* Washington, DC: Council of Chief State School Officers.

McTighe, J., & Ferrara, S. (1998). *Assessing learning in the classroom.* Washington, DC: National Education Association.

National Center on Educational Outcomes. (2005). Unpublished data.

Popham, W. J. (2000). *Modern educational measurement: Practical guidelines for educational leaders* (3rd ed.). Boston: Allyn and Bacon.

Sireci, S. G., Li, S., and Scarpatii, S. (2003). The Effects of Test Accommodation on Test Performance: A Review of the Literature. *Center for Educational Assessment Research Report*, 485, 0-0.

Thompson, S. & Thurlow, M. (2003). 2003 *State special education outcomes: Marching on.* Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved January 18, 2005, from the World Wide Web: http://education.umn.edu/NCEO/OnlinePubs/2003StateReport.htm

CHAPTER 4: A DECISION FRAMEWORK FOR IEP TEAMS RELATED TO INDIVIDUAL STUDENT PARTICIPATION IN STATE ACCOUNTABILITY ASSESSMENTS

The purpose of this chapter is to help Individualized Education Program (IEP) teams consider the most suitable way for an individual student with disabilities to participate in the annual statewide assessments. The IEP team currently has five options to consider[1]:

1. participation in the general assessment,
2. participation in the general assessment with accommodations,
3. participation in an alternate assessment judged against grade level achievement;
4. participation in an alternate assessment judged against modified achievement standards; and
5. participation in an alternate assessment judged against alternate achievement standards.

In Table 1.1 (Chapter 1) we listed the five assessment options and the similarities and differences among them. Each of the five options must produce achievement scores that can be used in calculating Adequate Yearly Progress (AYP) as required under the No Child Left Behind (NCLB) Act. Each option is derived from federal regulations and policies; at the present time, *states can decide whether to use modified and/ or alternate achievement standards in judging the performance of students with disabilities. N.B. As we have indicated earlier, no regulations have been developed for the alternate assessment judged against modified achievement standards.*

All the assessment options are based on the same content standards; there are no differences in the academic content standards used for the content of all five options. In the first three assessment options, student performance is judged against grade level achievement standards. These grade level achievement standards are designed to enable inferences about the breadth and depth of content proficiency for the grade level. Both the general assessment with

[1] See Table 1.1 in Chapter 1

and without accommodations and the alternate assessment judged against grade level standards would allow for comparable inferences to be made. In the fourth and fifth options, student performance is judged against modified or alternate achievement standards, respectively. With both modified and alternate achievement standards, there is an inference that accommodations and supports have been used, particularly those involving assistive technologies, prompting, or scaffolding[2] and/or that the grade level content has been changed in breadth, depth, and/or complexity. Modified achievement standards are designed to enable inferences about grade level expectations somewhat constrained in breadth, depth, and/or complexity. Alternate achievement standards are designed to enable very stipulated inferences about grade level expectations that have been extensively prioritized and narrowed and assume student performance is contingent on having the supports used in the assessment.

While the requirement that *all* students participate in these assessments may raise multiple issues for IEP teams to address, this chapter focuses only on those most closely related to the participation options. Specifically, we discuss the learning and behavioral characteristics of students likely to be appropriate for each testing option. IEP teams also might consider graduation requirements based on the option chosen, but because these differ by state, we do not address graduation implications. For alternate assessments judged against alternate achievement standards, states must specify the impact on meeting graduation requirements for each option in their state-level testing guidelines. Of course, the default option for all students with or without disabilities is the general assessment without accommodations unless the IEP team determines otherwise.

[2] Scaffolding is an instructional (and assessment) technique whereby the teacher models the desired learning strategy or task response, then gradually shifts responsibility to the students; in some cases, the teacher uses what is correct in the student's response but probes or cues the student, so as to lead the student to a more complete and accurate response (Clay & Cazden, 1992) (Also see Glossary for further definitions)

Clarifying the IEP Team's Role in Decision Making

DeStefano, Shriner, and Lloyd (2001) found that participation rates and accommodation patterns did not appear to be based on access to the general curriculum nor on the nature of instructional accommodations prior to training in a decision model. New federal regulations and additional assessment options have emerged since this research was carried out making the demands on IEP team decision making even more complex. Before describing the guidelines for selecting an appropriate assessment option for an individual student, it is critical to be clear about *inappropriate* criteria that might be applied in making the decision.

The IEP Team Can Not Decide to "Exempt" a Student from Participation in Statewide Assessment

The IEP team cannot decide ***if*** a student will participate in statewide assessments. Since the 1997 Amendments of IDEA, all students with disabilities must be included in state and district assessments. Those unable to participate with accommodations must be provided with an alternate assessment. Some states may specify certain conditions under which any parent may refuse to permit their child's participation; these exemptions apply to students without disabilities, as well as students with disabilities. For example, in Pennsylvania, the only grounds for parental refusal are religious ones[3]. Beyond that, for an IEP team to exempt a student from state assessments is inconsistent with federal law.

The IEP team must decide ***how*** a student with disabilities will take the statewide assessment. With increased options for the participation of students with disabilities in state

[3] In Pennsylvania, students may be excused from the assessment if (and only if) their parents refuse participation and that refusal is possible only after the parents have reviewed the test content and declared it to be inappropriate, on religious grounds.

testing[4], these decisions are complex. They also have important implications for school accountability, reporting, and graduation rates. IEP teams develop "individualized" education programs that consider the effects of the disability on the needs of the individual student. These programs include the recommendation for participation in statewide assessment. IEP teams may need guidance and training in recommending the most appropriate form of participation in statewide assessment for each individual child with a disability.

The IEP Team Should Not Select an Assessment Option Based on a Student's Participation in a Separate, Specialized Curriculum

Although test scores serve many purposes, we focus on the use of statewide assessment data for accountability (i.e., to estimate students' achievement of the state's academic content standards). All accountability assessments, including alternate assessments, must be linked to the states' academic content standards (U. S. Department of Education, 2003). To decide that students will participate in an alternate assessment judged against alternate achievement standards because they need a "functional" or other specialized curriculum is inconsistent with this purpose.

The 1997 Amendments of IDEA specified that all students must have access to the general curriculum. This access was reaffirmed in the 2004 reauthorization of IDEA. Additionally, NCLB (2002) requires that all students be assessed in reading and math for accountability purposes. An important principle of assessment is that students can learn the material on which they will be tested. English and Steffy (2001) call this the "doctrine of no surprises." Because all students will be assessed on grade level content standards, instruction for all students with disabilities should be aligned with grade level content in reading and math

[4] States are not required to use every assessment option. Some states may have alternate assessments but not choose to use alternate or modified achievement standards. When fewer options are used, state guidelines should clarify which options are available.

(albeit reduced in breadth, depth, and/or complexity, for some students). The IEP team may decide that certain students with disabilities need additional instruction in daily living and functional skills that do not link to academic content standards. Or they may decide that a particular student with a disability needs remedial work in reading or math, social skills instruction, learning strategies, or other unique curricula. Specification of these additional curricular needs is an appropriate role for the IEP team and specialized curricula may be prescribed to augment participation in general curriculum. These augmentative curricula do not, however, imply any assessment model. For example, augmentative work in life skills may be appropriate for students participating in the general assessment or for students participating in an alternate assessment measured against alternate achievement standards.

No Child Left Behind sets the high expectation that *all* students will achieve state standards in reading, math, and science. This expectation has substantially increased the educational (and specifically curricular) opportunities for all students with disabilities. Evidence exists that even students with the most significant cognitive disabilities can learn academic content, although continued research is needed to demonstrate how to teach the full breadth of the general curriculum (Browder, 2005; Browder, Wakeman, Spooner, Ahlgrim-Delzell, & Algozzine, 2005). Alternate assessments with strong links to grade appropriate content standards use academic tasks and responses, although functional activities also may be incorporated for the application of academic content (Browder, et al., 2003.) Alternate achievement standards and modified achievement standards may be applied to a small sample of the population, but the content of the assessment is still linked to the academic curriculum and the academic content standards for the grade level.

The IEP Team Does Not Select the Assessment Option Based on Current Placement

The decision about the most appropriate type of assessment for students with disabilities is also *not* a *placement* decision. As noted in the prior paragraphs, every student has the right to access to the general curriculum. Students also have the right to instruction from a highly qualified teacher who is trained in the content area, and the right to an appropriate education in the least restrictive environment (LRE) (IDEA, 2004). LRE does not define the way a student participates in a statewide assessment. There are students with disabilities in every type of educational setting who could be recommended for any of the assessment options. For example, a student with learning disabilities attending a separate day school might be participating in the regular state assessment without accommodations. Or a student with significant cognitive disabilities who is fully included in a public-school general education classroom might be participating in an alternate assessment judged against alternate achievement standards.

The IEP team determines placement for special education services and in doing so, they have implied that the student will receive access to the general curriculum in this setting from a highly qualified teacher. The decision about place does not constitute a decision about how a student will participate in the state assessment; this requires a separate decision based on the student's learning characteristics and educational needs.

The IEP Team Does Not Select the Assessment Option Based on Disability Classification

The type of assessment chosen for students is *not* based on the student's disability classification used to determine eligibility for special education services. The terms "significant cognitive disabilities" or "persistent academic difficulties" do not define new categories of students with disabilities. It is not appropriate to determine whether a student's performance will be judged against grade level, modified, or alternate achievement standards based on a disability

label. Instead, the IEP team assessment decision should be based on the types and intensity of support the student needs to show academic learning during ongoing instruction. For example, a student who is visually impaired may use specialized orientation strategies to access instructional materials and large print with high contrast to show what he/she knows and can do. The IEP team may decide to recommend an accommodated general education assessment in large print with high contrast for this student. Another student may be recently blinded and not yet reading Braille. The IEP team might recommend an alternate assessment that allows this student to be read to[5] and to respond orally; the student's performance would still be judged against grade level achievement standards. In each case, the decision is based on the individual student's needs.

The IEP Team Does Not Select the Assessment Option to Improve School AYP Reports

IEP teams may perceive public or school pressure to select an assessment option for a student that will most likely contribute to improvements in the school's accountability scores. Selecting an option for participation in statewide assessments, like all educational decisions made by the IEP team, focuses on creating an appropriate education for each individual student with disabilities. Although there are caps on the number of students who can be reported as proficient using modified or alternate achievement standards, there are no limits on the number of students who can be assigned to these alternate assessments. Thus, the IEP team should have the flexibility to select the assessment option that is best for each student. Similarly, if a student's needs are best met by participating in an assessment judged against grade level achievement standards, even if scoring proficient is a "long shot" for this student, the IEP team should recommend this option. By using evidence-based procedures and practices to teach the general

[5] Reading the reading test to a student changes the construct being measured from reading comprehension to listening comprehension; in most states, this is not an acceptable accommodation on the general statewide assessment

education curriculum to all students, more students will achieve proficiency in whichever assessment they are assigned to take, and school scores will improve.

The IEP Team Should Select the Assessment Option Based on Educational Needs

The most appropriate criteria to consider in selecting the assessment option for a student with disabilities are his/her current educational needs. The IEP team identifies the types of supports and interventions the student requires for educational success. They have information on how the student has participated in various types of assessments in the past. From this information, they address the specific educational needs that relate to statewide assessment and formulate a recommendation.

Considering Educational Needs to Select an Assessment Option

To arrive at a recommendation regarding participation in statewide assessment, the IEP team should consider several questions: In what way does this student access the general curriculum? What has been this student's response to academic interventions? How does this student interact with text? Do the supports required by this student to perform or participate meaningfully and productively in the general education curriculum change the complexity or cognitive demand of the material? What inferences can be made about how the student will generalize skills to different contexts (i.e., transfer information taught in one context to other)? In the next few pages, we discuss how the answers to these questions could lead an IEP team to a particular assessment recommendation.

Question 1. In what way does the student access the general curriculum?

Some students with disabilities access the general curriculum in the same way as students who are nondisabled; they are included in general education classes and/or are expected to master the general curriculum to the same level of depth and complexity as their nondisabled

peers, although they may need some accommodations to do so. Some of these students may have achieved grade level reading and mathematics skills but also have significant physical challenges that require extensive accommodations. Other students may have deficient reading and/or mathematics skills, but with learning strategies, behavior support systems, or other types of support that do not change the scope or format of the general curriculum, they can demonstrate success. An IEP team would recommend that students who are working towards grade level achievement in their ongoing instruction take the general assessment with or without accommodations. IEP teams will choose this option for most students with disabilities. In fact, in a recent summary of participation rates across states for the 2002-2003 school year, NCEO reports that 84% of special education students participated in the reading statewide assessment by taking the general assessment with or without accommodations, and 76% of special education students participated in the math statewide assessment by taking the general assessment with or without accommodations (Thurlow, Moen, & Wiley, 2005).

If a student has been taking the general assessment, but has not achieved proficiency, it is important to consider the type and quality of instruction the student has been receiving before recommending an alternate assessment option that applies modified or alternate achievement standards. Because of past traditions in special education of using separate, specialized curricula, the student may not yet have had access to the general curriculum for his or her assigned grade level. For example, an eighth-grade student with learning disabilities, reading on a second-grade level, might report to the Special Education resource room during English period for instruction on a specialized remedial reading curriculum and may not have had the opportunity to be exposed to the standard curriculum that is addressed in the general assessment. Or the student may have had the opportunity to go to eighth grade English in addition to receiving specialized

remedial help, but the remedial program may not have been based on evidence-based practice. Or the student may have received instruction from a teacher not highly qualified to teach eighth grade English. Before changing from the default assessment recommendation, the IEP team should consider these three questions:

- Has the student received instruction in the grade level academic content?
- Was this instruction evidence-based?
- Was the instruction delivered by a highly qualified teacher?

If the answer is "No" to any of the three questions, the IEP team should recommend an alteration in the student's *instructional* program before considering an assessment option based on modified or alternate achievement standards.

Some students have access to the general curriculum but still need to focus on priorities within the content because they cannot learn the full depth, breadth, and/or complexity of the grade level content. The important consideration here is how much of the grade level content the student is expected to learn. Some students may be able to absorb the full breadth of material presented, but not with much depth. Because their reading is painfully slow and laborious, they may need to focus on chapter summaries rather than the full chapter while they are concurrently learning to read. They may be able to recall facts and main ideas but not many supporting details. They may be able to access text with technology but be unable to decode it. They may participate in the curriculum using compensatory strategies like using a calculator rather than learning computation. On assessments of academic subject matter content like science, accommodations may be sufficient for these students to demonstrate competence and proficiency. However, even with appropriate, intensive, remedial instruction these students cannot improve sufficiently in one year in reading and/or mathematics to demonstrate grade-level proficiency on grade-level

reading and/or mathematics tasks. Expectations for what can be accomplished in certain subject areas in one year have to be substantially "streamlined" for these students. Because of this, an IEP team may consider them appropriate candidates for the statewide assessments judged against modified achievement standards.

Students with the most significant cognitive disabilities may need even more streamlining or prioritization within the general curriculum. The National Alternate Assessment Center has summarized some of the variables that relate to cognition for this population (Kleinert, et al., 2005 available at http://www.naacpartners.org), including challenges in short term memory and transfer of knowledge that students with significant cognitive disabilities to need ongoing instruction on specific target skills and frequent opportunities to learn new skills. Access to the general curriculum often involves ongoing team planning about how the student will participate in each instructional unit or weekly lesson.

Access also may require what might be called "dual instruction." The student with significant cognitive disabilities may be working on emergent literacy or numeracy skills in the context of grade level content. For example, a fifth-grade student with significant cognitive disabilities may be learning to identify pictures, using stories adapted from fifth grade books. Or the student may be learning the concept of number but concurrently gaining awareness of how numbers are used in simple equations. An extensive level of prioritization within the general curriculum may suggest to the IEP team the need for an alternate assessment judged against alternate achievement standards.

Question 2. What has been the student's response to academic interventions?

Many students with disabilities respond to appropriate, intensive interventions aimed at improving performance on academic, behavioral, or social skills. Their progress can be reliably

detected, measured, and reported. Once they "catch up" they can keep pace with the grade level curriculum and perform proficiently on the statewide assessments. In a summary of participation and performance rates on statewide assessments administered in the 2002-2003 school year, NCEO reports that the average proficiency rate among special education students taking the general assessment (taken without or with accommodations) was 21% for reading and 12% for math (Thurlow, et al., 2005).

Other students with disabilities, however, appear to have much more difficulty making progress academically, even if they have had access to a multi-tiered intervention system based on evidence-based practices (Fletcher, Denton, Fuchs & Vaughn, in press). These students demonstrate academic difficulties. They begin the school year with significant gaps in achievement and in the time available in a single school year, they cannot close the gap and master the entire curriculum. Before deciding on an assessment option for such students, the IEP team should consider several interrelated questions:

- How long has the student received academic instruction in grade level content?
- What is the student's age?

An IEP team must be certain that the student has had sufficient opportunity to receive intensive, consistent, and explicit academic instruction of sufficient duration to achieve results before considering an assessment option that substantially reduces expectations for student achievement. Further, the IEP team must be certain that the student has been exposed to grade level content before assuming the student cannot achieve grade level expectations. Younger students, students who have demonstrated even small amounts of progress in academic skill acquisition, and students who have had less than 2 years'

worth of opportunities to receive appropriate, evidence-based instruction should continue to be recommended for the general assessment taken with or without accommodations.

- What is the student's preference about continuing to try for grade level achievement?
- What are the parents' preferences about the student continuing to try for grade level achievement?

Because selecting an assessment option other than the general assessment with or without accommodations may have implications for graduation, for post-secondary school attendance, and for future employment, it may be important to consider the student's and his/her parents' preferences before making the assessment recommendation. For some, achieving proficiency of modified achievement standards may not be the most desirable outcome.

Students with significant cognitive disabilities can make academic progress with ongoing, systematic instruction[6] on prioritized skills (Browder and Spooner, In Press). To borrow a term used in functional skills instruction, the target goal is "meaningful partial participation" rather than mastery of the entire content or construct. Some states have developed curricular frameworks or statements of the critical essence of their state standards to give teachers examples of how this participation can be meaningful and aligned with the grade level content. The National Alternate Assessment Center also provides resources on making these grade level links with prioritized skills and focused on higher expectations for academic learning than assumed in the past (http://www.naacpartners.org). For students whose academic progress has been to learn the critical essence of the academic content, alternate assessments judged against alternate achievement standards may be appropriate.

[6] The term "systematic instruction" is used here to refer to the prompt fading procedures like time delay, least intrusive prompts, and similar methods used to promote skill acquisition for this population.

Question 3. How does this student interact with text?

A student who can learn with text and instructional supports that do not change the complexity or cognitive demand of the general curriculum should have the opportunity to demonstrate grade level achievement. Students who interact with textbooks and other general curriculum materials that do not change the cognitive demand of the task are most likely to be candidates for assessments judged against grade level achievement standards. Accommodations may be needed (e.g., Braille; enlarged text; extended time to read, etc.) and these should also be available for the general assessment. For this student it would be appropriate for the IEP team to recommend participation in the general assessment with accommodations.

In contrast, students with disabilities who need texts and instructional supports that *do* alter the complexity and cognitive demand of the grade level content may need to be assigned to an assessment that permits judging their proficiency against modified achievement standards. These students can engage with the grade level content only if the text is modified to a substantially lower reading level or if the reading requirement is reduced in other ways. For example, they may be able to understand concepts in biology using text summaries and simplifications. Students with poor reading skills may be able to demonstrate comprehension of content (albeit auditory comprehension, not reading comprehension) if technology or a human reader helps them to access the text, but not to the extent that they can simply absorb unedited chapters of the grade level textbooks read aloud. That is, they need simplified content as well as support for decoding. These students are probably candidates for an assessment judged against modified achievement standards.

Students with significant cognitive disabilities may have inconsistent or rudimentary literacy skills. They may be gaining word awareness while using symbols to fill in reports and

other student assignments. They may be able to understand a concept in biology if the text is not only simplified and made accessible with technology or a human reader but also supplemented with extensive picture or auditory cues. They may use a minimal sight word vocabulary to glean meaning from phrases, headlines, and other signs.

Unfortunately, to date, few researchers and schools have focused on how to teach students with significant cognitive disabilities to read (Kliewer & Biklen, 2001; Joseph & Seery, 2004). In the absence of an adequate research base to define evidence-based practice, descriptive resources provide some illustrative adaptations suitable for this group of students. For example, Koppenhaver, Erickson, and Skotko (2001) used adapted texts and communication to promote participation in storybook reading for students with Rett Syndrome. Similarly, Downing (2005) made text accessible using a tactile alphabet book. Several researchers have taught picture reading using assistive technology (Mechling & Langone, 2000; Musselwhite & King-DeBaun, 1997; Snyder, Freeman-Lorentz, & McLaughlin, 1993). And Musselwhite and King-DeBaun (1997) embedded pictures or auditory cues so students could keep pace with the reading of a story; they used assistive technology to "say" repeated phrases (e.g., to read a repeated story line). Students who access text primarily by focusing on key words, pictures, and auditory cues may be candidates for alternate assessments judged against alternate achievement standards.

Question 4. Do the supports required by this student change the complexity or cognitive demand of the material?

Some students with complex physical or sensory challenges can perform on grade level if given alternative ways to demonstrate learning. One student may have significant physical challenges and communicate in a way that requires extensive time and possibly a translator to convey the message. Another student may have serious medical challenges that require providing

assessments using unique responses and contexts. If students with these types of challenges have been able to access text and other instruction in the general curriculum in ways that do not change its complexity or cognitive demand, they may be candidates for either the general assessment with accommodations or an alternate assessment judged against grade level achievement standards.

For other students, the supports needed modify the breadth depth and/or complexity of the content. For example, the student may be able to produce the answer only with instructional support like scaffolding. Or the student may use strategies and devices to compensate for not having the complex thinking or memorization skills required by the task. Some students may be able to find the answer in an "open book" context but not be able to retrieve the information in a standard assessment. Supports that give the student a portion of the answer obviously reduce the cognitive demand of the content being assessed and imply the need for modified achievement standards.[7]

Students with significant cognitive disabilities may lack the symbolic communication skills typically used to engage in academic instruction and thus need extensive support for academic learning. Browder, Ahlgrim-Delzell, Courtade-Little, and Snell (2005) have described three levels of symbol use that have relevance for access to the general curriculum. Some students do not yet use symbolic communication, but instead make their needs known through gestures, sounds, or other means. To access the general curriculum, these students need support to use new symbols while learning academic concepts. For example, a student may be able to use a symbol for a story character on a voice output communication device to participate in the reading of a story. Some students with significant cognitive disabilities may use and comprehend

[7] Supports that require no thinking skills or give the full answer (e.g., color coding of the correct answers) are not appropriate for *any* students as they do not promote or assess learning.

a small vocabulary of spoken words or symbols. These students have more options to begin applying symbols in the context of grade level content. By using assistive technology, the student may be able to fill in simple charts using picture symbols or use numbers. Students with significant cognitive disabilities also may have more expanded symbol systems and speech, but still need extensive supports (pictures, visual cues, objects) to apply these skills in the context of grade level curriculum.

In addition to assistive technology to compensate for underdeveloped symbolic communication, some students with significant cognitive disabilities may need social supports and instruction in self-directed learning. Some may benefit from learning with peer models who are nondisabled (Fisher & Frey, 2001; Hunt, Staub, Alwell, & Goetz, 1994.) For example, the student may be more responsive to a peer's prompting than a teacher's, or more likely to imitate a peer's model than a teacher's model. Others may be more responsive if allowed to choose materials or activities (Kennedy & Haring, 1993; Stephenson & Linfoot, 1995). Some may be able to show partial learning (Shriner, 2000; Thurlow, Lazarus, Thompson, & Robey, 2002). Students with significant cognitive disabilities who need extensive assistive technology support for symbol use and who need environment supports such as peer models or teacher scaffolding are likely candidates for an alternate assessment through which they can demonstrate achievement using these supports and have their performance judged against alternate achievement standards.

An important caveat is to remember that all students need the opportunity to think through assessment tasks and make a response. Just as color-coding the answers in a test booklet is not an appropriate form of support, physically guiding a student with significant cognitive disabilities through every task provides no information on whether the student has learned the

material. This type of prompting may be important to early learning but should be faded so that students can demonstrate independence. While students with significant cognitive disabilities may need extensive supports to participate in assessment tasks, these supports should not compete with, or prevent the student from being able to show, some level of independence and problem solving.

Question 5. What inferences can be made about the students' generalization/ transfer of learning?

Generalization or transfer of learning always exists within some boundaries. What differs across students with disabilities is the scope of these boundaries. Some students with disabilities can demonstrate transfer of learning with little or no direct instruction on generalization *per se*. Most students with disabilities, however, have some difficulty generalizing learned information to novel situations. Students may master content material in one educational setting but fail to apply that information to another general education class or real-life setting (Scruggs & Mastropieri, 1984). For example, students who may have learned basic division in school may not automatically apply this skill to dividing the cost of a pizza among four friends. While such generalization problems also are encountered in students without disabilities they can be magnified in students with disabilities. However, many of these same students respond well to generalization training that helps them apply what they have learned to different life situations. When these students participate in assessments judged against grade level achievement standards, it is assumed that the assessment items sample their broader grade level achievement. For example, if students can solve a particular word problem in a math assessment, it is assumed that they would be able to apply that same math competence to other scenarios (real life or textbook) that require similar levels of mathematical reasoning and computation.

It would be riskier to make this inference for some students with disabilities who may only be able to do the math problem if it is one they have done before or if the problem on the math test is worded exactly the same way as the practice problem on which they were instructed. It also would be risky to make this inference for some students with disabilities who can only solve the problem using a step-by-step written or aural guide. Students for whom the inference about generalization is more restricted than that assumed for the grade level may need the opportunity to show proficiency judged against modified achievement standards.

For students with significant cognitive disabilities, the safest assumption is that they can demonstrate achievement, if assessed under the same conditions and with the same supports that were used in instruction. These students often need systematic instruction to demonstrate generalization of a concept. For example, Horner's research on general case instruction demonstrated that teaching students with multiple exemplars that sampled the range of variation for a task (not just those being taught) could lead to performance of untaught tasks (Horner & Albin, 1988; Horner & McDonald, 1982.) When achievement beyond the assessment items cannot be assumed, unless generalization itself is assessed, the student's performance should probably be judged against alternate achievement standards.

Summary and Recommendations

Table 4.1 summarizes the educational characteristics of students that IEP teams should consider in making their recommendations regarding student participation in statewide assessment. States may specify additional considerations in their individual statewide assessment guidelines and guidance regarding the development of IEPs. Also, states that do not choose to use all the assessment options available in the regulations and summarized here should offer alternative guidance for the students with significant cognitive disabilities.

Table 4.1. Decision Framework for IEP Teams to Use in Choosing Assessment Options[8]

<table>
<tr><td>Prerequisite Considerations</td><td colspan="5">• Has the student had access to grade level content?
• Has the student had evidence-based instruction?
• Was instruction by a highly qualified teacher?</td></tr>
<tr><td colspan="6">If the answer is NO to any of these questions, create access and continue to apply grade level achievement standards and evaluate response to intervention.</td></tr>
<tr><td colspan="6">If the answer is YES to all three questions, consider the options and questions below</td></tr>
<tr><td rowspan="2">Consideration</td><td colspan="5">Assessment Recommendation</td></tr>
<tr><td>General Assessment</td><td>General Assessment with Accommodations</td><td>AA Grade Level Achievement</td><td>AA Modified Achievement</td><td>AA Alternate Achievement</td></tr>
<tr><td>Question 1:
In what way does the student access the general curriculum?</td><td colspan="3">Shows progress in the full scope and complexity of the grade level curriculum, although may not yet be on grade level.</td><td>Does not show grade level achievement; needs changes in complexity and scope of curriculum to show progress in grade level content.</td><td>Due to significant cognitive challenges in memory and transfer of learning, needs extensive prioritization within grade level content.</td></tr>
<tr><td>Question 2:
What has been this student's response to academic interventions?</td><td colspan="3">Responds to grade-level instruction although may not yet be on grade level.</td><td>Academic problems despite appropriate and intensive instruction; multiple years behind grade level expectations.</td><td>Requires ongoing systematic instruction to learn prioritized skills; needs to focus on "critical essence" of content.</td></tr>
<tr><td>Question 3:
How does this student interact with text?</td><td colspan="3">On or near grade level in reading.</td><td>Needs controlled vocabulary/ reduced reading level; may also need text reader.</td><td>Needs key words, pictures, auditory cues embedded in adapted or controlled text; may need text reader to use these cues; may have some emerging reading skills.</td></tr>
</table>

[8] These options have not been endorsed by the Department of Education, as the regulations for alternate assessments judged against modified achievement standards have not been established as of the completion of this paper.

Figure 4.1, continued

Consideration	Assessment Recommendation				
	General Assessment	**General Assessment with Accommodations**	**AA Grade Level Achievement**	**AA Modified Achievement**	**AA Alternate Achievement**
Question 4: Do the supports required by this student to perform or participate meaningfully and productively in the general education curriculum change the complexity or cognitive demand of the material?	None needed.	Needs accommodation.	Needs modified presentations and responses but on grade level.	Needs support that reduce complexity or breadth of assessment items such as aids that reduce judgment needed to do task or teacher scaffolding during assessment.	Needs extensive supports such as simplified symbol system peer model, or motivation through choice making to retrieve response.
Question 5: What inferences can be made about the student's generalization/ transfer of learning?	Shows transfer of learning to the extent expected for the grade level during ongoing instruction			Transfer of learning is more limited in scope than grade level; may only transfer to similar or familiar content/ Contexts.	Needs systematic instruction to generalize; because generalization is especially challenging during instruction, should not be assumed unless assessed.

1. IEP teams need to use a systematic process for recommending how students with disabilities participate in large-scale assessments; the primary data for making this decision focuses on students' skills and needs.

2. The type of assessment in which the student participates should adequately focus on grade level standards that have been changed appropriately.

3. Such decisions should be evaluated annually as students begin to have better and more complete access to the general education curriculum.

References

Browder, D. (2005). *Research on academic learning by students with significant cognitive disabilities*. Presented to OSEP 15th Technical Assistance and Dissemination Conference. June 7, 2005. Retrieved June 9, 2005, from http://education.uncc.edu/access

Browder, D.M., & Spooner, F.H. (In press). *Teaching reading, math, and science to students with significant cognitive disabilities*. Baltimore: Paul H. Brookes.

Browder, D., Ahlgrim-Delzell, L., Courtade-Little, G., & Snell, M. (2005). General curriculum access. In M. Snell & F. Brown (Eds.). *Instruction for students with severe disabilities* (6th ed.) (pp. 489-525). Upper Saddle River, NJ: Merrill/Prentice Hall.

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 Person with Severe Disabilities*, *2,* 165-181.

Browder, D., Wakeman, S., Spooner, F., Ahlgrim-Delzell, L., & Algozzine, B. (2005). *Research on reading for students with significant cognitive disabilities*. Manuscript submitted for publication.

Clay, M., & Cazden, C. (1992). A Vygotskian interpretation of reading recovery. In L.C. Moll (Ed.), *Vygotsky and education: Instructional implications and applications of socio-historical psychology* (pp. 206-222). New York: Cambridge University Press.

DeStefano, L., Shriner, J. G., & Lloyd, C. A. (2001). Teacher decision making in participation of students with disabilities in large-scale assessments. *Exceptional Children*, *68*, 7-22.

Downing, J. E. (2005). *Teaching literacy to students with significant disabilities*. Thousand Oaks, CA: Corwin Press.

English, F. W., & Steffy, B. E. (2001). *Deep curriculum alignment: Creating a level playing field for all children on high-stakes tests of educational accountability*. Lanham, MD: Scarecrow Press.

Fisher, D., & Frey, N. (2001). Access to the core curriculum: Critical ingredients for student success. *Remedial & Special Education*, *22*, 148-157.

Fletcher, J. M., Denton, C. A., Fuchs, L & Vaughn, S. R. (in press). Multi-tiered reading instruction: Linking general education and special education. In J. W. Gilger & S. O. Richardson (Eds.), *Research-based education and intervention: What we need to know* (pp. 23-45). Baltimore, MD: The International Dyslexia Association.

Horner, R. H. & Albin, R. W. (1988). Research on general case procedures for learners with severe disabilities. *Education and Treatment of Children, 11*, 375-388.

Horner, R., H. & McDonald, R. S. (1982). Comparison of single instance and general case instruction in teaching a generalized vocational skill. *Journal of the Association for the Severely Handicapped, 7*, 7-20.

Hughes, C., Copeland, S., Agran, M., Wehmeyer, M., Rodi, M., & Presley, J. (2002). Using self-monitoring to improve performance in general education high school classes. *Education and Training in Mental Retardation and Developmental Disabilities, 37* (3), 262-272.

Hunt, P., Staub, D., Alwell, M., & Goetz, L. (1994). Achievement by all students within the context of cooperative learning groups. *Journal of the Association for Persons with Severe Handicaps*, *21*, 53-71.

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

Joseph, L. M., & Seery, M. E. (2004). Where is the phonics? A review of the literature on the use of phonetic analysis with students with mental retardation. *Remedial and Special Education*, *25*, 88-94.

Kennedy, C. H., & Haring, T. G. (1993). Teaching choice making during social interactions to students with profound multiple disabilities. *Journal of Applied Behavior Analysis, 26*, 63-77.

Kleinert, H., et al. (2005). *Implications of the "assessment triangle" for students with significant cognitive disabilities: The first vertex-models of student cognition*. Retrieved June 9, 2005, from http://www.naacpartners.org/

Kliewer, C., & Bilken, D. (2001). "School's not really a place for reading": A research synthesis of the literate lives of students with severe disabilities. *The Journal of The Association for Persons with Severe Handicaps*, *26*, 1-12.

Koppenhaver, D., Erickson, K., & Skotko, B. (2001). Supporting communication of girls with Rett syndrome and their mothers in storybook reading. *International Journal of Disability, Development and Education, 48*, 395-410.

Mechling, L., & Langone, J. (2000). The effects of a computer-based instructional program with video anchors on the use of photographs for prompting augmentative communication. *Education and Training in Mental Retardation and Developmental Disabilities*, *35*, 90-105.

Musselwhite, C., & King-DeBaun, P. (1997). *Emergent literacy success: Merging technology and whole language for students with disabilities*. Park City, UT: Creative Communicating.

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

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (Eds.) (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press.

Scruggs, T. E., & Mastropieri, M. A. (1984). Issues in generalization: Implications for special education. *Psychology in the Schools, 21,* 397-403.

Shriner, J. G. (2000). Legal perspectives on school outcomes assessment for students with disabilities. *Journal of Special Education*, *33*, 232-239.

Snell, M. E., & Brown, F. (Eds.). (2006). *Instruction of students with severe disabilities* (6th ed.). Upper Saddle River, NJ: Merrill.

Snyder, T. L., Freeman-Lorentz, K., & McLaughlin, T. F. (1993). The effects of augmentative communication on vocabulary acquisition with primary age students with disabilities. *B. C. Journal of Special Education, 17*, 73-93.

Stephenson, J., & Linfoot, K. (1995). Choice-making as a natural context for teaching early communication board use to a ten-year-old boy with not spoken language and severe intellectual disability. *Australia and New Zealand Journal of Developmental Disabilities*, *20*, 263-286.

Thurlow, M. L., Lazarus, S. S., Thompson, S. J., & Robey, J. (2002). *State policies on assessment participation and accommodations, 2001* (Synthesis Report 46). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved June 9, 2005, from the World Wide Web: http://education.umn.edu/NCEO/OnlinePubs/Synthesis46.html

Thurlow, M. L., Moen, R. E., & Wiley, H. I. (2005). *Annual performance reports: 2002-2003 State assessment data*. Minneapolis, MN: University of Minnesota, National Center on

Educational Outcomes. Retrieved June 22, 2005, from the World Wide Web: http://education.umn.edu/NCEO/OnlinePubs/

Westling, D. L., & Fox, L. (2005). *Teaching students with severe disabilities* (3rd ed.). Upper Saddle River, NJ: Merrill.

U. S. Department of Education. *Title I – Improving the Academic Achievement of the Disadvantaged; Final Rule, 68 Fed. Reg. 236* (December 9, 2003).

CHAPTER 5: STANDARDS AND ASSESSMENT APPROACHES USING A VALIDITY ARGUMENT

The purpose of this chapter is to apply the information from Chapter 2 (validity argument), Chapter 3 (approaches to assessment) and Chapter 4 (decision framework for participation options) with various examples from state standards and assessments. In the first section, we describe how accommodations can be validated. 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 using large-scale assessment systems. It is this process of taking a state standard and using a state test (item) to measure proficiency that needs to be carefully analyzed for the introduction of construct irrelevant variance.

In the second section, we compare two states on their 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 the second state is performance task-based. In each of these systems, we consider the kinds of procedural and empirical evidence that needs to be collected to back up the validity claim that performance on the large-scale assessment is an adequate indicator of proficiency. We focus on construct under-representation (and misrepresentation) in this analysis.

The third section of the chapter 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, a focus on what students can do and its relationship with scoring. These seven principles should allow states to develop an assessment system that is completely inclusive and seamlessly integrates all participation options.

In the fourth section, we operationalize these principles using an example of reading assessment in which we make slight changes to accommodate students participating in large-scale assessment 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: in supports (assistive technologies, prompts, and scaffolds) and in breadth, depth, and complexity. The three important components of the assessment model are drawn together in this chapter.

1. A validation process is explicated using an argument with claims. assumptions, and evidence to support the inferences that are made from tasks reflecting domains of knowledge and skills (Chapter 2).

2. An approach to assessment is analyzed, beginning with a typical statewide assessment and then, after working through appropriate accommodations, using approaches to assessments that rely on indirect teacher judgments (using rating scales and checklists), portfolios, performance events, and performance task collections (Chapter 3).

3. A range of options for participation is presented with various populations of students considered for any of several alternatives across subject areas based on the unique needs of the individual (Chapter 4).

In the fifth section, we present conclusions and recommendations by addressing all these issues. These recommendations simply make the validity argument explicit in the assessment approach, the options in which students participate, and the way changes are made.

Construct Irrelevant Variance and the Need for Accommodations

One of the first questions to consider in understanding a validity argument is to be clear about the construct being tested because it forms the basis of the claim. For example, in the following standards from Massachusetts and Oregon and mathematics problem from a 5th grade state practice test, a word story problem is presented as a multiple-choice item. The question is whether this item also has within it other features that are irrelevant to these mathematics constructs?

To answer the question, we must consider (a) the construct being assessed; (b) 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 (c) 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 is developed to adequately reflect the *domain of knowledge and skills* of the standards and can be used as the basis for the inference.

Table 5.1. Construct in an Example Mathematics Standard and an Assessment Problem

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

As noted in Chapter 2, two constructs are present in our assessment systems: a construct about achievement (and the domain of knowledge and skills associated with it); and a broader construct about cognitive skills applied to complex tasks that are learned over a lifetime. In the validation argument, then, we have two critical links in the chain: the first link is whether the task presented on the large-scale assessment appropriately measures the domain of achievement,

and the second is whether the achievement reflects anything more in terms of situated problem-solving that reflects cognitive skills. These two links represent construct misrepresentation and construct under-representation, as described in Table 5.2.

Table 5.2. Validation claim and questions supported by evidence.

Construct	**Misrepresentation**	**Under-representation**
Achievement (domain of tasks)	Does the math story problem include other constructs or relay on access skills that prevent student 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?
Cognitive abilities (complex tasks)	Does the situation contain elements or features that inadvertently distract the student from reflecting their abilities?	Does the situation reflect the full range of situations to reflect application of the student's ability?

In this simple mathematics problem, reading may be part of construct-irrelevant variance that impedes our efforts to measure the knowledge and skills of mathematics as applied in this limited situation (a printed word story problem). However, if we had used a performance task to measure achievement (open-ended problem requiring the student to write their answer), then writing may become part of the construct irrelevant variance. As a measure of cognitive ability in mathematics, we could have required a demonstration of money estimation in the community; however, a host of other factors could then become sources of construct irrelevant variance 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.). As noted in Chapter 2, construct irrelevant variance can arise from several sources including the unique needs of students with disabilities and how they participate in large-scale assessment systems.

In the math story problem either as a measure of achievement or of cognitive skill, the construct also could be seriously under-represented 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 these instances, the construct may have been under-represented.

The sufficiency of evidence needed to support the claims and the sheer presence of competing threats to support or explain the claim are virtually insurmountable. And in making the claim, serious social consequences are at stake. Egregious misinterpretations are likely to be made (e.g., the student is not proficient in mathematics). Resources are likely to be squandered (e.g., very complex tasks are used in which reliability-related evidence is found lacking). Tasks are misrepresented as constructs and the three communities of measurement, content, and special education begin to drift apart. In this volume, we emphasize the need for appropriate and credible assessment approaches, knowing their limitations in making inferences about cognitive skills using more complex tasks, and yet, maintaining the credibility of what is present and known. In the end it is likely that states need to focus primarily on the assessment of achievement from the first construct and then emphasize skill development needed to reflect achievement of the content standards.

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 to students who cannot read to eliminate reading as a construct irrelevant variable. Indeed, reading mathematics problems may be one of the most consistently supported accommodations in large-scale mathematics assessment research. 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). Indeed, Tindal and Ketterlin-Geller (2004) note the following in their review of mathematics accommodations research on four major classes of accommodations (using calculators, reading mathematics problems to students, having extended time, and using multiple accommodation packages). Notice, however, that these task (test) features, may be problem-specific 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 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 (p. 8).

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. At other times, however, accommodations may 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, 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 explicate the reasoning behind an accommodation that considers the construct and the evidence needed to support its measurement (see Thurlow & Bolt, 2001). We address that kind of evidence in Chapter 12 acknowledging the lack of findings or uniformity in findings (c.f.,

National Center on Educational Outcomes *Online Accommodations Studies*). 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 to be made in testing.

Table 5.3. Types of Accommodations

Presentation	**Presentation Equipment**	**Response**	**Setting**	**Scheduling**
Large print	Magnify 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 Cass	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. As we noted in Chapter 1, the general education large-scale assessment (with or without accommodations) allows 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 breadth, depth, and complexity, we can no longer make the same statements about score comparability; nevertheless, with a validity argument and in the context

of federal regulations, we can make comparable statements about proficiency for purposes of reporting Adequate Yearly Progress.

In the sample mathematics problem presented at the beginning of the chapter, changes could be made in the assessment approach by observing the student actually make change and using a checklist or rating scale to note the correctness of the response; by assembling into a portfolio materials that documents 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 these options could become part of an alternate assessment judged against modified or alternate achievement standards. Remember though, 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 misrepresented or under-represented), requiring both procedural and empirical evidence.

Validity Argument using Different Alternate Assessment Approaches

We integrate the validation 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 grade three to grade five content standards in mathematics. The selection of states and grade level was somewhat arbitrary though some research on them has been published previously (see Weiner, 2002, for a description of Massachusetts and Tindal, McDonald, Tedesco, Glasgow, Almond, Crawford, & Hollenbeck, 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) need 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), the stability and consistency in sampling behavior over replications, 'item' or task functioning, reliability of judgments and scoring, internal relations among items and tasks, response processes, as well as external relations with other measures. Just as in the validation of the general education test (with or without accommodations), attention needs to be given to construct irrelevant variance and even more importantly to construct misrepresentation or under-representation; this latter problem is particularly critical as changes are being made in depth, breadth, and/or complexity of the standards.

Massachusetts Standards and Alternate Assessments with Portfolios

We confine our analysis to the standards for grades 3 and 4 that focus on number sense (which has 7 objectives within it) and operations (which has 3 objectives within it), 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, a very reasonable reach to several mathematics constructs that focus on number sense and operations, fitting well with the standards from the National Council of Teachers of Mathematics (2000).

Table 5.4. Massachusetts Standards for Grades 3 and 4

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; demonstrating an understanding of the values of the digits; comparing and ordering the numbers.	♦ Manipulate numbers at a higher level by counting, writing, grouping, sorting, comparing, 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 X 100 + 5 X 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 and 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, 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 concretely ♦ Use all operations to solve problem situations related to money ♦ Understand commutative properties of addition and multiplication (order can be reversed)
4.N.9	Select, use, and explain the commutative, associative, and identity properties of operations on whole numbers in problem situations, e.g. 37 X 46 = 46 X 37, (5 X 7) X 2 = 5 X (7 X 2).	
4.N.10	Select and use appropriate operations (addition, subtraction, multiplication, and division) to solve problems, including those involving money.	

To provide a functional equivalence across standards, an 'essence' of the standard is distilled in Massachusetts, where standards are translated into some minimal specifications that eventually are used in guiding development of alternate assessments. These grade level standard 'essences' are then used to explicate the alternate assessment system to ensure it is aligned with them and to help structure the assessment approach (in this state, portfolios). For each grade level standard, the state has illustrations posted on their assessment web site (retrieved on May 30, 2005) for structuring *activities* related to the standard (left side) and documentation or product portfolio entry that eventually is judged as proficient (or not)

Table 5.5. 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 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 peers. 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 assistance from the teacher, demonstrating her accuracy in identifying the 'next highest dollar' for purchases of $10.00 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 peer 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 peer • 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 products aligned with the 'essence' of standards. The primary issue we address here is the need for making a validity claim and then collecting both procedural and empirical evidence to support 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 really do the work that is displayed in the portfolio or is the teacher part of this process?

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 address the question of how clear and consistent the results are reported to the public.

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 score 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 scored like generalization, independence, and achievement?
- How well do work samples 'hang together' to reflect a standard?
- Which work samples in the portfolio are related to each other and therefore reflect a consistent structure?

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

- Are there consistencies in how students respond reflecting 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, content related, 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 and teachers as well as be related to content-grade level standards.

Oregon Standards and Alternate Assessments with Performance Tasks

We now turn to another state in which comparable standards have been articulated, though with a slightly different grade level (4-5). In Oregon, the standards have been affixed to common curriculum goals (CCG) and in this particular example three CCGs have been 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 is slightly different, which eventually has bearing on any alignment analysis. In Massachusetts, there are two standards with 10 objectives (7 in one standard and 3 in another) while in Oregon, there are three standards (with 19 objectives or almost double the number in Massachusetts): 5 objectives address *numbers*, 11 objectives address *computation and estimation*, and 3 objectives address *operations and properties*.

As we noted with the Massachusetts standards, those developed in Oregon have a certain breadth and depth, with reach into the grade level curriculum. 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 5.6. Oregon Mathematics Standards: Numbers, Computation, and Operations- Grades 4-5

Common Curriculum Goals	**Oregon Grade-Level Standards** **Grade 4**
Calculations and Estimations	

Common Curriculum Goals	Oregon Grade-Level Standards Grade 4
Understand numbers, ways of representing numbers, relationships among numbers, and number systems.	NUMBERS Read, write, order, model, and compare whole numbers to one million, common fractions, and decimals to hundredths. Identify the place value and actual value of digits in a number to one million. 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 0-9. 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 x 70 or 2700 / 30). Identify the most efficient operation (add, subtract, multiply, or divide) 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 1000 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 5.1. For each standard, two items are displayed: (a) a practice test that is used by teachers to get students familiar with the test format, and (b) an actual performance task used in the alternate assessment.

Standard: *Supply a missing element in or determine a rule that extends number patterns involving addition or subtraction of decimals.*
Standard*: Describe, extend and generalize 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 to one million, common fractions, and decimals 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 one million*

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?"*

Figure 5.1 Oregon Mathematics Standards: Numbers, Computation, and Operations-Grade 4-5

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 students are presented the items and the way they can respond. What is the functional equivalence of these alternate assessments? 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 support the inference that is being made. Have the tasks been adequately developed and assembled into an alternate assessment? Are the directions for administration clear and understandable? Are teachers sufficiently trained in their administration and scoring (especially because responses are scored as partially correct not just correct vs. incorrect)? In this alternate assessment, teachers are trained on how to make accommodations (vs. modifications): Are the directions for distinguishing between these two types of administration sufficiently clear? Are student performances accurately recorded and entered in the computer?

Empirical evidence also needs to be collected to support the inferences being made about the construct (the standards) being assessed. Are enough tasks present to reflect the domain (and avoid construct under-representation)? What is the alignment of the alternate assessment with the state standards at this grade level: in categorical concurrence, depth of knowledge, range of knowledge, and balance of representation? How reliable are teachers in test administration 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 Item-Task Development and the Validity Argument Applied to Populations

Given these two states, each with their own standards and assessments, we have emphasized a process for collecting evidence that is both procedural and empirical. In Massachusetts, the evidence focused on the use of portfolios to make inferences about proficiency on state standards, while in Oregon, performance tasks were used to make the inferences about proficiency on the state standards. This evidence, however, can be collected only from extant assessment systems, which begs the question of how to develop items and tasks from the beginning. How can an assessment approach developed to both reflect grade level content standards *and* fit the specific unique needs of individual students?

We present an overview of universal design in the belief that many changes can be made in the manner that tests are developed so that a sizable proportion of students with disabilities can be included appropriately and meaningfully in making inferences to state standards. Nevertheless, for some students, changes in supports as well as breadth, depth, and complexity are also needed so they can meaningfully participate in the large-scale assessment system. The

most important issue is the application of principles of item development that can be used by states to fit within their assessment approach.

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

Grade level curriculum should be the primary resource for developing alternate assessment items and performance tasks. Weak alignment for academic content can occur when items are developed from resources other than the general curriculum 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[1] are likely to be too broad or too narrow, vague, age inappropriate, or not representative of the academic content (Browder, Flowers, et al., 2004). In a study Browder, Spooner et al. (2003) found that, for state alternate assessments nominated as being most closely aligned with national standards in reading and mathematics, more academic tasks *and* contexts were used for their items.

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 age/grade appropriate. Alternate assessments may sample knowledge forms and skills typically mastered at earlier grades by using age/grade appropriate items and performance tasks. For example, a 5th 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

[1] 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 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* (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 (Claudia Flowers, Diane Browder, & Lynn Ahlgrim-Delzell (2004) "An Analysis of Three States' Alignment Between Language Arts and Mathematics Standards and Alternate Assessments."

adaptations of the literature from the 5th grade reading series for the assessment rather than preschool materials. Or, for example, the 5th 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 5th 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 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 in understanding the repeatability or stability of behavior: Enough items or performance tasks need to be included to ensure that the student understands the concept rather than simply able to perform the 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 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.

Principle 2: The alternate assessment should parallel the breadth of the general curriculum for the grade level.

Alternate assessments should be developed to address sufficient breadth of the curriculum. For example, in the elementary grades, mathematics or science curricula often cover a wide breadth of content with less 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 reflect specific breadth, depth, and complexity. That is, if the 3rd 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 assessment of fewer strands and how the additional strands are sampled in subsequent grades. Alternate assessments judged against modified and alternate 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 caveat to 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 implicitly implied (if not explicitly provided). Rarely do we just expose students to content without task analyzing the skills needed to be successful in the content. Often earlier skills are needed to express later skills (for example, sufficient 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 towards simpler levels of knowledge compared to the more complex levels reflected in the state standards. Although alternate assessment as judged against modified achievement and 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 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. A student in an alternate assessment judged against alternate achievement standards would have items designed to demonstrate proficiency judged against modified achievement standards.

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

Although assessments can be developed so they are aligned with state standards, they also may distort that standard or be inappropriate for the student. Most skills are developed within a progression, and it is unlikely that good intentions alone (e.g., getting to grade level content standards) are a sufficient rationale for subjecting students with disabilities to items and tasks that aligned. Learning to read implies a progression of skills that are vertically aligned. For

example, phonemes and morphemes comprise the basic units; words can be read in ways that allow for segmenting and blending phonemes with graphemes; sentences and passages reflect both local and global dependence; comprehension is a complex construct with multiple features that build in complexity. Mathematics is similarly 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). This skill sequence needs to be considered in making changes to assessments so that specific items are still aligned with the standards, perhaps as part of a skill sequence that is vertically aligned and reflect both an appropriate assessment for the student and is an honest 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 8th grade compared to 3rd grade, some added achievement should be expected in 8th 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 8th grade as compared to the 3rd grade alternate assessment.

Principle 6. Alternate assessment items and performance tasks should be universally designed and 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 skill. 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 to supplement or supplant changes in breadth, depth, and/or complexity needs to be made carefully. Often, a simple change in supports can be made with little change in the construct (which is much more preferred than making 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., learning to read 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 affect this practice has on ensuring comparability of scores.

Example Assessment Development (in Reading)

With these guiding principles, we now present examples of items and tasks that have been changed (in supports as well as breadth, depth, and/or complexity) to reflect accommodations or alternate assessments to be judged against modified or alternate achievement standards. In reading through these examples, it is important to remember that accommodations reflect changes that provide comparable inferences being made (in relation to proficiency of state standards). That is, if the change results in measurement of the same construct, then it would be considered an accommodation, and the scores could be aggregated with the general education test. With alternate assessments, however, the inferences that can be made are somewhat constrained (in the case of performance judged against modified achievement standards) and quite stipulated (in the case of performance 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 will become part of the alternate assessment as judged against modified or alternate achievement standards. Therefore, the *scores* could not be aggregated with the general education *scores* but need to be reported separately. Furthermore, limitations would be placed upon the inferences we can make.

We describe the use of two types of changes to ensure appropriate participation in large-scale testing: (a) use of supports in the manner that tests are administered or taken (e.g., use of assistive technologies, prompts, or scaffolds) and/or (b) 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-complexity, as these categories are not meant to be exclusive. We present these examples only as illustrations. Clearly, these examples need to be embedded within a state system and then subjected to the very analysis we have proposed in the first four chapters, using a model that includes a validation argument in which evidence is collected using an approach to assessment on a specified population of students.

An important caveat to these changes is that accommodations are based only upon changes in supports and not changes in the breadth, depth, and/or complexity of the grade level content standards; for alternate assessments judged against grade level content standards, the changes can include changes in supports and or breadth-depth-complexity but the inference to standards is comparable. The reason for this caveat is to provide comparable inferences to these standards; the assessment needs to be parallel to that used in the general education administration without accommodations. Table 5.7a describes a 4th grade level standard and the two kinds of changes that could be made in the assessment.

Table 5.7a. Decode or Comprehend Meaning of Words in Text

Standard	*Changes using assistive devices, scaffolds, and prompts*	*Changes in breadth, depth, or complexity*
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*: Because of the verb ‘reproduce’ in the standard, the student needs to produce the word. Nevertheless, the word can be presented 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 *2% alternate assessment using modified achievement standards*: In this scenario, two types of changes could be made:

- *Support changes* (assistive devices, prompts, and scaffolds) could be used to make the inference to this standard limited. For example, the student could be asked to recognize words (e.g., receptive mode rather than expressive mode). All 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.
- *Breadth, depth, and/or complexity 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 in 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 *1% alternate assessment using alternate achievement standards*: Again, two types of changes could be made:

- *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).
- *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% of the English language text: *the, of, and, a, too, in, is, you, that, it.*

Table 5.7b. 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 (number of unique words to total words)

Table 5.7b describes a second 4th grade standard and the two kinds of changes that could be made in assessment of 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 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 *2% alternate assessment using modified achievement standards*: In this scenario, two types of changes could be made:

- *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 their 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 is the same while 'no' would signify they are different. A second example might be to present student with two passages, one of which has numerous syntactic and semantic errors. Student would identify the passage that is most correct.
- *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 *1% alternate assessment using alternate achievement standards*: In this scenario, two types of changes could be made:

- *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 3–5-word sentences that are displayed one sentence per line (not wrapped as in normal text).
- *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 chapter by considering construct irrelevant variance and construct under-representation. In both, the starting point is to be clear about the construct and in two kinds of distortions that can arise in its measurement. The former comes about when items and tasks are given or taken in a manner that prevents a clear inference about achievement; often other knowledge and skill also is being assessed in addition to that viewed as central to the construct. The latter 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 intertwined with this central construct).

Several 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 5.3, grouped into five major types (presentation, presentation equipment, response, schedule, and setting), considering them as accommodations. Using these accommodations results in a score that can be aggregated with the general education scores and a comparable inference made 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 central construct, this list of 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 used a portfolio approach as their alternate assessment; the second state used performance tasks. After describing the approaches to

assessment, we considered the kinds of procedural and empirical evidence that needs to be collected in supporting 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 principles for developing items and tasks that should be considered in developing any alternate assessments, whether judged against 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 (a) the supports used as part of an assessment as well as (b) 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 assessment can be comparable (when accommodations are made) or become either somewhat constrained (alternate assessment judged against modified achievement standards) or quite stipulated (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 assessment approaches (with various task formats) of the content and to systematically provide a variety of participation options for students with disabilities.

References

Algozzine, R. (2004). A content analysis of the curricular philosophies *and Practice in Severe Disabilities., 28,* 165-181.

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 Person with Severe Disabilities*, *2,* 165-181.

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

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

National Council of Teachers of Mathematics (2003). *Standards*.

Roid, G., & Haladyna, T. R. (1989). *The technology of item writing*. Boston: Kluwer Publishing.

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 [today's date], from the World Wide Web: 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, D.C.: National Assessment Governing Board. Available at http://www.nagb.org/pubs/conferences/tindal.pdf

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 [today's date], from the World Wide Web: http://education.umn.edu/NCEO/OnlinePubs/Synthesis48.html

CHAPTER 6: TEST DESIGN, DEVELOPMENT, SCORING, AND ANALYSIS

Introduction

The purpose of this chapter is to describe the planning, constructing, implementing, and scoring process that needs to be used so large-scale assessments are universally accessible and result in valid inferences. We address several topics in this chapter, including the use of state standards to initiate development of achievement measures; the use of test specifications (blueprints) for selecting specific types of tasks (item formats) that reflect breadth and depth of objectives, and cognitive complexity; consideration of universal design to ensure appropriate access; the need for item and test reviews to ensure no bias is present; the actual test construction and administration in classrooms with consideration of various accommodations; specific designs applied to alternate assessment approaches; a follow-up alignment analysis of the test with state grade level content standards; application of the 1999 *Standards for Educational and Psychological Testing* for students with disabilities (SWD); test scoring; and finally fairness in testing and test use. In keeping with the validity argument, we describe each of these components by addressing the evidence that needs to be collected to ensure appropriate claims and inferences can be made.

State Standards to Guide Achievement Measures

The first step in test design is to be clear about the focus of content and the decisions to be made from the outcomes. In this volume, we assume that the content of the tests is driven by grade level academic standards that have been adopted in each state. Furthermore, we assume that the purpose of the test is to ascertain proficiency on those standards.

As noted in Chapter 2, we distinguish between two critical constructs: (a) the domain of knowledge and skills as measured by the large-scale assessment, and (b) domain of cognitive

abilities that are further assumed to develop over a lifetime. We use tests composed of specific tasks to measure knowledge and skills. Also, we use learning objectives related to knowledge and skills as proxies that reflect cognitive abilities. In this process two types of inferences are made.

1. The tasks used to measure achievement reflect the domain of knowledge and skills within state standards.

2. The knowledge and skills of displayed on these tasks adequately reflect the cognitive abilities that are developed within students over time (both within a year and over years).

For purposes of developing an adequate measure we focus primarily on knowledge and skills stated in state standards. For example, Oregon grade 8 literature standards state that the *student will read, construct meaning, and respond to a wide variety of literary forms. Specifically, in the second standard on literary elements, the student must demonstrate knowledge of literary elements and techniques and how they affect the development of a literary work.*

a. *Analyze and explain elements of fiction including plot, conflict, character, mood, setting, theme, point of view, and author's purpose.*

b. Identify and explain various points of view and how they affect a story's interpretation.

In chapter 5 we described how the tasks (content and format) served as a fulcrum bridging the construct of state standards and the inferences we could make. We emphasized that the inference was only as good as the measurement tasks and the degree to which the format was free of irrelevant variance and the content had been adequately sampled (and not misrepresented or under-represented).

These state standards (in grade 8) leave considerable latitude for developing actual tasks, with three topics that need to be addressed. First, the response format of the task itself needs to

be considered with a host of possible options (see Table 1 below) using constructed responses (CR) or selected responses (SR). Second, the cognitive complexity needs to be considered in developing a task. Third, because of the limited time that is available for measuring students, the breadth and depth of standards must be considered, requiring decisions about the number of items or tasks within and across various standards. In part, the format of the tasks may have bearing on the number of items or tasks that can be included (e.g., constructed responses may result in fewer tasks being presented). In this chapter we focus primarily on tasks reflecting the domain of knowledge and skills within these state standards, but standards can be very broad (with verbs like *analyze* and *explain*) reflecting even broader cognitive abilities that become part of larger constructs within life-long learning.

Test Specifications (Blueprint) – Mapping Format, Cognitive Complexity, Breadth-Depth

In the glossary, we have included the definition of 'test specifications' to help guide the process for developing a test: "A detailed description for a test, often called a test blueprint, that specifies the number or proportion of items that represent each content and process/skill area; the format of items, responses, and scoring rubrics and procedures; and the desired psychometric properties of the items and test such as the distribution of item difficulty and discrimination indices" (Standards, 1999, p. 183).

Generally, the blueprint provides a direct mapping of the standards (and objectives) with items and tasks. Because item format is such a critical issue in considering the number of items or tasks as well as their cognitive complexity, the first step may be to consider this dimension. Haladyna (1999, p. 24) provides a very thorough table of possible formats, some of which can be incorporated into the assessment approaches depicted in chapter 3.

Table 1. Item Type for Constructed and Selected Response Formats

Response Format	
Constructed Response	**Selected Response**
Anecdotal	Conventional
Critique or review	Alternative
Demonstration	Matching
Discussion	Complex multiple choice (not recommended for use)
Essay	True-False (not recommended for use)
Exhibition	Multiple True-False
Experiment	Comprehension Item set
Fill-in-the-blank/short answer	Pictorial Item Set
Writing Sample	Problem solving item set
Interview-Observations	Interlinear Item set
Instrument Aided Observation	Experimental Innovative
Oral Report	
Performance	
Portfolio	
Project	
Research paper	
Visual Observation	

In making a choice about format, "the type of items, the response formats, scoring procedures, and test administration procedures should be selected based on the purposes of the test, the domain to be measured, and the test-test takers" (Standards, 1999, p. 44). Most state educational agencies determine the assessment approach rather than the Individualized Educational Program (IEP) team; nevertheless, in building a large-scale assessment system, it is important to consider these issues immediately so that the various options are available for individuals with disabilities.

As noted in Chapter 3, the typical approach used in large-scale assessments employ multiple-choice tests (with perhaps performance task in writing). The primary assessment approach used in alternate assessments is either a portfolio or a performance task (or collection of performances). If we use the literature example above, we could develop the following blueprint.

Table 2. Example of Basic Test Specification Associating Content Standards with Format

	Type of Assessment	
Standard or Objective	**Typical Assessment (MC)**	**Portfolio Alternate Assessment**
Fiction elements (plot, conflict, character, mood, setting, theme, point of view, and author's purpose).	16 multiple-choice items (2 per element)	3 work sample entries
Point of view influence on interpretation	1 short critique or review	2 work sample entries

An important issue, therefore, in choosing a format is to be clear on what constitutes an 'item' within any of the approaches to assessment from Chapter 3. As Tindal (2005) notes, a 'behavioral event' (e.g., an entry from a portfolio) may need to be used as a 'stand-in' for an item. Behavioral events have four attributes: (a) they reflect routines (with a beginning-middle-end), (b) they are captured in one session (e.g., sitting/setting), (c) they comprise a skill or multiple skills, and (d) they contain one or more constituent parts. This term is introduced for two reasons.

1. The analysis of validity evidence is more straightforward. With brief tasks, more events can be counted; with longer tasks, fewer events are considered. The critical effect is on the reliability of scores: In general, it is easier to attain higher levels of reliability.

2. The alignment process can be realistically operationalized: it doesn't become too granular on the one hand or too vague on the other hand. If every single behavior of an assessment approach is task analyzed and placed into the alignment process, it may become too cumbersome. However, specificity of behaviors comprising the assessment needs to be applied to ensure the alignment process is informing.

We also can manipulate the breadth and depth of standards being sampled. For example, we could include some but not all the fiction elements (only focus on plot, conflict, character, mood, and setting, but not theme, point of view or author purpose). The number of items for each

of these standards and objectives also can be reduced or their proportionate point value rearranged in the computing total score.

The problem with this blueprint is that the complexity is still unknown. We could use Web's (1999) depth of knowledge definitions to guide cognitive complexity (e.g., factual recall, conceptual analysis, strategic thinking, and extended thinking). In this standard then, the entries are described explicitly resulting in three entries, one for each level of depth: (a) factual, (b) conceptual, and (c) strategic. This would require considerable description so that teachers would be clear on how to document work samples that reflect them.

In the end, this step of deciding the 'item' format and mapping it into the grade level content standards being addressed is to assure that appropriate inferences are being made. The 1999 standards are very clear about this component of test design and development:

Table 3. Applicable APA Testing Standards for Test Specifications

	Applicable APA Standards
3.3	*"Test specifications should be documented, along with their rationale and the process by which they were developed. The test specifications should define the content of the test, the proposed number of items, the item formats, the desired psychometric properties of the items, and the item and section arrangement. They should also specify the amount of time for testing, directions to the test takers, procedures to be used for test administration and scoring, and other relevant information" (p. 43).*
3.15	*"When using a standardized testing format to collect structured behavior samples, the domains, test design, test specifications, and materials should be documented as for any other test" (p. 46).*

Universal Design (UD) for Assessments

Recently, the development of tests has been scrutinized for their universality in being accessible to ALL students. The original conception of UD described seven features: (a) equitable use, (b) flexibility in use, (c) simple and intuitive use, (d) perceptible information, (e) tolerance for error, (f) low physical effort, and (g) size and space for approach and use. In a summary of UD applied to large-scale *assessments*, Thompson, Johnstone, and Thurlow (2002) describe seven elements.

1. Inclusive assessment populations – All students need to participate in large-scale assessment systems regardless of personal characteristics. This requirement might mean that items need to be better pilot tested to ensure equitable access and that in this pilot testing, the full range of students and student needs be considered.

2. Precisely defined construct – Because an item or task includes both content and format, the target construct needs to be clearly emphasized and then the format can be manipulated appropriately.

3. Accessibility, non-biased items – There should be multiple ways to interact with the item or task. For example, many constructed responses are difficult to adapt for various interactions while selected responses have many different possibilities (e.g., students point to the correct option, blink when the correct option is pointed to, or a switch to indicate the correct option).

4. Amenable to accommodations – It should be possible to make slight changes in the supports used to give or take a test and still consider the outcome as comparable to that attained without the support. This kind of interpretation emphasizes the functional equivalence of the response relative to the construct of interest.

5. Simple, clear, and intuitive instructions – When students are being tested on knowledge and skills, directions in how to respond need to be clear so poor performance is not confused with misunderstanding the task (requirements).

6. Maximum readability – Critical information for responding to the task needs to be obvious in the most and straightforward manner with information accessible.

7. Maximum legibility – Both text and graphic illustrations need to be clear and easy to understand. Plain language strategies may be necessary to ensure the content is clearly available.

While the current wave of interest in universal design appears to be taking a central role in test development in special education, most good test design has always included many of these issues (and more). For example, Haladyna (2004) wrote an entire book on how to create multiple-choice items and provides very specific steps in how items (and options) should be worded and formatted. Nevertheless, some of the features of UD extend these measurement principles to be more inclusive and catholic.

Test Review for Bias and Sensitivity

Once items and tasks are selected and formatted, they need to be reviewed for bias (gender, disability, race-ethnicity, and other stereotypic issues). Obvious words and phrases that are sexist or politically incorrect need to be removed; subtle issues, however, may arise when considering content, however. For example, the story could contain words that are only known by sighted individuals (e.g., "a flash of light pierced through the open door") or only by hearing individuals (e.g., an analogy that notes "his words rolled out like thunder"). In both instances, the figures of speech may be biased against students with sensory impairments.

Table 4. Applicable APA Testing Standards for Bias Review

	Applicable APA Standards
7.3	*"When credible research reports that differential item functioning exists across age, gender, racial/ethnic, cultural, disability, and/or linguistic groups in the population of test takers, test developers should conduct appropriate studies when feasible. Such research should seek to detect and eliminate aspects of test design, content, and format that might bias test scores for groups" (p. 81).*
7.4	*"Test developers should strive to identify and eliminate language, symbols, words, phrases, and content that are generally regarded as offensive by members of racial, ethnic, gender, or other groups, except when judged to be necessary for adequate representation of the domain" (p. 82).*

Test Construction and Administration (Static and Interactive Features)

Items and tasks are displayed with students and teacher directions so that the essential knowledge and skill is being measured and not irrelevant components (which would be considered construct irrelevant). The items and directions are static features of test construction

and administration. In addition, however, test administrators typically interact with students while giving a test. Although most large-scale assessments do not require formal training in their administration, they typically assume a standard administration: All test administrators give the test to students in a consistent fashion as directed by the test manual. We consider this as an interactive feature of test construction and administration. In either the static or interactive component of test construction and administration, several accommodations are possible for making changes in the way the test is given or taken.

Static features (the items and tasks as well as the directions on the test). The test booklet needs to provide information in a logical and easily accessible manner both for the teacher/tester as well as the student. For example, the test administrator manual should contain all the necessary information for giving the test, so the administrator does not have to go somewhere else to find it. If accommodations are allowed for certain portions of the test, they should be noted in a place where they can be found in advance of giving the test (e.g. as a prelude to the directions as opposed to an appendix at the back or published as a separate document). The student booklet should have plenty of white space and the information formatted to fit the page (with no page breaks occurring within a problem or task).

Dynamic features (the way tests are given or taken). The directions for teachers to consider in giving the test need include both helpful strategies for interacting with students as well as explicit directions in what to say. For example, teachers may be given suggestions on how to break up the test, how to handle interruptions, what to do with various questions from students, and suggestions for dealing with student conative factors (motivation and efficacy). Likewise, the test needs to provide the test administrator with the actual wording to use in giving the test.

Design Applications to Alternate Assessment Approaches

Because of the unique assessment approaches to alternate assessments (particularly in applying performance assessments and portfolios), several steps can be taken to ensure the validity of inferences made based on the performance of students. These are noteworthy issues because they are likely to (or should) appear in the actual test booklet when it is initially designed. Certainly, these issues are not exhaustive but should serve as a reminder that the focus on the assessment is to develop an item or task, assemble within a test, and administer it with enough integrity to support an inference based on students' performance.

- When using indirect teacher judgment with rating scales, the environmental context of the behavior being judged needs to be clearly stated. Teachers need to know (a) when and how to consider a behavior and (b) make a judgment. For example, when reflecting on the independence of a student in participating in a reading discussion (think of the grade 8 literacy standards noted at the beginning of this chapter), teachers need to be directed to consider the setting and the interactions of the student as well as what the judgment of independence entails (e.g., use of guided prompts).
- The activity (or process) resulting in a work sample for a portfolio needs to be specified, including the level of assistance provided to the student. Teachers often are participant observers when collecting documents from students and placing them in a portfolio. Therefore, the nature of this interaction needs to be documented along with the product. For example, when creating a graph depicting changes in amount for different groups (e.g., a voting record for girls and boys that has been enacted in the classroom), the context of this activity needs to be documented.

- Judgments of products should be made as close as possible in time to the actual execution or completion of a product. Because judgments reflect complex scoring influenced by many factors (social, psychological, environmental, etc.), they need to be conducted when all the relevant information is present and not contaminated by events that intercede between the behavior and the judgment.

Alignment of Tests with Grade Level Content Standards

Assuming the test is designed and developed under optimal conditions, a need exists for ensuring that the items and tasks align with the content standards. We have described Webb's (1997, 1999) alignment in several other chapters of this white paper as a system to ensure appropriate and balanced coverage of the standards. In this system, the following definitions apply. Achieve provides another excellent model for alignment (Resnick, Rothman, Slattery, & Vranek, 2003-2004) that is very close in the constructs and only slightly different in the process.

Categorical concurrence. The focus is on whether the standards and the assessments address the same content categories. In Webb's (1999) system, a requirement is made that the assessment needs to contain at least six items from a standard, the rationale being related to the need for making reliable judgments for ascertaining mastery. In considering the categorical concurrence of behavioral events, however, it may be as important to consider the amount of behavior reflected in each event (in which case the sheer number of events is somewhat less relevant than the type of behavior). The goal is to achieve a fair count (percentage) of standards that have stable representation in the assessment at the level in which scores are reported. In whatever manner behavioral events are defined within any assessment format, the key issue is whether they address a standard.

Depth of knowledge (DOK). The focus in DOK is on the cognitive complexity required by the standard and the assessment with four levels possible. If an assessment item fits more than one objective, it needs to be coded as primary or secondary. According to Webb, at least 50% of the target items need to be at or above the level required by the objective. Other taxonomies could be applied but this one provides a simple demarcation between factual recall, conceptual analysis, strategic thinking, and extended thinking; large-scale assessments seldom focus on strategic thinking and virtually never address extended thinking (which is more likely to tap in the domain of tasks reflecting cognitive abilities).

Range of knowledge. This feature address how well represented each standard is in the alternate assessment. This category refers to the breadth or spans of knowledge required by the assessment and matched to standards with at least 50% of the objectives needing at least one related item. Range of Knowledge can be calculated from extant data (the number of objectives with items divided by the total number of objectives in the standard) and reflected as a percentage. This index is averaged across standards.

Balance of representation. This index reflects the degree to which more than one objective on an assessment is given emphasis. A *hit* is a standard with a corresponding item. This index is defined as the proportion of objectives MINUS the proportion of hits assigned to objectives. Balance of Representation in the alignment system, therefore, reflects proportion of target skills with multiple representations on the standards. This multiplicity usually comes about through different 'items' obtained as unique tasks, performances, work samples, or observations. Balance should reflect the certainty of inference in being proficient on standards as a function of the 'items' present (tasks, performances, work samples, or observations).

Although a test format may have been thoughtfully applied to the development of a test and even if the content of the test is aligned with the grade level content standards, a number of other issues need to be considered to ensure it provides a fair and technically adequate measure that can be used in making decisions for students, particularly for students with disabilities. Because this topic is so critical for both the volume and current practice, we have turned to the standards (AERA, 1999) directly and summarized each one below.

Application of 1999 Standards for Testing Individuals with Disabilities

This section of the book is reprinted from Tindal (2005). In 1999, three organizations collaborated to update the *Standards for Educational and Psychological Testing* (1999): American Educational Research Association, American Psychological Association, and the National Council on Measurement in Education. As note earlier, this updated version created a seismic shift in the way validity was approached. Validity became defined as a unified concept that was inherently part of inferences and evidence, of decision-making and not necessarily a feature of the test. These standards have been published in a book with three parts: (a) validity, (b) testing individuals of diverse linguistic backgrounds (including individuals with disabilities), and (c) psychological testing and assessment. Each of these sections includes chapters that highlight specific standards, many of which we have included in this section. In the most pertinent section for students with the most significant disabilities, 12 standards are presented that highlight the following issues.

1. Steps should be taken to avoid that having a disability will influence the inferences that can be made about the intended construct.

2. Those who make accommodations should be aware of existing research on the effects of disabilities.

3. Tests developed for students with disabilities should be pilot tested on students with similar disabilities.

4. Modifications to tests should be described in detail along with the rationale; validity of the inference should be provided.

5. Technical materials should describe the steps taken to modify tests and aide interpretation.

6. Extended time limits should be established empirically.

7. The validity of inference for students from various disabilities should be documented when sample size permits.

8. Responsible decision-makers need to have access to accurate and current information about potential accommodations and modifications.

9. Appropriate norms should be based upon the decisions to be made.

10. Changes in testing should be appropriate to the individual and maintain as much as possible the "feasible" features of the standardization.

11. Score comparability between regular and modified assessments should preclude flagging.

12. Multiple sources of information should be used in testing individuals with disabilities.

These twelve principles are designed to guide interpretation for students with disabilities who have taken tests. Beyond these overall proscriptions, the 1999 standards have only begun to have sway in the field of special education. Alternate assessments require a complex mix of descriptions that provide explanations of their purpose and function as well as various instruments for measuring and scoring students' performance. Yet, they should follow the standards being applied to all other tests and measures in education.

Because alternate assessments include a range of formats (portfolios, observations, and performance tasks) and are used in such varying environments, evidence cannot be documented simply by comparing the classroom materials with the state alternate assessment materials. Rather, a full description of the state's alternate assessment needs to be recorded prior to collecting other evidence and as a prerequisite step to organize the test content. In this initial phase, all critical components need to be identified and operationalized; subsequently, evidence can be collected as part of curricular presence or as alignment with grade level content standards.

Test Scoring

Once all student responses have been 'captured' and recorded, the last step in the administration cycle is to score the assessments. Again, the 1999 standards are quire clear.

Table 5. Applicable APA Testing Standards Applied to Scoring

Applicable APA Standards	
3.14	*"The criteria used for scoring the test takers' performance on extended-response items should be documented. This documentation is especially important for performance assessments, such as scorable portfolios and essays, where the criteria for scoring may not be obvious to the user" (p. 46).*
3.24	*"When scoring is done locally and requires scorer judgment, the test user is responsible for providing adequate training and instructions to the scorers and for examining scorer agreement and accuracy. The test developer should document the expected level of score agreement and accuracy" (p. 48)*

> "A test score is a summary of the evidence contained in an examinee's responses to the items of a test that are related to the construct or constructs being measured. The sort of summary desired, and the extent to which that summary is intended to generalize beyond the specific responses at hand, depend strongly on the theoretical orientation of the test scorer. One point of view is the purely empirical: It sees the test score as a summary of the responses to the items on the test and generalizes no further. An alternative approach views the item responses as indicators of the examinees level on some underlying trait or traits; in this context, it is appropriate to draw inferences for the observed responses to make an estimate of the examinee's level on the underling dimension(s)" (p. 1).

Scoring alternate assessments can be on any number of dimensions, e.g., independence, generalization, accuracy, or fluency. Although most large-scale assessments are based on the number or percent correct, with alternate assessments, the dimensions being addressed are considerably more varied and more complex. As we mentioned in Chapter 5, if independence is being scored, then a fully prompted response should not count (no behavior, no score).

Two components are necessary for rubric scores: (a) a definition of the dimension being scored, and (b) anchor descriptions of values on this dimension. For example, the AIMS writing assessment used in Arizona high schools uses five traits, each of which is defined and scored with six levels: ideas and content, organization, sentence fluency, word choice, conventions, and Sentence fluency is defined at the highest level (6) as writing that has an effective flow and rhythm, at (5) as writing that has an easy flow and rhythm, at (4) as writing that flows, at (3) as writing that tends to be mechanical rather than fluid, at (2) as writing that tends to be either choppy or rambling, and at (1) as writing that is difficult to follow or read aloud.

Further issues may arise in scoring performances or work samples. For example, in the Oregon writing assessments, several codes are used to depict various non-scorable samples.

Table 6. Non-scorable Writing Performances in Oregon

BL – Blank.	The student writing folder has ***no*** writing in it. (The student may have been absent or exempted from the assessment.)
TS – Too short to score.	The paper is so short that scores would be meaningless (e.g., "Tuesday I had my cat de-clawed and I will always remember it.")
TL – Too long to score.	The paper exceeds the length allowed in the Administration Manual: "An extra sheet may be attached to allow a student to complete a thought – no more than finishing the sentence or paragraph begun on the final page of the writing folder." This code is not used if the student was tested under modified conditions or if a student has a disability which requires wider spaces in which to write.
NE – non-English.	The paper is written mostly or totally in a language other than English. (Note: This does not apply to papers written in Spanish. Spanish language papers are scored at a site where raters are specifically trained to score them).
IL – Illegible.	Several raters have looked at the paper and ***cannot decode*** the words. This code is not used to indicate messiness. Papers that are coded as illegible are not readable ***at all.***
PV – Profane or Graphically Violent.	The paper includes extreme violence or profanity well beyond usual community standards for school writing.
MISC – Miscellaneous.	These papers have one or more of the following characteristics: (a) poetry, or (b) plagiarism

Fairness in Testing and Test Use

In the 1999 standards, several issues are presented about test fairness and eventually 12 standards are invoked for judging this dimension. Four views of fairness are promulgated, of which the first two address freedom from bias, the third pertains to outcomes, and the fourth focuses on opportunity to learn.

1. The technical term for bias is "different meanings for scores earned by members of different identifiable subgroups" (p. 74).

2. It is not just the item or test that can be biased, however, but the context and purpose also: "the use of the test in a particular circumstance or with particular examinees may be fair or unfair…fairness also requires that all examinees be afforded appropriate testing conditions" (pp. 74-75). In a break with the purely and previously held notions of the critical need for standardization, the current standards are clear in supporting the need to consider the individual need of students: "In some instances, greater comparability may sometimes be attained if standardized procedures are modified" (p. 75). This notion of fairness also includes the

opportunity to prepare for tests and have access to resources that are deemed critical to participation of the test. Finally, fairness addresses reporting of outcome (e.g. flagging results when accommodations are used).

3. Fairness can be considered in terms of outcomes and the degree to which individuals or groups are provided or denied resources as a function of performance. Although unequal outcomes do not necessarily raise suspicion of fairness, it may be important to minimize such differences across relevant subgroups.

4. Finally, fairness is viewed in the context of opportunity to learn. "When test takers have not had the opportunity to learn the material, the policy of using their test scores as a basis for withholding a high school diploma, for example, is viewed as unfair" (p. 76). As noted in the Turlington case (Pullin, 1983), two years of participation in the content often is considered necessary before such high stakes (of awarding high school diplomas) can be considered.

To ensure bias and fairness, then, it is important to consider the content of the test to ensure specific offending language is avoided and alignment with the curriculum (and standards) is assured. Likewise, test instructions for students scoring directions for teachers needs to be clear and unbiased. These forms of bias may be revealed in the way the test is administered or in the results that may indicate items or groups functioning differentially.

Nevertheless, success in virtually all real-world endeavors requires multiple skills and abilities, which may interact in complex ways. Testing programs typically address only a subset of these. Some skills and abilities are excluded because they are assessed in components of the selection process (e.g., completion of course work or an interview); others may be excluded because reliable and valid measurement is economically, logistically, or administratively infeasible" (p. 80). Ironically, the nature of alternate assessments may provide the very source of

fairness that would normally not be possible with the mass testing used in general education. In the end, using individually administered tests with extensive interpersonal interactions (much like mirroring most other social environments in the family, work, and community), it may be possible to develop a testing program that is fair and equitable.

Summary and Recommendations

1. The test design needs to begin with clear articulation of the constructs, which are usually developed from the grade level content standards.
2. With these constructs in mind, the test can be developed using any number of formats. Principles of universal design should be used to make sure the format is appropriate for students with disabilities.
3. Bias review needs to include content experts as well as others from the special education community to ensure the measures are an appropriate rendering of content and accessible to all students. Other obvious issues of bias related to race and gender need to be considered as well.
4. The test needs to be aligned to the content standards using some formal system that considers multiple dimensions of categories and representations.
5. Organizing the test itself should provide users clear directions for administration and scoring.
6. Ensure that both offices of special education and assessment-evaluation are working together and using each other's expertise as well as others in the field. An excellent resource is information from Downing (in press).

References

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

Downing, S. M. (in press). Twelve steps for effective test development. In S. M. Downing and T. M. Haladyna (Eds.), *Handbook of test development.* Mahwah, NJ: Lawrence Erlbaum Associates.

Haladyna, T. M. (1999). *Developing and validating multiple-choice test items*. Mahwah, NJ: Lawrence Erlbaum.

Haladyna, T. R. (2004). *Developing and validating multiple-choice test items-2nd edition*. Mahwah, NJ: Lawrence Erlbaum.

Pullin, D. (1983). Debra P. versus Turlington: Judicial standards for assessing the validity of minimum competency tests. In G. Madeaus (Ed.), *The courts, validity, and minimum competency*. Hingham, MA: Kluwer-Nijhoff Publishing

Resnick, L. B., Rothman, R., Slattery, J. B., & Vranek, J. L. (2003-2004). Benchmarking and alignment of standards and testing. *Educational Assessment*, *9*(1 & 2), 1-27).

Thissan, D., & Wainer, H. (2001). *Test scoring*. Mahwah, NJ: Lawrence Erlbaum.

Thompson, S. J., Johnstone, C. J., & Thurlow, M. L. (2002). *Universal design applied to large-scale assessments* (Synthesis Report 44). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved [today's date], from the World Wide Web: http://education.umn.edu/NCEO/OnlinePubs/Synthesis44.html

Tindal, G. (2005). *Alignment of alternate assessments using the Webb system: An abbreviated version*. Washington, DC: Council of Chief State School Officers.

Tindal, G., Cipoletti, B., & Almond, P. (2005). *Alternate assessments: Evidence based on Test Content and Alignment with Standards*. Eugene, OR: University of Oregon, Behavioral Research and Teaching.

Webb, N. L. (1997). Determining alignment of expectations and assessments in mathematics and science education. *NISE Brief.* National Center for Improving Science Education, University of Wisconsin- Madison.

Webb, N. L. (1999). *Alignment of science and mathematics standards and assessments in four states. (NISE Research Monograph No. 18).* Madison, WI: University of Wisconsin-Madison, National Institute for Science Education. (ERIC Document Reproduction Service No. ED440852).

CHAPTER 7: EVIDENCE OF ITEM QUALITY

Introduction

Given the various assessment approaches and participation options described in Chapters 3 and 4, questions arise about how to assemble evidence that supports the quality of items. Chapter 3 of the *Standards* (AERA, APA, & NCME, 1999, pp. 37-48) provides complete details about steps in test development and revision, including assessing and documenting the quality of items to be included in a test. This chapter includes techniques for investigating item quality for typical statewide assessments as well other assessment formats that might be used in large-scale assessment of students with disabilities. What constitutes an assessment "item" is clearer in typical large-scale assessments than in some of the other assessment approaches described in this volume. Achievement tests have historically been comprised of selected response (SR) items, such as multiple-choice, or short constructed response (CR) items, such as a brief writing sample (Haladyna, 1999). However, for students with significant disabilities,

> "...an item needs to be more broadly interpreted as an opportunity to respond or a 'behavioral event' or performance that can be evaluated as more complex than simply correct or incorrect. This 'behavioral event' or performance also needs to be part of a complete cycle of teaching and learning that relies on a measurement system with internal integrity and yet fits into a statewide testing program," (Tindal, Cipoletti, & Almond, 2005, p. 14).

Several approaches within the range of assessment options for students with disabilities employ some type of CR items such as portfolios that contain collections of student work, observations of student performance (performance events or performance tasks), and observation checklists. CR items are either high-inference performance measures that require a rubric and are

subjectively scored, or low-inference performance measures that require observation and are objectively scored (Haladyna, 1999).

The techniques described in this chapter serve as sources of procedural and empirical validity evidence. First, indicators of the technical quality of items are introduced and applied to a range of assessment types, from classical and modern test theories. Second, methods are presented for collecting and evaluating validity evidence based on item content. Third, important questions are highlighted when modified test items are used. Fourth, issues are examined with the use of existing cross-grade items and computer-adaptive testing (CAT) for assessments based on modified achievement standards. In all four of these analyses, new questions are posed related to the use and interpretation of item quality indicators. Questions about the quality of individual assessment items should be answered before the technical adequacy of the entire instrument and validity evidence to support inferences about the assessment scores are considered. This chapter focuses exclusively on evidence of item quality that might be collected *once items are written*. The methods suggested here should be considered in conjunction with guidelines for test design (Chapter 6), reliability evidence (Chapter 8), and validity evidence collected at the assessment level (Chapter 9), in evaluating the quality of assessments for students with disabilities.

I. Statistical Indicators

Along with the movement toward assessing higher-order thinking skills and increased flexibility in reaching students with significant disabilities, assessment programs have broadened the mechanics of assessing via SR items to include CR items or tasks. This section of Chapter 7 covers five major indicators of technical adequacy that can be used as empirical validity evidence at the item level. Given the vast spectrum of item types and test delivery modes, not all

indicators are applicable to all items. Some indicators may be appropriate for SR items whereas others may be more suitable for CR items.

The first two indicators pertain to appropriateness of testing time and item difficulty. The third indicator assesses the statistical relevance of the item as part of the total assessment. The fourth indicator describes the degree to which performance on the item varies considerably from one group of students to the next (an issue of fairness). Lastly, the fifth indicator addresses the level of local item dependency (LID) among items. LID may occur when several items are tied to a common paragraph (such as in reading) or a theme (such as in math). LID may also be a concern in some performance events, when a series of items are scored based on the level of independence with which the student can respond correctly. States will make judgment calls about the adequacy of the statistics obtained using these indicators in the context of their unique, large-scale assessment programs.

Testing Time

For assessments that are timed, it is important to be sure most students have enough time to complete the test/task. In other words, we want to be sure that there are not many missing responses. Extending the testing time is allowed in an unrestricted manner in 37 states and 32 states allow it with restrictions; some states allow it under both conditions (Thurlow & Bolt, 2001). This situation pertains to the case of students under regular test administration conditions as well as students with disabilities who are given a finite (such as an additional 50%) time accommodation. With enough test takers, such indicators are reflected in the number of items left blank or unanswered at the end of a standardized testing session. These indicators are typically used to either shorten the test or to give students longer testing time.

In some situations that involve accommodations such as Braille or large print, the allotted testing time may not be appropriate primarily due to the very nature of and logistics surrounding the accommodation. Examples of these conditions are setting and presentation accommodations (Thurlow & Bolt, 2001). Other tests may not have time limits on administration sessions, although there may be limits on the window of testing time due to test security policies. However, for assessments that are not timed (e.g., portfolio approaches), time factors associated with these logistical issues are psychometrically irrelevant.

Item Difficulty

Under normal circumstances for standardized testing, an optimum assessment is one that is comprised of items that are at the ability level of the test taker. For typical, large-scale achievement testing, most items tend to be answered correctly by 40% to 60% of all students. For achievement tests that assess specific content standards and primarily have criterion-referenced score interpretations, content appropriateness and item quality as judged by content experts takes precedence over item difficulty (AERA et al., 1999, p. 40). For example, an arithmetic test used to assess basic operations must contain items on division regardless of the difficulty level of these items.

For binary items, item difficulty is traditionally assessed by the proportion of examinees who respond correctly (item p-value). For partial credit items, the item p-value is a ratio (mean/MPS), where the mean is the average item score for examinees and MPS is the maximum possible score on the item. For standardized, content-based assessments such as the South Carolina Palmetto Achievement Challenge Tests (PACT; Huynh, Meyer, & Barton, 2000), the mean p-value ranges from 47% to 70% for binary SR items and from 26% to 65% for CR items.

In many state and national testing programs, common (fixed form) tests are often used. When forms are designed for students with varying levels of ability, it often is recommended that items be selected to cover the range of student ability. For example, the test may be comprised of a portion of items considered to be "easy," a portion "challenging" or difficult, and a group of moderately difficult items. Table 7.1 illustrates this range of average p-values for CR items on one state's large-scale assessments in fourth grade.

Table 7.1. Major Indices for Constructed Response and Extended Response Items on Fourth Grade Palmetto Achievement Challenge Tests (South Carolina Department of Education, 2003)

			Mean	
Subject	**Type of Item**	**Number of Items**	***p*-value**	**Item/Criterion Correlation**
ELA	Constructed response	3	0.507	0.562
ELA	Extended response	4	0.728	0.677
Mathematics	Constructed response	4	0.301	0.550
Science	Constructed response	2	0.399	0.511
Social studies	Constructed response	1	0.269	0.561

Of course, judgments about item difficulty must be made with knowledge about the characteristics of the students tested, as a group with unusually high or low ability may make items of average difficulty appear easier or challenging than they are. An alternate assessment based on modified achievement standards may, for example, include a larger proportion of "easy" items that reflect less complexity.

When new content standards are introduced, p-values for assessment items corresponding with those standards may be somewhat low initially. P-values also may be initially lower than expected when groups of students who have been excluded from testing or assessed using other methods are included in an assessment for the first time. However, we would expect p-values to move to a typical range once instruction and assessment are aligned with the new standards.

Scores for students should improve in succeeding years as students' instruction in the new standards and experience with taking tests increases. For this reason, changes in the breadth, depth, and complexity of "items" may need to be reconsidered after their initial use as the expectations may be shifting (McGrew & Evans, 2003).

When assessment scores are based on the level of independence with which students can complete a task or give a correct response (such as in some performance-based alternate assessments), one simple indicator of item difficulty might be the frequency distribution of scores on each item. For example, if an item is scored on a four-point scale, with the maximum number of points being awarded for an independent, correct response and the lowest number of points representing direct physical assistance to respond, the frequency distribution of each point value would allow stakeholders to understand the proportion of students who were able to respond with independence or with varying degrees of support. P-values for partial credit items also may be calculated as described above.

The partial credit p-value was used in a field test of Colorado's performance-based alternate assessment, where the results were used to assess individual item difficulty as well as adjust the length of the assessment (Fahey, Filbin, & Connolly, 2004). While these are appropriate uses of item difficulty indicators for alternate assessment, it is important to remember that responding only with physical guidance should not be considered proficient in alternate assessments using alternate achievement standards because this level of assistance requires little to no understanding of the assessment task (see Chapter 5).

Statistical Item Relevance

To contribute to assessment of the specified content standards, each item should show some level of relevance with the other items of the total assessment. In other words, there should

be some level of correlation among the items as well as between a single item and the subtest defined by the other items. These correlations contribute to measures of internal consistency (or test reliability) such as the KR20 or Cronbach alpha indices. However, judgments about items made based on these internal consistency measures should be made with caution when the items measure a construct that is not unidimensional, as the indices may underestimate reliability (Cortina, 1993). Chapter 8 provides more detailed information about reliability measures.

Correlations among items are typically assessed via the inter-item correlation matrix. Item/test correlation is often referred to as an index of item discrimination. This can be reported as a biserial correlation or point biserial correlation for binary items. For partial credit items, it can be the regular Pearson correlation or the poly-serial correlation. Negative or near zero values for correlations are not acceptable. For binary items of most standardized tests, point biserial correlations are typically above 0.20. Higher threshold values are used for biserial or poly-serial correlations. It is important to note that biserial correlations may be higher than point biserial correlations, simply because of the formula used (Crocker & Algina, 1986).

The above correlational statistics may be computed for checklists and for performance assessments consisting of CR items that measure discrete competencies and are scored with analytic rubrics. However, measures of statistical item relevance may not be computed when the entire assessment (such as a portfolio or an extended response item in writing) is scored with a holistic rubric that yields a single score. While there are no “items” on this type of assessment, reviews of portfolio contents may be conducted to confirm the relevance of the evidence to the achievement standards upon which the portfolio is based.

Differential Item Functioning

Differential item functioning (DIF) is a statistical indicator that specifies the degree to which performance on the item varies considerably among groups of students with similar ability. DIF alerts to the potential lack of fairness in using the test item for all students and may be used to determine whether student background variables may be introducing construct-irrelevant variance to the assessment. DIF should at least be checked for some major reporting groups with enough cases for the test to hold. Reporting groups are typically based on gender, ethnicity, and socioeconomic group. Groups also may be formed based on disability labels; however, for students with disabilities, and especially those with significant cognitive disabilities, various reporting groups may not be large enough for traditional large-sample DIF methods (such as the Mantel-Haenzel chi-squares used in NAEP) to work. A list of DIF procedures (including Item Response Theory [IRT] and logistic regression) appropriate for moderate to large samples is provided in Clauser and Mazor (1998). DIF methods based on generalizability theory (Brennan, 1992; Shavelson & Webb, 1991) have also been proposed. Nonparametric DIF procedures for small-sample situations are addressed by Meyer, Huynh, & Seaman (2004). An exposition of Rasch-based procedures that are also appropriate for small samples can be found in Bond and Fox (2001, pp. 170-171). Although we are personally aware that small-sample Rasch-based methods have been used with some level of success in testing programs including the South Carolina High School Alternate Assessment (HSAP-Alt; SCDE, 2005) and the Michigan Alternate Assessment Program (M. Rodriguez, personal communication, July 14, 2005), to our knowledge, no *published* reports exist that apply these small-sample methods to assessments based on modified or alternate achievement standards. Thus, there is not yet an empirical basis for criteria used to identify DIF using small samples in these assessments.

A presence of DIF does not necessarily mean the item is unfair, as fairness judgments are *ultimately* rendered through content examination by experts. For large-scale assessments, an expert panel is typically convened to examine items with substantial DIF values to determine if the items are inherently unfair against the student group under testing. Decisions of these panels are typical taken as the *final decisions* on the items.

Local Item Dependency

This section discusses the level of local item dependency (LID) among items. LID may occur when several items are tied to a common paragraph (such as in reading) or theme (such as in math); in writing assessments where intra-rater ratings are highly correlated due to potential halo effects; or in performance-based alternate assessments when an assessor individually administers a series of tasks to a student. For example, one of South Carolina's alternate assessments features a series of performance tasks administered by a teacher (South Carolina Department of Education, 2005). Teachers use standardized scripts to determine the proper series of prompts and responses to students' answers to give students opportunities to respond with the least amount of assistance possible. Scores are assigned based on the level of assistance required to make a correct response. LID is a potential concern in this situation because the assessor's judgment of correctness and independence are embedded within the student's score.

Estimation of student achievement is statistically optimized when a test is comprised of items that do not "cue" to each other. This property is often referred to as "local item independence," an essential condition in applications of item response theory (IRT). Items that are locally dependent essentially share common information about student ability in some overlapping manner. Such item overlapping would render the test less efficient (in terms of test

administration and scoring time) if both items were used in the assessment. An optimum goal in creating an efficient assessment is to select items that are as least locally dependent as feasible.

When the student populations are large enough to allow IRT calibration, LID may be probed via indices such as Yen's Q3 (1993). Other indices based on partial correlations such as those used by Ferrara and Huynh and their associates (1997, 1999) in the context of the Maryland School Performance Assessment Program (MSPAP) may be used.

II. Validity Evidence at the Item Level

Validity evidence is collected from a variety of sources, to support the inferences made from test scores. This section focuses on validity evidence gathered at the item level, rather than the test level. Miller & Linn (2000) identified six aspects of validity, two of which feature evidence collected at the item level. *Content*-related evidence for standardized assessments typically comes from state standards, while *substantive* evidence is based on the cognitive processes required to complete the item.

Content-Related Evidence

The primary question for assessing content match is *to what extent does the item reflect the content standard?* The *Standards* (AERA et al., 1999, Standard 3.11) call for investigations of content coverage to make inferences about the extent to which performance on test items generalizes to the larger domain of interest (the content standard). As modified and alternate achievement standards now allow assessment items with changes the breadth, depth, and complexity of the content assessed, careful documentation of content coverage is necessary to support the kinds of score interpretations appropriate for each of the assessment options in the model.

Traditional investigations of content-related validity evidence focus on stakeholder ratings of the degree to which items reflect the intended domain -- in this case, the state standards. For example, general assessment items used in South Carolina's PACT tests are reviewed for "alignment and quality" (South Carolina Department of Education, 2003, p. 30) by staff from the company that develops the test, by State Department of Education staff, and by educators and content specialists prior to use. With SR and low-inference CR items, raters may assess all the assessment items. With high-inference CR assessments (i.e., portfolios and some states' performance assessments), ratings may be based on samples of items or portfolios. As discussed in Chapter 8, care must be taken in the sampling of students to ensure a reliable result.

Existing validity studies on alternate assessments have relied on expert and stakeholder judgments. In a study on Washington's portfolio-based alternate assessment, Johnson & Arnold (2004) collected validity evidence by having raters review portfolios and determine whether they contained links to the appropriate content standard in four areas. Three-fourths of all portfolios were categorized as being connected to state content standards in some way, although the connection was based on the teacher's interpretation of the content standards and may have been superficially linked. Browder, Flowers, Ahlgrim-Delzell, Karvonen, Spooner and Algozzine (2004) used a combination of surveys and focus groups, with reviewers including content area experts, experts in severe disabilities, and special education teachers and administrators (stakeholders). Reviewers assessed samples of performance indicators from 31 states' alternate assessments for their alignment to national standards (e.g., National Council of Teachers of English, National Council of Teachers of Mathematics), access to the general curriculum, reflection of functional skills, and age appropriateness. Nearly all the math experts (86%) and 70% of stakeholders said some states' alternate assessments were aligned with national standards

in math. While there was not agreement on *which* states' alternate assessments were aligned to national language arts standards, 86% of language arts experts and 100% of stakeholders found some states' alternate assessment performance indicators were aligned. Reviewer opinions were mixed when it came to identifying states that were very well (or very poorly) aligned to national standards.

Both the Johnson and Arnold (2004) and Browder et al. (2004) studies were based on global judgments of content match. More precise methods of assessing content match may be found in alignment procedures, such as those described by Bhola, Impara, & Buckendahl (2003). (See Chapter 9 for a more detailed description of alignment procedures, including statistical calculations.) One method of investigating alignment between expectations for student knowledge and skills, and general education assessments was developed by Webb (1997) in the context of math and science. In Webb's approach, the degree of match between state content standards and assessment items is referred to as *categorical concurrence.* Webb (1999) suggested that having at least six items match each content standard is an acceptable degree of categorical concurrence. This guideline was based on consideration of reliability of the scale, the mean scale score, and cutoff score used to determine mastery on that scale. What is considered an "acceptable" number of items may vary depending upon the test format (fixed vs. adaptive), and by score use and reporting purposes. When scores are reported for individual students, and when subs cores are reported for strands within content areas, more items may be needed to sufficiently sample the intended domain (M. McCall, personal communication, July 13, 2005).

In an analysis of four states' mathematics and science assessments, an acceptable level of categorical concurrence was found for as few as 38% of standards on one test, to 100% of standards on another test (Webb, 1999). Acceptable levels of categorical concurrence according

to Webb's criteria may be difficult to obtain for alternate assessments. With fewer items on the assessment, many alternate assessments do not have six items per content standard to begin with. In recent studies on the alignment of alternate assessments with state standards, evidence of categorical concurrence was mixed. On Wisconsin's Alternate Assessment, based on teacher rating scales, Roach, Elliott, & Webb (2005) found the lowest percent of standards with acceptable levels of categorical concurrence in science (13%) and the highest in reading (100%). An analysis of three states' alternate assessments revealed greater degrees of categorical concurrence for performance-based assessments in reading and math, while assessments that had fewer items tended to have lower percentages of acceptable categorical concurrence (Flowers, Browder, & Ahlgrim-Delzell, in press). In that study, categorical concurrence ranged from 0% to 67%.

Substantive Evidence

Substantive validity evidence refers to the cognitive demands made on a student when completing an assessment task. Analyses of the examinee's response processes "can provide evidence concerning the fit between the construct and the detailed nature of performance or response actually engaged in by examinees" (AERA et al., 1999, p. 12). State content standards often include indicators of cognitive demand (e.g., "analyze," "explain"), and assessment items can be reviewed to determine the cognitive behavior required to respond to the item.

In Webb's (1999) alignment model, cognitive demand is referred to as depth of knowledge (DOK). DOK is rated on a four-point scale, ranging from recall and reproduction (level 1) to extended thinking (level 4). Webb (1999) suggested that at least 50% of the items corresponding to each standard should be at or above the level of cognitive demand required by the standard. This criterion is "weakly" met if that figure falls between 40% and 50%. The 50%

guideline was based on the assumption that students would have to get at least half of the assessment items correct in order to be proficient on that standard, and students would need to be able to work at or above the level of cognitive demand represented by those test items in order to get them correct.

Despite the 50% guideline, both general and alternate assessments have been found to have a wide range of DOK matches. Webb's (1999) analyses of four states' science and mathematics tests found acceptable DOK match at the standard level to range from 0% to 100%. Using Webb's 4-point scale, Herman, Webb, & Zuniga (2005) found the average rating on a statewide high school mathematics exam averaged only 1.6. Recent alignment studies on alternate assessment have also indicated that the assessment objectives tend to have lower levels of cognitive demand than the standards to which they are aligned (Flowers, Browder, & Ahlgrim-Delzell, in press; Roach et al., 2005).

As Johnson & Arnold (2004) note, the actions of teachers largely determine what is included in portfolios. When examining substantive evidence for portfolio assessments, teachers' response processes as well as the students' must be considered. For example, some teachers are better prepared to design tasks or arrange conditions that require students to demonstrate greater levels of cognitive demand, whereas other teachers may limit student response opportunities to reiteration of factual knowledge (e.g., names of state capitals) or demonstration of basic skills (e.g., mathematic calculations in absence of mathematical application). Other effective instructional practices (e.g., generalization of a skill to other materials or contexts) can increase the cognitive demand on students.

At a more fundamental level, if teachers do not include evidence that matches the skills the portfolio is supposed to represent, raters are not able to make sound judgments about student

performance. Standardization of portfolio contents and assembly procedures, and effective professional development for teachers on the link between instruction and the portfolios, should serve to increase expectations for student performance as well as improve the quality of information reviewers may use to assess cognitive demand. For example, Johnson and Arnold trained teachers to focus on the alignment of task objectives to standards, and on recording behavior that served as evidence of cognitive attainment. They also provided teachers with a checklist for designing and recording tasks (M. McCall, personal communication, July 13, 2005).

While alignment studies provide some evidence of cognitive demand, they rely on reviewer judgment of the students' response processes. Think-aloud procedures, in which students talk about the process by which they reach an answer to an item, provide more direct substantive evidence. Think-aloud procedures are described in more detail in Chapter 9.

Item Quality Evidence from Curriculum Alignment

While validity evidence at the item level focuses on the relationship between test items and the standards they are intended to measure, it is important to remember another element in students' education: the curriculum. If tests are aligned with standards, but not with what is taught, students' scores will not reflect their growth on the intended curriculum, and the inclusion of those scores in accountability formulas will not be tenable. Curriculum guides have traditionally been available to help general education teachers align instruction with content standards and typical standardized assessments. However, special educators are still learning how to access the general curriculum, especially for their students with the most significant disabilities. In a review of alternate assessment portfolios conducted in 2001-02, Johnson and Arnold (2004) counted the number of content areas in the alternate assessment that contained

evidence matching students stated IEP skills. Roughly half of the portfolio evidence (ranging from 50% in Communication to 55% in Reading) matched the stated IEP skills.

While IEPs do not reflect the entire curriculum for students with disabilities, they are one source of information about what students receiving special education services are taught. Especially for students with significant cognitive disabilities, IEPs may contain objectives that show how students are accessing the general curriculum. Karvonen & Huynh (2005) described a method of analyzing IEPs for evidence of alignment based on content, depth of knowledge, expected level of performance, and other indicators. Some states are addressing curriculum-assessment alignment directly, through guidance to the IEP team on aligning IEP objectives with state standards reflected in the alternate assessment (Idaho State Department of Education, 2004). However, caution should be exercised in using the IEP as a proxy for the taught curriculum, as IEP contents are often unstandardized and the full academic curriculum that is taught may not be reflected in IEP goals. Another source of evidence about curriculum alignment may come from extending methods developed in general education for examining both curriculum materials and instructional practices (Porter, 2002).

Other Criteria for Rating Assessments

Curriculum for students with the most significant cognitive disabilities has for many years emphasized functional life skills; the shift to academics in response to IDEA 1997 and NCLB has not eliminated the values that teachers, students, and parents have for a comprehensive education. Evidence of these values is seen in a 2003 national survey of state special education directors, who indicated that alternate assessments were scored not only on the basis of independent performance of the academic and functional skills, but also on criteria such as generalization of the skill to other settings, age appropriateness, self-determination, social

relationships, and inclusion in general education settings (Thompson & Thurlow, 2003). As described in Chapter 5, we recommend that scoring be based on student performance which can include generalization of skills and responses that reflect self-determination (e.g., choice making, problem solving). Social relationships and inclusion in general education settings are important forms of support that should be considered in inferences made about the students' score. For example, if the assessment of a student with significant cognitive disabilities is conducted with a peer model, a conservative assumption is that the student will need similar models to perform at this level in the future. Just as students with mild disabilities receive testing accommodations that are consistent with the accommodations they receive throughout the year, students with significant cognitive disabilities who receive supports during alternate assessments should have opportunities to use similar supports during instruction throughout the year. It is important to note that these supports should not be used as flags on student scores used for other purposes, as has been the practice on score reports of students tested with accommodations.

Item Reviews (Including Bias and Sensitivity)

During the design and initial implementation of test items, it is recommended that test developers have items (test specifications, including content and cognitive demand) and response formats reviewed. Item writers' content expertise, backgrounds, and item writing experience should be documented, along with a description of the item writing training that is provided (Downing & Haladyna, 1997). More detailed information about item writing may be found in chapters 6 and 11. Once items are written, professional editing helps uncover inconsistencies, vagueness, and other problems with content that may confuse the test taker and introduce construct-irrelevant variance to the testing situation (Downing, in press).

When students with disabilities participate in an assessment, it is important to include both reviewers with a special education background and content specialists who understand the state content standards across grade levels. Specialists in assistive technology should also review items that may be accessed through adaptations (e.g., Braille, enlarged print), or adaptive or alternative communication systems. This review process provides an additional check on student background variables that potentially introduce construct-irrelevant variance and weaken arguments for validity of the assessment scores.

Baranowski (in press) provides a thorough description of writing and review procedures for multiple-choice items. These procedures can be applied to open-ended performance tasks and portfolio assessments as well. Many of the features from universal design also can be considered in the development of items: (a) inclusive assessment population; (b) precisely defined constructs; (c) accessible, non-biased items; (d) amenable to accommodations; (e) simple, clear, and intuitive instructions and procedures; (f) maximum readability and comprehensibility; and (g) maximum legibility (Thompson, Johnstone, & Thurlow, 2002). When employing universal design principles to develop or modify assessment items, it is important to be sure the underlying construct is not changed in the process.

Related to questions about validity evidence is the issue of bias based on cultural background (e.g., race, ethnicity, nationality, language, gender, ability/disability, socioeconomic status). Zieky (in press) provides clear examples and nonexamples of bias in item content based on cultural background, as well as a thorough overview of procedures for conducting fairness reviews based on written guidelines. While potential item bias should be investigated through a systematic review process, it is also helpful to train item writers on writing unbiased items and select item writers with diverse backgrounds (Downing & Haladyna, 1997).

Items on an assessment may be well aligned with content standards, yet inferences based on scores for members of underrepresented groups may not be tenable (Crocker & Algina, 1986). Although reviewing items for content match and sensitivity is an important step in collecting validity evidence at the item level, this review process is not sufficient to claim items are aligned and free of bias. When student response data are available, a statistical analysis such as DIF may be useful in alerting to items where there is potential for bias (see section 1 of this chapter). It may be recalled that a presence of DIF does not necessarily mean the item is unfair, as fairness judgment is *ultimately* rendered through content examination by experts.

Selecting and Training Reviewers

Those who have conducted validation studies with expert reviewers' caution against common pitfalls in the selection and training of those reviewers. For example, some stakeholders (teachers) may be overly generous in finding items to be aligned to standards (Bhola et al., 2003). Reviewers with different backgrounds (e.g., educators, test publishers) may also disagree about the degree of alignment between test items and standards. Those conducting the study should ensure the "experts" truly possess content expertise, provide adequate training and corrective feedback, make rating systems easy to use, and allow frequent breaks to avoid rater fatigue (Sireci, 1998). Care should be taken in conducting sufficient training, especially in areas where raters may have difficulty identifying problematic items. For example, raters may be more accurate at identifying cultural flaws than technical flaws (Engelhard, Davis, & Hansche, 2000). As with all types of large-scale assessments, validity investigations that rely on rater judgment should clearly report rater selection procedures, qualifications, and rating tasks (AERA et al., 1999, Standard 1.7).

III. Considerations in Using Modified Items

In the context of alternate assessment based on modified achievement standards, items aligned to content standards at a certain grade level may be *modified* for use with students who are assessed based on these modified standards. The modified items are essentially new items and therefore should be subjected to the same procedures for determining technical adequacy (See Section I on statistical indicators) and content match (See Section II on validity at the item level) as new, on-grade assessment items.

When essential concepts from content standards evolve gradually across grades, there is no clear distinction between items written for a given grade ("on-grade items") and items intended for many grades ("cross-grade items"). An example of this situation may be seen in the consensus framework based on the content standards from nine member states of the Alternate Assessment Collaborative (AAC) (Bechard, 2005). In this consensus framework, curriculum specialists used organizational frameworks from national content organizations to build a matrix and arrange key standards by essential concepts into a common document. The matrix displayed the grade levels in which topics were introduced. The development of a concept within that framework across grade levels, as seen in Table 7.2, may inform development of modified items.

Table 7.2. Evolution of Comprehension Strategies/Reading Process Benchmarks Across Grade Levels (Bechard, 2005)

Grade 3	***Grade 4***	***Grade 5***	***Grade 6***	***Grade 7***	***Grade 8***	***HS***
Make connections between text and own life, and within text	Make connections between text and own life, within text, and across texts	Make connections between text and own life, within text, and across texts	Make connections between text and own life, within text, and across texts	Make connections between text and own life, within text, across texts, and between text and world	Make connections between text and own life, within text, and across texts between text and world	Make connections between text and own life, within text, and across texts between text and world

Note. The bold text highlights where greater complexity of the concept first appears.

Simplifying content as might be done in tests built on modified achievement standards could make an item difficult to classify as a grade 4 or grade 5 item since the statement of content is identical for both grades. As described in Chapter 5, this simplified content can be linked to grade level content at the item level through applications to grade level materials and activities in developing the assessment task. For example, does the context within which the item is set align with the student's chronological age so that high school students encounter characters, plots, and settings of interest to themselves and other high school students even when the reading difficulty is below grade level? Or, at the test level, this linkage may be demonstrated through other items created for grade level specific content. Pulling items from assessments at lower grade levels in developing alternate assessments may be appropriate if they are upgraded to have these appropriate grade level links. Simply recycling items from a lower grade level or giving the full assessment from a lower grade level would not be appropriate.

As described in earlier chapters, these items are aligned with grade level content standards but may be modified by reducing the complexity, breadth, or depth of the content. These items also may be revised to feature age-appropriate content, materials, or contexts. Modified items are *not* "cross-grade" items taken from a test written for a higher or lower grade level than the one in which the student is enrolled. The use of cross-grade items is discussed in the next section.

IV. Considerations in Using Cross-Grade Items

In several state assessment programs, test forms are specifically created for each grade to tap on-grade content standards and then linked through a vertical scaling/linking process so that student growth can be tracked longitudinally. In this situation, it is feasible to use existing items

originally intended for one grade to assess students at other grades. These items are referred to as "cross-grade" items in the remainder of this chapter.

The use of cross-grade items may be appropriate where (a) assessments (such as in reading or math) have been vertically linked to cover the common cross-grade core of the curricula, and (b) the modified achievement standards do not constitute a major breach of the construct being assessed. Under these two conditions, the assessment closely resembles "computer-adaptive testing" (CAT; Weiss, 1983), in which items are selected sequentially to closely mirror the ability of the student. One well-known CAT system is the Measures of Academic Progress (MAP) developed by the Northwest Evaluation Association (NWEA, n.d.). CAT requires that all items be calibrated and linked to a common latent trait (construct), such as reading or mathematics, which needs to be assessed. The major difference between use of cross-grade items (as described in this section) and CAT lies in the fact that the modified achievement standards may require that the cross-grade items be in some *pre-defined content strands,* whereas in CAT items are (theoretically) allowed to spread out *in some pre-specified algorithm* across the entire content spectrum. The difference mentioned here may be mediated to some degree by requiring several checks to ensure the acceptability of cross-grade items as part of a modified assessment system. The rationale for determining which subsets of items will be administered to adequately assess certain content should be well-documented (AERA et al., 1999, Standard 3.12).

Many assessment programs have been designed to measure students' achievement in a specified subject area like reading and mathematics, and to track their progress across several grade levels. One example is the *Iowa Tests of Basic Skills* (*ITBS*; Hieronymus & Hoover, 1986). Within the *ITBS* context, Petersen, Kolen, and Hoover (1989) indicated that "the content

represented in all levels of a test from an elementary achievement test battery can be viewed as defining a *developmental continuum* for a particular area of achievement" (p. 231, emphasis added). A developmental scale, therefore, may be appropriate for this situation. The construction of such a scale requires a statistical process called "vertical scaling" or "vertical linking." In the construction of an item bank for test form construction, vertical scaling brings all items together on a common scale. Theoretically, any item in the operational form may be replaced by another easier or more difficult item without changing the construct under measurement. Although tests over a wide range of grades can be technically linked vertically, their use is typically constrained to yearly growth measures from one grade to the next.

The discussion about content overlap between grades and vertical scaling for tests leads to the expectation that some items that were originally created to assess students at one grade level may be used for students at a different grade level. However, before this can be done, the following questions need to be addressed. The first two questions are posed within the context of item alignment in which an assessment is made of the degree to which cross-grade items can be used for an assessment. The questions also are posed mainly for assessments that are based on traditional SR or short CR items administered under standardized conditions (with or without accommodations).

Question 1. Do content standards form a single construct (dimension) for all grades?

This question needs to be asked if *any* cross-grade item can be used in a modified assessment. As pointed out in the example from the AAC (Bechard, 2005), organizational frameworks from national content organizations can provide a unifying structure for sorting out the content. Since the content structure of a domain is typically operationally defined by a test plan, answering this question may require empirical as well as statistical investigation. For tests

that are vertically scaled, the very strong assumption of unidimensionality across grades (the full range of the vertical scale) is made. Thus, incorporating multiple grades' content within a given, single test plan may help to meet such assumptions.

The assumption of a common construct over a large span of grades is many times very hard to meet. Therefore, the use of cross-grade items that are originally written to assess performance several grades lower may not be totally justified. However, as mentioned previously, commercial tests and many state assessments are vertically linked over many grades; vertical scores are often used only to track yearly progress between adjacent grade levels (Patz & Hanson, 2002). This leads to the expectation that it might be easier to justify the use of cross-grade items taken from adjacent grade levels. For this expectation to hold, the following, less restrictive question may be asked.

Question 2. Is the assumption of unidimensionality more appropriate for adjacent grades?

By defining a test plan on content of adjacent grades, rather than a full range of multiple grades (i.e., grades 3 – 10), the assumption of unidimensionality might be much more easily met. Often, within even a three-grade span (e.g., grades 3, 4, and 5), content tends to overlap across grade specific tests/tasks and test plans. Given that fact, validating the assumption of unidimensionality may prove far easier when evaluated qualitatively as well as empirically and statistically.

There are a variety of methods by which the assumption of unidimensionality can be probed. Item response theory (IRT) provides the bulk of these methods. Other classical (non-IRT) methods include use of the screen test based on either principal component analysis or factor analysis. When IRT models for evaluating items are used by test developers, it is

important to document the model and estimation procedures used, and evidence of model fit (AERA et al., 1999, Standard 3.9).

As mentioned previously, the above two questions about unidimensionality are posed mainly for assessments that are based on traditional SR or short CR items administered under standardized conditions (with or without accommodations), judged either against grade level or modified achievement standards. How unidimensionality factors into performance assessments, collections of evidence, observations, and other assessment approaches is a different question to answer. In some cases, a common dimension may be assumed if these types of CR items are scored using a set of *rubrics* that are *identical* or *qualitatively equivalent* over a relevant grade span.

Once the assumption regarding unidimensionality is settled, questions may be asked regarding the similarity of technical characteristics of the same item when it is administered under on-grade and cross-grade conditions. Granting that a group of students taking assessments with cross-grade items can be found who share the same ability as the group of students taking the on-grade assessment (but who may differ on other characteristics such as disabilities and accommodations), the following additional question, consisting of five additional subtopics (See Section I) may be posed.

Question 3. Does the item behave similarly on the following technical characteristics (as appropriate) for students of similar ability as reflected in the total test score on the same items: appropriateness of testing time, appropriateness of difficulty, relevance of the item to the entire assessment, relevance of item within curriculum and instruction, DIF, and local item dependency (LID)?

Answers to the DIF question would be found by comparing the responses of the group of students tested with cross-grade items to the group of students tested on similar items at the appropriate grade level. Question about each of the other criteria may also be answered using the procedures described earlier in this chapter.

General Remarks on the Use of Cross-Grade Items

The above three questions are asked to gauge the degree to which a cross-grade item is suitable for an assessment judged against modified achievement standards. Along with other items (including modified items discussed in Section III) probing grade level content standards, the suitable cross-grade items play a major part in enlarging the item bank for the construction of such assessment. It is important for test developers to carefully modify grade-level items to reflect the modified achievement standards and subject those items to scrutiny based on construct- and content-related validity evidence. Simply substituting cross-grade items for an assessment based on modified achievement standards is not an acceptable basis for inferences made about these assessments. As item pools grow, it is important for test developers to document procedures for developing, reviewing, and trying out these items (AERA et al., 1999, Standard 3.7); and monitor the stability of the scale on which scores are reported across test forms (AERA et al., 1999, Standard 4.17).

General Remarks about Computer-Adaptive Testing for Assessment with Modified Achievement Standards

In computer-adaptive testing systems (Weiss, 1983) such as the MAP of the Northwest Evaluation Association (NWEA, n.d.), item developers think in terms of all students in a grade. There are some items that address content described in more than one grade level. When this occurs, the item may appear in adjacent grade level banks. *If all items measure the same*

construct so that the item bank is created on the same dimension, items that have been presented at different grade levels maintain the same item characteristics and have the same IRT (Rasch) difficulty level. In this case, vertical scaling is not a requirement of CAT. Also, *if all items measure the same construct even if they are from different strands*, then scores from each strand subtest may constitute a measure of the student's achievement on the entire construct. Under these assumptions, assessments with modified achievement standards may be based only on any subset of all strands to be assessed for the grade level. The subset may be selected to reflect the on-grade content standards as modified for the student and the subtest score could legitimately be used under these conditions. The assumption of unidimensionality across strands and grade levels may need to be checked before a system of CAT can be implemented for assessments with modified achievement standards. Overall, it is important for CAT users to adhere to Standard 3.12 of the Standards (AERA et al., 1999, p. 45) that states that "The rationale and supporting evidence for computer adaptive tests should be documented." Among other things, "the documentation should include procedures used in selecting subsets of item for administration."

Conclusions and Recommendations

It is the responsibility of those who develop and implement large-scale assessments to compile evidence to support test users in determining the quality of the assessment and the information generated by the assessment (AERA et al., 1999). This chapter provides the reader with some suggestions for item quality indicators when students with disabilities are included in states' large-scale assessment systems. Table 7.3 summarizes the types of item-level analyses that may be conducted with the various assessment options in the model. Information in this table is based on a review of literature and state assessment websites conducted in June 2005. In many cases statistical indicators based on classical and modern test theories may be used, but

new questions must be posed and answered about the meaning of those indicators across student groups and assessment options. In other cases, methods for investigating the quality of items on some types of assessments have not yet been developed. It is possible that some states are collecting new types of item quality evidence but have not yet released their methods or results to the public.

Table 7.3. Item-Level Quality across the Range of Assessment Options (based on items or samples of items)

	"Typical" State Content Area Assessments	**Rating Scales and Observational Checklists**	**Portfolios**	**Performance Events**	**Performance Task Collections**
Statistical Indicators					
Incomplete items (timed tests)	✔	N/A	N/A	Usually N/A	N/A
Binary p-value	✔	✔	N/A	✔	?
Partial credit p-value	✔	✔	?	✔	?
KR20 or Cronbach alpha	✔	✔	?	✔	
Item-test correlations (discrimination indices)	✔	✔		✔	
DIF analysis	✔			?	?
Local Item Dependency	✔		N/A	?	N/A
Content Indicators					
Stakeholder review: item content	✔	✔	✔	✔	✔
Alignment studies - content	✔		?	?	
Alignment studies – cog. demand	✔		?	?	
Think aloud procedures	✔	✔	?		
Curriculum alignment studies	✔			?	?
Item editing	✔			✔	
Item bias/sensitivity review	✔			✔	
Indicators for modified and cross-grade items*					
IRT investigations-unidimensionality	✔				

Note. Established method for this type of assessment? Some application to this type of assessment. N/A Not applicable to this type of assessment. Empty cells reflect areas in which methods for evaluating items do not yet exist or have not been applied to this type of assessment for students with disabilities. Judgments about the use of indicators with various assessments are based on a review of published literature and state websites in June 2005.
*Also includes statistical indicators mentioned above.

Based on the methods available at this time, the following recommendations are made for establishing evidence of item quality.

1. All new items (including modified items) should be subject to classical item analysis that provides data on difficulty, time to complete the item, relevance of the item to the total test, differential item functioning, and local item dependency.

2. New information needs to be included in technical documentation about the process of narrowing and simplifying grade level achievement standards to create items for assessments based on modified achievement standards. When the process of modifying items begins with typical items from different grades, evidence is needed to ensure that the modified items still measure the intended construct. For alternate assessments based on alternate achievement standards, technical documentation should describe procedures that were used to make sure items reflect the highest possible expectations while still providing students who need the most extensive support and have minimal symbolic communication the opportunity to perform at all grade levels.

3. When considering validity evidence at the item level, long-standing procedures such as item bias and sensitivity reviews are still useful. In recent years more complex and rigorous ways have been developed to examine content-related validity evidence for assessments (Bhola et al., 2003). States should no longer rely simply on global judgments of content match by experts and other stakeholders.

4. It is recommended that cross-grade items and grade-level items be linked to a vertical scale and that the technical characteristics of cross-grade items be compared between students who take these items on grade-level and others who take them as cross-grade.

5. Special considerations should be made when investigating the quality of items used to assess students on modified achievement standards. Where an item bank exists that links several grades on a common core of curricula that reflect a common construct, a computer-adaptive testing system may be considered. Where various content strands (as measured by subtests) can be assumed to measure a common construct and where it is justifiable to use a few content strands to assess students under modified achievement standards, it is recommended that the scale score system for the total test be used for the selected strands.

6. States should evaluate the other types of assessments that are emerging for this population (e.g., performance and portfolio assessments) by the same standards of technical quality as traditional tests (AERA et al., 1999, p. 42). Developers and users of performance-based assessments need to strive for solid evidence of technical adequacy while also retaining the level of contextualization inherent in these assessments, which is important for the education of students with significant disabilities. Careful selection of a representative sample of students to participate in item try-outs, and careful descriptions of the sample's characteristics, will help stakeholders evaluate the appropriateness of the items for the intended population (AERA et al., 1999, Standard 3.8).

Along with the suggestion for methods of collecting evidence of item quality in the newer assessments comes a need to establish a sufficient base of empirical evidence to provide guidance about what constitute "good" items. For example, alignment benchmarks established under Webb's (1997) model were designed for general education tests in math and science. We are now seeing guidelines for applying alignment procedures to alternate assessments, based on Webb's model (Flowers, Browder, Ahlgrim-Delzell, & Spooner, in press; Hoffman & Arick, 2003; Tindal et al., 2005). However, empirical studies using Webb's method are still limited in

number, and more information is needed about expectations for those alignment criteria under varying conditions (e.g., alternate assessments based on modified or alternate achievement standards). Those studies need to address pitfalls in the alignment process, such as attempting to align when content standards are stated globally (Bhola et al., 2003; Resnick, Rothman, Slattery, & Vranek, 2003). As Downing and Haladyna (1997) noted,

> "Testing programs that examine small numbers of examinees may have very limited resources to devote to the rigorous item development validation process…Nonetheless, such programs have an obligation to make an effort to utilize high-quality test items…" (p. 63).

It is important to develop the new technologies needed to engage in rigorous empirical studies of the full range of assessment options for *all* students with disabilities.

References

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

Baranowski, R. A. (in press). Item editing and item review. In S. M. Downing and T. M. Haladyna (Eds.), *Handbook of Test Development.* Mahwah, NJ: Lawrence Erlbaum Associates.

Bechard, S. (2005). *Developing alternate assessments using expanded benchmarks from a nine-state consensus framework in reading, writing, mathematics, and science.* Paper presented at the 2005 National Council on Measurement in Education (NCME) annual conference, Montreal, Quebec, Canada. Retrieved May 1, 2005, from http://www.measuredprogress.org/Resources/Publications/DevAltAssessments.pdf

Bhola, D. S., Impara, J. C., & Buckendahl, C. W. (2003). Aligning tests with states' content standards: Methods and issues. *Educational Measurement: Issues and Practice, 22*(3), 21-29.

Bond, T. G., & Fox, C. M. (2001). *Applying the Rasch model*. Mahwah, N. J.: Lawrence Erlbaum Associates.

Brennan, R. L. (1992). *Elements of generalizability theory*. Iowa City, IA: ACT Publications.

Browder, D., Flowers, C., Ahlgrim-Delzell, L., Karvonen, M., Spooner, F., & Algozzine, R. (2004). The alignment of alternate assessment content with academic and functional curricula. *The Journal of Special Education, 37,* 211-223.

Clauser, B. E., & Mazor, K. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and Practice*, *17*, 31-44.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 1,* 98-104.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Forth Worth, TX: Harcourt Brace Jovanovich College Publishers.

Downing, S. M. (in press). Twelve steps for effective test development. In S. M. Downing and T. M. Haladyna (Eds.), *Handbook of Test Development.* Mahwah, NJ: Lawrence Erlbaum Associates.

Downing, S. M., & Haladyna, T. M. (1997). Test item development: Validity evidence from quality assurance procedures. *Applied Measurement in Education, 10,* 61-82.

Engelhard, G. Jr., Davis, M., & Hansche, L. (2000). Evaluating the accuracy of judgments obtained from item review committees. *Applied Measurement in Education, 12,* 199-210.

Fahey, K., Filbin, J., & Connolly, T. (2004). *Colorado's performance-based alternate assessment: Historical perspective and lessons learned technical report*. Denver, CO: Colorado Department of Education. Retrieved June 23, 2005, from http://education.umn.edu/NCEO/StateForum/default.htm#manuals

Ferrara, S., Huynh, H. & Baghi, H. (1997). Contextual characteristics of locally dependent open-ended item clusters in a large-scale performance assessment. *Applied Measurement in Education*, *10*, 123-144.

Ferrara, S., Huynh, H., & Michael, H. (1999). Contextual explanations of local dependence in item clusters in a large-scale hands-on science assessment. *Journal of Educational Measurement, 36,* 119-140.

Flowers, C., Browder, D., & Ahlgrim-Delzell, L. (in press). The alignment of three states' alternate assessments to state standards. *Exceptional Children.*

Flowers, C., Browder, D., Ahlgrim-Delzell, L, & Spooner, F. (in press). Promoting the alignment of curriculum, assessment, and instruction. In D. M. Browder & F. Spooner (Eds.), *Teaching reading, math, and science to students with significant cognitive disabilities*. Baltimore: Paul H. Brookes.

Haladyna, T. M. (1999). *Developing and validating multiple-choice test items* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Herman, J. L., Webb, N. M., & Zuniga, S. A. (2005, May). *Measurement issues in the alignment of standards and assessments: A case study*. (CSE Report 653). Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing. Retrieved June 16, 2005, from http://cresst96.cse.ucla.edu/reports/R653.pdf

Hieronymous, A. N., & Hoover, H. D. (1986). *Iowa Tests of Basic Skills manual for school administrators*. Chicago: Riverside.

Hoffman, T., & Arick, J. (2003). *Oregon statewide assessment system extended career and life role assessment system alignment with Oregon standards*. Salem, OR: Oregon Department of Education. Retrieved June 23, 2005, from http://www.ode.state.or.us/teachlearn/testing/admin/alt/ea/pubs/casiasxclras1021033504.pdf

Huynh, H., Meyer, P., & Barton, K. (2000). *Technical documentation for the South Carolina PACT-1999 tests*. Columbia, SC: South Carolina Department of Education. Retrieved July 16, 2005, from http://www.myscschools.com/offices/assessment//Publications/1999_Pact_document.doc

Idaho State Department of Education (2004). *The 2004 administrator's guide to Idaho alternate assessments in reading, language arts, and mathematics*. Boise, ID: Author. Retrieved June 24, 2005, from http://www.sde.state.id.us/SpecialEd/AltAssessment/iaamanual.pdf

Johnson, E., & Arnold, N. (2004). Validating an alternate assessment. *Remedial and Special Education, 25,* 266-275.

Karvonen, M., & Huynh, H. (2005). *Relationship between IEP characteristics and test scores on an alternate assessment for students with significant cognitive disabilities*. Paper submitted for publication.

McGrew, K. S., & Evans, J. (2003). *Expectations for students with cognitive disabilities: Is the cup half empty or half full? Can the cup flow over?* (Synthesis Report 55). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved July 16, 2005, from http://education.umn.edu/NCEO/OnlinePubs/Synthesis55.html

Meyer, P., Huynh, H., & Seaman, M. (2004). Exact small-sample differential item functioning methods for polytomous items with illustration based on an attitude survey. *Journal of Educational Measurement*, *41*, 331-344.

Miller, M. D., & Linn, R. L. (2000). Validation of performance-based assessments. *Applied Psychological Measurement, 24,* 367-378.

Northwest Evaluation Association (n.d.). *Assessments should make a difference*. Lake Oswego, OR: Author. Retrieved July 2, 2005, from http://www.nwea.org/assessments/

Patz, R. J., & Hanson, B. A. (2002, April). Psychometric issues in the construction and maintenance of vertical scales. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, Louisiana.

Peterson, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling, norming, and equating. In R. L. Linn (Ed.) *Educational Measurement* (3rd ed., pp. 221-262). New York: American Council on Education.

Porter, A. C. (2002). Measuring the content of instruction: Uses in research and practice. *Educational Researcher, 31*(7), 3-14.

Resnick, L. B., Rothman, R., Slattery, J. B., & Vranek, J. L. (2003). Benchmarking and alignment of standards and testing. *Educational Assessment, 9*, 1-27.

Roach, A. T., Elliott, S. N., & Webb, N. L. (2005). Alignment of an alternate assessment with state academic standards: Evidence for the content validity of the Wisconsin alternate assessment. *The Journal of Special Education, 38*, 218-231.

Shavelson, R. J. & Webb, N. M. (1991). *A primer on generalizability theory*. Newbury Park, CA: Sage Publications.

Sireci, S. G. (1998). Gathering and analyzing content validity data. *Educational Assessment, 5,* 299-321.

South Carolina Department of Education (2003). *Technical documentation for the 2003 Palmetto Achievement Challenge Tests of English language arts, mathematics, science, and social studies*. Columbia, SC: Author. Retrieved June 24, 2005, from http://www.myscschools.com/offices/assessment/Publications/Index_of_Technical_Reports.htm

South Carolina Department of Education (2005). *HSAP-Alt Spring 2005 test administration manual*. Columbia, SC: Author. Retrieved June 23, 2005, from http://www.myscschools.com/offices/assessment/Publications/HSAPAltTAM030805.pdf

Thompson, S. J., Johnstone, C. J., & Thurlow, M. L. (2002). *Universal design applied to large-scale assessments* (Synthesis Report 44). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved July 16, 2005, from http://education.umn.edu/NCEO/OnlinePubs/Synthesis44.html

Thompson, S., & Thurlow, M. (2003). *2003 state special education outcomes: Marching on*. Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved June 16, 2005, from http://education.umn.edu/NCEO/OnlinePubs/2003StateReport.htm/

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 June 16, 2005, from http://education.umn.edu/NCEO/OnlinePubs/Synthesis41.html

Tindal, G., Cipoletti, B., & Almond, P. (2005). *Alternate assessments: Evidence based on content and alignment with standards*. Manuscript in preparation.

Webb, N. L. (1997). *Criteria for alignment of expectations and assessments in mathematics and science education*. (NISE Research Monograph No. 6). Madison, WI: University of Wisconsin-Madison, National Institute for Science Education. (ERIC Document Reproduction Service No. ED414305)

Webb, N.L. (1999). Alignment of science and mathematics standards and assessments in four states. (NISE Research Monograph No. 18). Madison: University of Wisconsin-Madison, National Institute for Science Education. (ERIC Document Reproduction Service No. ED440852)

Weiss, D. J. (Ed.). (1983). *New horizons in testing: Latent trait test theory and computerized adaptive testing*. New York: Academic Press.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30*, 187-213.

Zieky, M. (in press). Fairness review in assessment. In S. M. Downing and T. M. Haladyna (Eds.), *Handbook of Test Development.* Mahwah, NJ: Lawrence Erlbaum Associate

CHAPTER 8: RELIABILITY ISSUES AND EVIDENCE

Introduction

In this chapter we discuss assessment *reliability*, with emphasis on the benefits of careful assessment design and administration when used for measuring students with disabilities (SWD). Our discussion is structured around the concept of measurement error, specifically in the context of, (a) the process for collecting responses, (b) the scores assigned to observed responses, (c) the decisions made from these scores, and (d) the reliability of an assessment system. The importance of reliability pivots around the need for assurances that assessments are designed and used in ways that minimize unstable response patterns, and corresponding examinee scores (individually and collectively). Reliable measurement cannot be overstated as a necessary, albeit not sufficient, condition for measurement validity (Chapter 8). Essentially, the challenge we address involves offering flexible assessment protocols without compromising measurement accuracy.

Perhaps the most psychometrically technical aspect of assessment, reliability is generally described in terms of score 'stability'. The *Standards for Educational and Psychological Testing* (1999) defines reliability as "the consistency of [such] measurements when the testing procedure is repeated on a population of individuals or groups" (p. 25). Reliability typically refers to the measurement error that is introduced into the "entire measurement process" (p. 27) and limits the degree to which generalizations can be made beyond the specific testing event and quantifies the confidence that can be held in the value assigned to any performance. "Reliability data ultimately bear on the repeatability of the behavior elicited by the test and the consistency of the resultant scores" (p. 31). Specifically, for the purposes of this paper, we are concerned about the reliability (dependability, replicability, etc.) of behavior, scores, and inferences.

Reliability requires quantifying the measurement error associated with (a) observed behaviors, and (b) numeric scores assigned to our observations. The situation becomes complex when observed behaviors depend on the sampling of items (Chapter 7) and how items 'stimulate' observed behavior. This is true irrespective of whether an item response format is selected response (SR) or constructed response (CR). Furthermore, assigning numeric values to observed behaviors, i.e. scoring and scaling, impacts the reliability of our measurement system. Scoring issues pertain whether the score is very specific or very general, e.g., with three score levels as in conventional classification standards of 'does not meet', 'meets', or 'exceeds'. Ultimately, we need some indication that our error is diminished with careful assessment design (item sampling, administration, and scoring).

There are basically two kinds of error: systematic and unsystematic (random). Obviously, the use of performance assessments and testing accommodations will introduce a host of additional sources of error beyond the student and item. Scoring via raters is one such source. The performance tasks, while considered comparable, may render alternate forms nonequivalent. The wide variety of approaches that comprise a state's assessment of diverse populations require multiple types of evidence that reliable measures are being obtained. As discussed throughout this paper, a range of assessments are designed so that students can freely participate in various options of the general education test (with or without accommodations). Furthermore, this variety of assessments is theoretically aligned with grade level content standards. Obviously, this scenario suggests the potential for an array of 'nuisance' factors that may threaten the psychometric reliability of our assessments.

A reasonable organization of our discussion proceeds from first providing a formal definition of reliability, and then identifying sources of error, and finally describing evidence that informs us about measurement reliability and measurement designs that are apt to attenuate the measurement error. In preface to this presentation, most of the relevant, technical psychometric issues pertaining to measurement reliability cannot be thoroughly covered in this paper without loss of practicality. Readers wishing for more technical documentation are referred to the references providing throughout.

Formal Definition of Reliability

It will be difficult to technically appraise the presence of reliability without having some formal definitions. Most important are the concepts of error scores, parallel forms, reliability coefficients, and standard error of measurement.

Error Scores

One of the most traditional conceptualizations is in terms of the *true score*, Unfortunately, we never actually know a person's true score; it must be estimated from the *observed score*, which provides imperfect information. Therefore, in addition to the observed score, we must theorize an *error score*. A very simple concept of observed score, true score, and error score is captured in equation 1.

$$\textit{observed score} = \textit{true score} + \textit{error score}. \quad (1)$$

Very simply, the observed score is composed of two components, (1) the true score and (2) the error score. Obviously, both the true score and error score are unobserved and must be estimated.

The concept of error scores is at the heart of reliability. The goal of good measurement design is to minimize the error component. Note, in our simple model (equation 1), error is thought to occur randomly. The importance of random error may be recognized if we use an

assessment repeatedly to measure the same individual. The obtained measures would not be identical. In fact, they will be variable, depending on the reliability of our assessment instrument. Our best estimate of the examinee true score will be the average of the repeated measures. The variability around the mean is our theoretical concept of error, also called error variance.

Parallel Forms

A formal concept of error is developed largely around assumptions pertaining to *parallel forms*. Clearly, to estimate our error scores we do not want to administer the same assessment repeatedly to the same examinee. We would want to use parallel forms of the assessment. Parallel forms are assessments comprised of different tasks, but the tasks are assumed to be randomly sampled from the same domain, and they are of comparable difficulty. The correlation of any two parallel forms and will be highly correlated **only if the assessment is reliable**. The concept of correlated parallel forms lets us continue the definition of psychometric reliability. Equation 2 describes in terms of observed score variances and, and their covariance. Equation 2 can be written in terms of true score and observed score variance (Feldt and Brennan, 1989; Chatterji, 2003). Equation 3 shows that the observed correlation of two parallel forms provides information for estimating assessment reliability. Based on equation 1, equation 4 shows that observed score variance is composed of true score and error score variance. As error score diminishes, the ratio of true score and observed score variance will approach the value 1. So, if the correlation of parallel forms, approaches 1, then the error variance must be small.

While the assumption of parallel forms (items or tests) is generally necessary psychometrically, it is extremely difficult to accomplish. Below we identify sources of error, nonequivalent (nonparallel) forms are identified as one of the most important and difficult to control.

Standard Error of Measurement and Information

Another conceptualization of measurement accuracy is developed in terms of the standard error of measurement (SEM). As described above, the concept of random error around the true score results from administering repeated parallel forms, the SEM of a measure is essentially the average deviation of error scores around the true score. As with reliability, SEM can be estimated in terms of correlated observations.

According to equation 5, as the correlation of parallel forms increases, the standard error of measurement diminishes to zero. It is very important to keep in mind that our measures are estimations, and that theoretically, each time the assessment is administered we will obtain a different measure. How different depends on the reliability or error in measurement?

The preceding SEM estimation is classical, in contrast to an alternative Item Response Theory (IRT) perspective. In IRT, items are calibrated with respect to difficult and sensitivity among many other possible item characteristics. Using the item calibrations, it is possible to estimate the amount of *information* provided by a test and its items. Furthermore, the amount of information depends on ability. For instance, a difficult item will generally be very uninformative when administered to someone with low ability. Response to an easy item provides much more information about someone with low ability. As a rule, items are most informative when responded to by a person with comparable ability level. (Note, in IRT, person ability and item difficulty are on the same scale, which makes it possible to match items to persons.) A highly informative test is one that is not too easy nor too difficult for the respondent.

Sources of Random Error

The alternate assessment poses numerous challenges associated with measurement error. Some sources of random error pertain to examinee characteristics, item and test design, administration, and scoring protocols. State large-scale assessments typically use both selected response (SR) items and constructed response (CR) items or tasks, either with or without accommodations. CR is used in performance measures that require a rubric (subjectively scored) or performance measures that require observation of student performance, or collection of student work portfolios. Likewise, several approaches within the participation options employ some form of constructed response such as observations (or reflections) of student performance, (structured or unstructured) performance tasks, or collections of student work. The opportunities for measurement error seem to expand with increased flexibility. Therefore, assessment design and reliability estimates need to consider the multiple factors that can attenuate measurement accuracy. The challenge with isolating and controlling sources of measurement error is complicated by the relationships among error sources, as described below.

Student Variables

All students come into school situations from a variety of home environments, all of which have bearing on their performance in school. For example, students come to school hungry, tired, or fatigued, etc. As they interact with classroom tasks and receive feedback, they come to have expectations of success or failure, reflecting motivation and self-efficacy that may interact differentially with the kinds of tasks they are given. All these conative factors may influence the results from large-scale assessments in very unsystematic ways (McGrew, Johnson, Cosio, & Evans 2003).

In addition, for students with disabilities, several personal and behavioral characteristics also may unsystematically influence their performance. For example, with some disabilities, medications are used (attention deficit-hyperactivity disorders) and depending upon the dosage or uptake, performance on large-scale tests may be inconsistent. Even without the use of medications, students with disabilities may exhibit behavioral tendencies that distract them from attending to tasks (tendencies to perseverate, be distractible, be inattentive, etc.). Whenever such behavior influences their performance unsystematically, the validity claim is weakened (that the outcomes reflect what the student knows and can do and therefore, the inference of proficiency is less certain). The format of the test may be very unlike anything done in the classroom and influence students in uneven ways (negatively affect some and be neutral for others).

Therefore, participation in large-scale assessment systems is not just a matter of students being scheduled to take a test at the end of the year. Rather, the assessment needs to be considered an important part of the school's annual cycle of activities. And the large-scale assessment program needs to consider such behavioral factors when collecting performances from students. Although tests may be given only once during the year (typically in the spring), plans for its administration should be introduced early in the year to allow students and teachers a fair opportunity to participate.

For example, many states require teachers to use accommodations in testing that have been part of the accommodations used in the classroom. If these accommodations are not a regular part of either the teaching or testing, unsystematic variance may be introduced. Furthermore, a watchful teacher during the year may be able to better understand critical student behaviors to recommend specific accommodations for the test at the end of the year.

Task Sampling

Samples of performance tasks are prepared in the spirit of parallel forms. That is, the tasks are ideally comparable to the extent that a student would not prefer one over another as they are both of equal difficulty. As with any sampling of items, the sample of tasks is apt to be variable with respect difficulty and representation of the performance domain. Using multiple tasks are alternate forms for either comparing an individual over time or one examinee to another, the extent to which the tasks differ is of obvious consequent. Score variability attributable to task differences can be identified with carefully controlled studies.

With portfolio assessments, it often is the teacher who selects the pieces of student work that to be included in a student's collection or portfolio. Although selection criteria may be specified both in the test administration manual and in test administration training, ultimately teacher judgment is involved. Consequently, the portfolio or collection of work may represent the grade level content broadly or narrowly. What is included or not included in the collection can, therefore, affect the adequacy of the evidence in representing what the student knows and can do. From the perspective of replicability, another collection, assembled at a different time or by a different teacher, may or may not support inferences drawn from the original collection.

Item Calibration

Increasingly, assessment developers are recognizing the value of item calibration in response to the reality that assessment items are not necessarily equivalent (Thissen & Wainer, 2001). Whether CR or SR, assessment items provide differential amounts of 'information' depending on the respondent's true ability. Item calibrations are estimates of item characteristics such as item difficulty, item sensitivity among many other possibilities (van der Linden &

Hambleton, 1997). *With accurate item calibrations*, estimation of true scores becomes considerably more accurate, i.e., minimized standard error of estimation.

Model choice and calibration accuracy pertain directly to measurement reliability. As emphasized above, the value of item calibrations for ability estimation depends on the appropriate choice of IRT model and proper calibration procedures. The technicalities involved in these decisions far outreaches the scope of this paper, but the importance of good calibration should be noted. First, the calibration process requires adequate sampling of examinee response patterns. Ideally, a range of abilities should be represented in the calibration sample. Second, an appropriate IRT model should be applied. For instance, alternate assessments rely heavily on performance tasks. Usually, observed performance is scored polytomously, i.e. more than correct/incorrect. This requires using a rating scale, partial credit, or graded response model. Numerous other possible models are available in the literature (van der Linden & Hambleton, 1997; Boomsma, van Duijn & Snijders, 2001).

Another aspect of IRT item calibration must be mentioned in the context of reliability. Assessments are ideally unidimensional, i.e., measuring a single construct. This is very difficult to accomplish on rigidly constructed measures. With alternate assessments, the challenge increases dramatically as we apply flexible CR tasks and risk the involvement of multiple ability factors, not to mention factors such as time constraints, and rater severity, etc. Multidimensional IRT models can be used to accurately calibrate performance tasks, thereby yielding more reliable ability estimation. Local item dependency (Chapter 7) is generally attributable to multidimensional problems. Good assessment development will identify and provide corrections for this situation.

Finally, when developing measures for diverse populations, it is very important to understand if assessment tasks function identically across the populations. Yovanoff and Tindal (in press) were able to understand how well the Oregon Early Reading Extended Assessment tasks function irrespective of whether students were special education or general education. Ideally, tasks should perform equivalently across populations, though one population may have higher mean ability than another population. This type of analyses helps maintain a quality control over assessment development.

Scaling and Equating

In addition to IRT item calibrations, assessments are often rescaled or equated with an external measure. This occurs in state measurement programs when scores on alternate assessments are most desirably on the same scale as the general assessment scores. Yovanoff and Tindal (in press) scaled the Oregon Early Reading Extended Assessment with the general Oregon Statewide Assessment. Using the appropriate IRT model and necessary research design, the investigators demonstrated the early reading performance tasks do function appropriately for students unable to be accurately measured with the general benchmark one statewide assessment. In effect, this type of scaling and measurement calibration makes the assessment both more accurate and more informative. Equating standard errors are extremely important for appraising the accuracy of score equivalents (Kolen & Brennan, 1995). Whenever assessments are equating, the standard error should be reported.

Scoring Process

Irrespective of any errors made in collecting assessment data or as estimated with reliability coefficients, different or unique errors can be made when making judgments. In this category of unsystematic error, we are referring to ratings and classifications being made for students such as pass/fail or below basic, basic, proficient, and advanced. In this instance, the focus is less on the actual score consistency than on the consistency of judgments into states of mastery ("proficient"). In judgment error at the score level, the focus is on rubrics (or partial correct responses); at the classification level, the judgment lies within not only the final decision to classify a student student's performance but also in the standard setting process itself. The analysis, therefore, needs to consider both the individual judgments made for a student as well as the overall process for making classification decisions.

"Where the purpose of measurement is classification, some measurement errors are more serious than others. An individual who is far above or far below the value established for pass/fail or for eligibility for a special program can be mis-measured without serious consequences. Mismeasurement of examinees whose true scores are close to the cut score is a more serious concern. The techniques used to quantify reliability should recognize these circumstances. This can be done by reporting the conditional standard error in the vicinity of the critical score" (1999 Standards, p. 3). For example, Hollenbeck and Tindal (1999) reported that although judges were in considerable agreement at the exact or adjacent score values, they were in the greatest disagreement with respect to judgments of proficiency (at the cut score); as a consequence, the state educational agency began reporting 'conditional' proficiency (in essence noting that disagreement occurred at the cut-score) to acknowledge this type of error.

We address classification decisions in detail in chapter 11 when we present information on score reporting, including standard setting. Suffice it to say here that reliability at the classification level involves attention to proper selection of content experts as well as training with feedback. And the 1999 Standards are very clear: "When subjective judgment enters test scoring, evidence should be provided on both inter-rater consistency in scoring and within examinee consistency over repeated measurements. A clear distinction should be made among reliability data based on (a) independent panels of raters scoring the same performances, (b) a single panel scoring successive performances or new products, and (c) independent panels scoring successive performances or new products" (p. 34).

A hybrid of reliability of the scoring process and the decisions made from the score come about in the case in which one judgment is made based on all evidence in the portfolio, which is somewhat unusual, as most portfolios are comprised of several dimensions needing judgment and the score eventually is translated into a proficiency decision. Oregon offers both alternate assessments judged against grade level achievement standards, the Juried Assessment, and alternate achievement standards, the Extended Assessments, the former of which reflects this confluence of reliability in scoring and making judgments.

The Juried Assessment administration manual provides the following description of a Juried Assessment: "Juried assessment is completed under non-standard conditions – either through the Collection of Evidence process or through a Collection to Jury a Modification. . .. Each option answers the question: "Does the evidence provided by the student meet the Oregon content and performance standards for a particular subject?" (Oregon Department of Education, 2005, p. 5). Training in scoring is provided to assure reliability; in addition, collections that are rated above the standard must be verified through a secondary source. The requirements for a

collection of evidence in Oregon's Juried Assessment are clarified in subject specific documents on the website along with the administration manual. A description of the juried process was extracted from the manual for inclusion here (see Figure 1). The Juried Assessment is designed for students who are literate in a language other than English, with physical disabilities preventing participation in the writing assessment, or any disability affecting the student's ability to read and write.

The Moderation Panel would consider the evidence and determine whether test results using the translation in the first example, the word prediction software in the second, or the auditory methods in the third, are reliable and valid in addressing a specific standard. If the panel determines that the change does not affect the validity of the test score for this student, the student's score would then be considered for meeting that standard.

Juried Modifications are approved one student and one assessment at a time. A panel of experts makes the final determination. It is believed that there may be a student with a significant learning disability who uses assistive technology, screen readers, and recorded text to perform the task of understanding text and interpreting "meaning". The panel might approve this modification as an accommodation for the student after reviewing the student's case if:

- The student is skilled in using the read aloud adaptations
- The measure of comprehension reflects the student's own knowledge and understanding
- The student achieves the same standards for interpreting text required of all students

If approved, the student would be permitted to use the "read aloud" modification with the Reading/Literature Knowledge and Skills assessment and can "meet" (e.g. be determined "proficient" on the standard.) There is the possibility that the decision could be made after testing was completed if there was sufficient documentation of the process to assure that it was the student's own work.

Figure 1. Example Moderation Process in a Juried Assessment

If Oregon's approach with the juried assessment is typical of what states develop to meet the "alternate assessment judged against grade level content standards," there is likely to be heavy reliance on two categories of judgment, one addressing the *sufficiency* of evidence to make a determination and another of *proficiency* based on the collection as a whole. To achieve reliability in these judgments, Oregon relies on systematic procedures and structured criteria for making the determination. Oregon's option was considered by Oregon's national Technical Advisory Committee (TAC) and considered a reasoned and prudent approach to avoiding false negative judgments (e.g., denying proficiency when it was deserved). Some members conjectured that the juried assessment might present a higher standard of achievement than the general multiple-choice assessments. Their view was noteworthy in that they considered the approach on its merits and weighed against both the *Standards* for testing and the consequences to the student. In the end, the TAC considered it a promising strategy to which they had nothing better to offer that would meet the same demands for integrity and fairness.

Alignment

Another source of unsystematic error in the data collection process is introduced in the participation of students in alternate assessments when conducting an alignment analysis between grade level content standards and portfolio assessment approaches. When states use portfolios as part of the alternate assessment that are defined by student need (e.g., a fixed set of entries are not specified a priori but teachers select unique entries for each student), the process of alignment between the alternate assessment and the standards cannot take place without sampling students. If the sampling of standards is isomorphic with the sampling of students, any statements about alignment on content coverage, breadth of knowledge, and depth of knowledge are primarily a function of those who participated. In this source of error (more like survey

sampling error), stable statements about alignment are difficult to make. Because each student samples only a pre-specified set of standards by design, alignment at the student level is inherently skewed. As a result, the process needs to sample enough students to determine the coverage of standards being addressed at the systems level. At this level, sampling of students needs to be considered using not only ages but disabilities, geographic region, and type of program to make any inferences about alignment.

Assessment Administration

A final source of nuisance error relates to assessment administration. One of the reasons for using standardized procedures in large-scale assessment systems is to minimize the error from external sources. Testing personnel (most often teachers), however, can introduce error (unsystematic variance) in the way they administer or score the test. Ironically, few states have training systems for test administration. We assume that the conditions as noted in the test booklets are the same as those enacted in the classroom. Significant deficits are evident in teachers' knowledge concerning high-stakes testing. Most of teachers' knowledge about testing and measurement comes from "trial-and-error learning in the classroom" (Wise, Lukin, & Roos, 1991, p. 39). This problem, however, is rarely addressed through any in-service programs, even though these authors attributed this lack of assessment knowledge to teacher certification agencies at the state-level (not requiring assessment/measurement courses for initial teacher certification).

This kind of unsystematic error is best addressed by state educational agencies (SEAs) by rigorous training and monitoring throughout administration of the large-scale assessment system. "Measurements derived from observations of behavior or evaluations of products are especially sensitive to a variety of error factors. These include evaluator biases and idiosyncrasies, scoring

subjectivity, and intra-examinee factors that cause variation from one performance or product to another (Standards, 1999, p. 29). As we noted in previous chapters, evidence can be both procedural and empirical for documenting the reliability associated with (a) test administration, and (b) response rating.

Participation options present a somewhat different challenge regarding the sources of error. The estimate of reliability for the first option, taking the general assessment *without* accommodations, is likely to be a characteristic that has already been addressed in most states' technical reports. There is still some question about which students participated in the assessment or more importantly which subgroups did not participate and the possible effect participation might have had on estimates of reliability.

Participation in the general assessment with accommodations has been increasingly studied (Sireci, 2003) and the use of accommodation has been increasingly utilized (Clapper, Morse, Lazarus, Thompson, & Thurlow, 2005), though in both instances, we know little about the effect of reliability on participation or the use of accommodations (see the later section on standard error of measures). Nevertheless, the 1999 standards are quite clear: "When significant variations are permitted in test administration procedures, separate reliability analyses should be provided for scores produced under each major variation if adequate sample sizes are available" (p. 36).

Therefore, we know little about estimates of reliability in any testing program in which changes are made (as accommodations or for the remaining three options: alternate assessments judged against grade level, modified, or alternate achievement standards). These latter options present challenges because the quantity of data resulting from the option is likely to be limited in scope and therefore making inferences beyond an instance of testing difficult. These are alternate

assessment judged against grade level and alternate assessment judged against alternate achievement standards. It is possible that the alternate assessment based on modified achievement standards can follow the course of research followed by accommodations with enough and adequate standardization to allow generalizations.

Ironically, standardization may be the antithesis of the solutions for controlling measurement error. By forcing tests to be taken in the same way across all students, both internal and external sources of error may be exacerbated [maintained] rather than controlled. For example, a student with hyperactive tendencies (on medications to control sources of error internal to the student due to attentiveness) and given a test in one session (to control sources of error external to the student due to time and setting), may actually need to have accommodations made in order to make appropriate inferences about their performance that is not influenced by construct irrelevant variance. When accommodations are made, then, reliability related evidence is needed to support the consistency of administration and scoring across replications (time, items, and raters).

Generalizability Theory and Differentiated Error

So far, all nuisance error has been undifferentiated and treated as one source. However, our discussion has noted many plausible sources of this undesirable error. Using generalizability theory (G theory), this single 'error' term is decomposed into various *facets*, i.e. factors, that influence performance (Brennan, 2001). The most usually studied facets are judges, tasks, and occasions. Using carefully planned research designs, assessment developers can better understand to what extent the assessment facets influence the reliability of observed behavior.

Generalizability studies (G-study) are used to differentiate the error and identify how much of the examinee score is attributable to lack of rater agreement, or task variability, for

instance. This is extremely valuable information as it casts light on where our assessments need adjustment. Obviously, alternate assessment can benefit from this effort identify exact sources of error. Once the error term is partitioned into specific sources, the assessment development research can proceed to estimate measurement reliability. Using Decision Studies (D-study), multiple assessment scenarios can be constructed along with the corresponding reliability estimate. For instance, with G-study information, it is possible to know how much reliability will improve if more raters were used, or responses to more tasks were obtained. The implications of this for time and money considerations are very clear. If examinee time is not possible, then the assessment would not include more tasks. Instead, the addition of another rater could be considered to the extent that it would bring reliability up to an acceptable standard.

Types of Reliability

We have identified some of the primary sources of random error that can jeopardize the quality of alternate assessments. Corresponding to many of these error sources there is an appropriate type of reliability.

Conventional reliability indices such as Cronbach's alpha and Kuder-Richardson formulas, KR20 and KR21, are based generally on the concepts of observed score variance, true score, and error score variance (Feldt & Brennan, 1989). According to our formal definitions above (equation 3 and equation 4), reliability, is considered as the ratio of true score variance to observed score variance (true score plus error variance). Ideally, we want to diminish error and maintain an observe score that is largely composed of true score. Generally, as error score variance diminishes, the correlation of observed and true scores approaches the maximum value '1'. The correlation of parallel forms is conceptually identical to the correlation of true score and observed score, and it is one reliability index of popular interest. Of course, this depends on the

equivalence of the two assessments as required by parallel forms. Conventional reliability indices and estimates of standard error allow us to understand the stability (consistency) of the score within the distribution and further calculate confidence intervals around the true score.

In the typical large-scale assessment, four types of reliability coefficients are considered associated with different sources of error: test-retest, parallel forms, internal consistency, and inter-rater agreement. We use test-retest if we believe error is due to occasion or time; we use parallel form if we think error is due to the form we used; we use internal consistency if we believe we have introduced error from our specific sample of items, tasks, or behaviors; finally, we use inter-rater if we have any reason to question the judgment or rating of performance.

In alternate assessments, the approach taken determines the type of reliability that is most important. And as noted in the 1999 standards "a reliability coefficient or standard error of measurement based on one approach should not be interpreted as interchangeable with another derived by a different technique unless their implicit definitions of measurement error are equivalent" (p. 32).

Teacher indirect observations (judgments) when using rating scales need to address inter-judge primarily to ensure the same ratings would be given by someone (anyone) else. It also may be important for the ratings to be made at two different times to ensure that the behavior is stable. When ratings are completed with sub-scores reported, internal consistency may be a possible necessary dimension to ensure that the individual judgments hang together within a sub score. Portfolios include a compilation of work samples usually collected in the 'natural' environment (of the school, community, work, or family) and are judged on some dimension (such as generalization, independence, accuracy, etc). Because a judgment is being made, inter-judge agreements are critical. Because, however, the collection of work samples often includes

slightly different forms, it may be important to consider the variance that accrues from these 'parallel' forms. Finally, it is possible that internal consistency is needed to support the claim that these various work samples are all related equally to the total score. For performance events, the most critical dimensions of dependability are the internal consistency (for brief tasks in which many 'items' comprise a task) or parallel form (a particularly important facet in generalizability theory); test-retest possibly may be important with extended tasks. Finally, performance task collections need to emphasize the same types of reliability as used with portfolios.

Table 1 and Table 2 summarize much the foregoing discussion. Table 1 provides a simple summary of appropriate reliability indices for various assessment designs. Table 2

Table 1. Assessment Methods and Appropriateness of Reliability Indices.

	Reliability Indices				
Assessment Method	**Test-Retest**	**Parallel Form**	**Internal Consistency**	**Inter-Judge**	**IRT Calibration/ Standard Error**
Selected Response	Very Appropriate	Very Appropriate	Very Appropriate	Very Appropriate	Very Appropriate
Checklists and Rating Scales	Appropriate	Inappropriate	Somewhat Appropriate	Very Appropriate	Appropriate
Portfolio	Inappropriate	Appropriate	Somewhat Appropriate	Very Appropriate	Somewhat Appropriate
Performance Event	Somewhat Appropriate	Appropriate	Very Appropriate	Not Applicable	Appropriate
Performance Tasks	Appropriate	Inappropriate	Somewhat Appropriate	Very Appropriate	Appropriate

For most assessment approaches listed in Table 1, the standard error of measurement (SEM) or conditional standard error of measurement (CSEM) should be provided. The value of SEM estimates is that they enable the computation of confidence bands around our examinee ability and item characteristic estimations. Table 2 lists some of the standards associated with reporting SEM and CSEM (AERA et al., 1999).

Table 2. Standards Relevant to the Standard Error of Measurement

Standard 2.1 "For each total score, sub score, or combination of scores that is to be interpreted, estimates of relevant reliabilities and standard errors of measurement or test information functions should be reported" (p. 31). Standard 2.2 "The standard error of measurement, both overall and conditional (if relevant), should be reported both in raw score or original scale unite and in units of each derived score recommended for use in test interpretation" (p. 31). Standard 2.14 "Conditional standard errors of measurement should be reported at several score levels if constancy cannot be assumed. Where cut scores are specified for selection or classification, the standard errors of measurement should be reported in the vicinity of each cut score" (p 35).

According to Nitko (1996), these different methods for obtaining reliability coefficients generate different results, the composition and variability of the group affects the reliability coefficient because the method is based on correlations, better reliability results from more items, behaviors, or products, objective scoring generally results in more reliable results, and the amount of error associated with an examinees performance depends (is conditional) on their

score level. Fleming, Taylor, and Carran (2004) compare two methods for determining inter-rater reliability, highlighting the effects when correlation coefficients are used versus percent agreement and conclude that judges can have judgments that are highly correlated and in disagreement.

As Haertel (1999) also points out that several sources of error are not considered at the group statistics level and may become important when making statements about aggregate level outcomes. Absences influence performance at the score level and likely provide very unsystematic variance at the distribution level. Invalid or below chance scores influence the mean of a group with students having varying levels of 'attemptedness' in completing an item. Invalid identification of students at the aggregate level influences the scores that are included in the distribution. Finally, several other mistakes in the calculation of group statistics have an obvious influence on the reliability of results.

While these forms of reliability are traditionally reported in the general education assessment, they have rarely been considered for alternate assessments. Rather, most reported evaluations of reliability for alternate assessments have focused primarily on inter-rater agreement (Browder, et al., 2003). Crawford and McDonald (2003) analyzed inter-rater agreement by counting the numeric difference between teachers' scores and the scores of a trained external rater. Each indicator was scored for each student assessment in the study, for level of agreement using score as an exact match between the teacher and trained rater and as a difference of one on a 5-point scale. They also examined scores that differed by two points, scores that three points, score that differed by four points, and indicators for which no core was recorded by at least one of the raters. Totals were calculated and percentage of agreement for all indicators recorded was recorded.

When alternate assessments judged against alternate achievement standards, it may be necessary to address local item dependency when student scores can reflect both "real" student performance as well as the level of assistance provided by the individual administering the assessment. Tests must also be reliable to be valid measures of a student's performance. An increase in the number of items (tasks) to be assessed might make activities more homogenous (Dunbar, Koretz, & Hoover, 1991) and therefore, influence the degree of reliability (see Fahey, Filbin, & Connolly, 2004), who reported that the development of the Colorado Student Assessment Program Alternate (CSAPA) was built by employing standardized assessment items and materials. In addition, numerous adaptations were accepted during the administration of the CSAPA. Content scaffolding, systematic instructional strategies, and the scoring of indicators depended on the teacher. To gauge the impact of these factors on score variance, the correlation between individual items (internal consistency) was calculated using Cronbach's alpha. The reliability coefficients spanned from .96 to .99.

Rudner and Schafer (2001) reported that most large-scale assessment coefficients ranged between .80 and .90. Using their yardstick the coefficients for the CSAPA suggested very strong internal consistency. They observed, however, that it was difficult to be sure that the strong coefficients were indicative of the homogeneity of the items and not a result of integrating skill performance and level of prompting. In addition, the strong coefficients might suggest a restriction in the range of items, since most students taking the CSAPA scored in the upper two performance categories. They concluded that the area warranted further investigation to assess effects of combining accuracy and degree of support on reliability measures.

Reliability of an Assessment System

An assessment system may be comprised of several alternatives, all with the goal of reaching a conclusion for the student. It is a good idea to see how the various components supplement each other and how reliable the system is toward the end. In this volume, we refer to an assessment system as a combination of various assessment approaches that provide a full range of participation options. Two definitions of "system" apply here.

- A complex whole formed from related parts: a combination of related parts organized into a complex whole, *a social system (MSN Encarta, captured 070205)*
- A set of things working together as a mechanism or interconnecting network. (Compact Oxford English Dictionary of Current English, Third Edition, June 2005 (captured at http://www.askoxford.com/concise_oed/ on 07/02/05).

An example of an assessment system is described on the Washington Department of Education website (http://www.k12.wa.us/assessment/default.aspx):

> The Washington State Assessment System (WSAS) is composed of three broad programs: statewide standardized testing; classroom-based assessments; and assessment staff development. The statewide testing program focuses on the Essential Academic Learning Requirements (EALRs), which are Washington's content standards, and provides broad achievement indicators for the state, districts, schools, and individual students. . .. The Washington Assessment of Student Learning (WASL) currently is comprised of a series of criterion-reference tests in reading, writing, mathematics, and science. These standards-based assessments incorporate three item types: selected response (multiple-choice); short constructed response; and extended constructed response. Performance

standards for the assessments in reading, writing, mathematics, and science have been set using an item mapping technique designed after that developed by researchers at CTB/ McGraw-Hill.

In this example, one aspect of the system contains three components: standardized testing, classroom-based assessments, and staff development. Within this broader system, the statewide assessment of achievement, the WASL, contains three formats across four subjects. In this paper, we extend the concept of assessment system beyond the subject areas and assessment formats to include assessment options within the cascade. Several issues appeared in this expanded view:

- Does one component of the system support the inferences of another? In the conjunctive/compensatory sense? That is, do the separate decisions within each component have to be present at some level to be counted toward the whole OR can some components of the system be used to compensate for others when coming to judgment.
- Must the technical adequacy among the elements of the system be comparable—whether made up of selected response, short constructed response, and extended constructed response as in the Washington model or made up of the general assessment with or without accommodations and the three forms of alternate assessment (based on grade level, modified achievement, or alternate achievement standards) as presented in the decision-framework?
- Regarding this system reliability, one wonders how the weakest link, e.g. least reliable or least replicable component, affects the entire system? Do estimates of reliability within one component compromise confidence in the other components? How are the reliabilities of constructed responses considered along with the reliabilities of selected

responses (given they often are lower)? How is the value of each component weighted in relation to the time spent in collecting the information and the resulting score?

- How do we assess fairness when some components of the system warrant greater confidence than others in the inferences that can be made?
- When determining Adequate Yearly Progress of individual student proficiency, do judgments based on components of the system benefit or disadvantage some subgroups more than others?

References

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

Boomsma, A., Van Duijn, M.A.J., & Snijders, T.A.B. (2001*). Essays on item response theory.* New York: Spring-Verlag.

Brennan, R. L. (2001). *Generalizability theory*. New York: Springer-Verlag.

Browder, D.M., Spooner, F., Algozzine, R., Ahlgrim-Delzell, L., Flowers, C., Karvonen, M. (2003). What we know and need to know about alternate assessment. *Exceptional Children*, 70(1), 45-61.

Chatterji, M. (2003). *Designing and using tools for educational assessment*. Boston: Pearson Education

Clapper, A.T., Morse, A.B., Lazarus, S.S., Thompson, S.J., & Thurlow, M.L. (2005). *2003 state policies on assessment participation and accommodations for students with disabilities* (Synthesis Report 56). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved [today's date], from the World Wide Web: http://education.umn.edu/NCEO/OnlinePubs/Synthesis56.html.

Crawford, L., & McDonald, M. (2003). *A reliability study of the 2002 administration of the third grade Colorado State Assessment Program Alternate (CSAPA) assessment in reading and writing.* Denver, CO: Colorado Department of Education Report.

Fahey, K., Filbin, J., and Connolly, T. (2004). *Colorado's Performance-Based Alternate Assessment: Historical Perspective and Lessons Learned Technical Report.* Retrieved from the world wide web at http://education.umn.edu/NCEO/StateForum/CSAPAtechmanual.pdf on June 18, 2005.

Feldt, L. S, & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 105-146). New York: American Council on Education/Macmillan.

Fleming, J.A., Taylor, J. M., & Carran, D. (2004). A comparison of two methods of determining interrater reliability. *Assessment for Effective Intervention*, *29*(2), 39-51.

Haertel, E. H. (October 1999). *Reliability: How to quantify error and how to communicate about it.* Handout for the Interactive Lecture Series. Dover, NH: National Center for the Improvement of Educational Assessment.

Hollenbeck, K., Tindal, G. (1999). Reliability and decision consistency: An analysis of writing mode at two times on a statewide test. *Educational Assessment*, *6*[1], 23-40.

Johnson, Evelyn (2003). *Analysis of Washington Alternate Assessment System Portfolios 2003*. Contact Nancy Arnold, Office of the Superintendent of Public Instruction, Washington Department of Education, NArnold@ospi.wednet.edu.

Kolen, M. J. & Brennan, R. L. (1995). *Test equating: methods and practices*. New York: Springer.

McGrew, K. S., Johnson, D. R., Cosio, A., & Evans, J. (2003). *Increasing the chance of no child being left behind: Beyond cognitive and achievement abilities*. Mpls, MN: University of Minnesota. Unpublished Paper.

Nitko, J. (1996). *Educational assessment of students (2nd edition)*. Englewood Cliffs, NJ: Merrill.

Oregon Department of Education (2005). *Juried Assessment 2004 – 2005: Guidelines for Meeting Oregon's Standards Using the Juried Assessment Process & Guidelines for Jurying a Modification.* Retrieved from the world wide web at http://www.ode.state.or.us/teachlearn/testing/admin/juried/juriedassmtmanual0405.pdf on June 15, 2005.

Rudner, L.M. & Schafer, W.D. (2001). *Reliability. ERIC Digest*. ERIC Clearinghouse on Assessment and Evaluation, Washington, D.C., downloaded 1/26/04 from http://wwww.ericfacility.net/databases/ERIC_Digests/ed458213.html.

Sireci, Stephen G., Li, Shuhong, and Scarpati, Stanley (2003). The Effects of Test Accommodation on Test Performance: A Review of the Literature. *Center for Educational Assessment Research Report*, 485, 0-0.

Thissen, D. & Wainer, H. eds. (2001). *Test scoring*. New Jersey: Lawrence Erlbaum Associates.

van der Linden, W.J., & Hambleton, R.K. eds. (1997). *Handbook of modern item response theory*. New York: Spring-Verlag.

Wise, S. L., Lukin, L. E., & Roos, L. L. (1991). Teacher beliefs about training in testing and measurement. *Journal of Teacher Education, 42*(1), 37-42.

Yovanoff, P., & Tindal, G. (in press). Scaling early reading alternate assessments with statewide measures. *Exceptional Children*.

CHAPTER 9: VALIDITY EVIDENCE

As elaborated by Messick (1989) in his extensive essay on test validity and rephrased in Chapter 2 of this volume, the validity argument involves a claim with evidence evaluated to make a judgment. Various preceding chapters describe three essential components of assessment systems: constructs (what to measure), the assessment instruments and processes (approaches to measurement), and use of the test results (for specific populations). To put it simply, validation is a judgment call on the degree to which each of these components is clearly defined and adequately implemented. Validity is a unitary concept with multifaceted processes of reasoning about a desired interpretation of test scores and subsequent uses of these test scores. In this process, we want answers for two important questions. Regardless of whether the students tested have disabilities, the questions are identical: (1) How valid is our interpretation of a student's test score? and (2) How valid is it to use these scores in an accountability system?

Validity evidence may be documented at both the item and total test levels. Since the adequacy of the items has been explored already in Chapter 7, this chapter focuses only on documentation of validity evidence at the total test level. At this level the *Standards* (AERA et al., 1999) call for evidence regarding content coverage, response processes, internal structure, and relations to other variables. Examples for each source of validity evidence are provided using illustrations from some large-scale assessment programs' technical documentation. We subsequently provide a discussion of each of these sources.

Evidence Regarding Content Coverage

In part, evidence of content coverage is concerned with judgments about "the adequacy with which the test content represents the content domain" (AERA et al., 1999, p. 11). The test is comprised of sets of items that sample student performance on the intended domains. The

expectation is that the items cover the full range of intended domains, with enough items so that scores credibly represent student knowledge and skills in those areas. Without enough items, the potential exists for a validity threat due to construct underrepresentation (Messick, 1989).

Once the purpose of the test and intended constructs are determined, the foundation of validity evidence from content coverage may come in the form of test blueprints or test specifications (see Chapter 6). Among other things, the *Standards* (AERA et al., 1999) suggest specifications should "define the content of the test, the number of items on the test, and the formats of those items" (Standard 3.3, p. 43). Test blueprints that include this information often are released to educators (and the public) via test manuals. Figure 1 (on the next page) provides an example of a test blueprint for fifth grade mathematics on a fixed format (pencil and paper) general assessment, given with or without accommodations.

What might test blueprints look like for alternate assessments? Those based on standardized tasks or items may include information like general assessments. For example, administration manuals for Texas' State-Developed Alternate Assessment II, a paper and pencil test, include test blueprints for reading, writing, and math by instructional level (grade or grade band). The blueprint for each subject indicates the number of items that are used to assess each objective in the Texas Essential Knowledge and Skills (Texas Education Agency, 2004). The manual briefly describes item formats elsewhere. As an example with less standardized tasks or items, the Educator Manual for Massachusetts' portfolio-based alternate assessment instructs teachers to include portfolio evidence that relates to specific grade level standards within the state's curriculum framework (Massachusetts Department of Education, 2004). Teachers have some flexibility in determining the specific standards for which evidence is provided (e.g., selecting three of the five possible standards). Maximizing the specificity and clarity of

instructions to teachers may help standardize the type of evidence for alternate assessments in general and portfolios in particular, thus allowing for more consistent interpretations of scores (as would be desired in large-scale assessment programs).

Figure 1. Palmetto Achievement Challenge Tests (PACT) Blueprint for Grade 5 Mathematics. (South Carolina Department of Education, n.d.)

Distribution of Items

Type of Item	Constructed Response	Multiple Choice	Total on Test
Number of Items	4-5	32-34	**35-39 items**
Value of Each Item	2-4	1	
Total on Test	**11-13 points**	**32-34 points**	**45 points**

Distribution of Points by Strand

Strand	Constructed Response Points	Multiple Choice Points	Points per Strand
Number & Operations	At least one CR item in at least 4 of the 5 areas for a total of:	The balance of the points will be multiple choice and distributed through the strands according to the total points listed in the next column.	11-13
Algebra			7-9
Geometry			8-10
Measurement			8-10
Data Analysis & Probability			7-9
TOTALS	**11-13 points**	**32-34 points**	**45**

When performance assessments, checklists, rating scales, or portfolios are used, student test scores are based on fewer items than on general assessments (and therefore a smaller sample of the intended domain). In these cases, it is especially important that the assessment items or pieces of portfolio evidence be representative of the intended domain (AERA et al., 1999, Standard 4.2) and specified in the assessment administration instructions. Otherwise, construct under-representation or misrepresentation is a likely threat to validity.

Evidence related to content coverage may become increasingly complex and nuanced, as

state assessment systems use universal design principles to shift from a finite series of discrete testing options to a broader continuum consistent with the cascade of options. Tables of specifications may become tailored to groups of students who have multiple paths of participation in the cascade of assessment options. In that case, high-quality items will serve as a foundation for content-related validity evidence at the assessment level. This topic represents an area in which considerable empirical evidence is needed.

Methods for Examining Content Coverage

The *Standards* (AERA et al., 1999) and other resources (see Downing & Haladyna, in press) provide overviews of procedures for examining the match between test specifications and a batch of items that comprise a test. The procedures test developers follow in specifying and generating test items should be well documented (Standard 1.6). In large-scale assessment systems, where multiple forms of each test are needed, item writing and reviewing may continue in an ongoing basis. Trained item reviewers with content area expertise may be given a set of items and asked to indicate which standard the item matches. Comparisons would then be made between item reviewers' ratings and the item classifications initially made by the test developer. A critical consideration in collecting evidence related to content coverage is the achievement standards upon which the assessment is based. The most tenable judgments of content match will be made if items are compared to specific objectives (i.e., the smallest "grain size") described in the appropriate grade level, modified, or alternate achievement standards.

Evidence of content coverage also may come from methods for investigating the alignment of standards and assessments. For instance, Webb's (1999) alignment criteria consider not only alignment at the item level (described in Chapter 7), but statistical indicators about the degree to which the test matches the standards. The *Range of Knowledge* correspondence

criterion indicates the number of objectives within the content standard with at least one related assessment item. In applying this indicator to statewide mathematics and science assessments, Webb determined that an acceptable Range of Knowledge exist when at least 50% of the objectives for a standard had one or more assessment items. In that study, states' assessments had anywhere from 0% to 100% of their standards meet the criterion for acceptable Range of Knowledge. Resnick, Rothman, Slattery, and Vranek's (2003) study of general assessments in five states used the test blueprint as the basis for the intended content standards and calculated a proportional range statistic like that of Webb. They found that states tended to have lower range scores than the criterion they established as acceptable.

The Range of Knowledge indicator also has been applied to alternate assessments. In an analysis of two states' alternate assessments, Flowers, Browder, and Ahlgrim-Delzell (in press) found very low levels of Range of Knowledge match (0% to 37% of standards met the criterion). In contrast, Roach, Elliott, and Webb (2005) found that most standards on one state's alternate assessments in reading, language arts, math, science, and social studies met or weakly met the criterion for acceptable Range of Knowledge.

While Webb's (1999) Range of Knowledge criterion indicates the extent to which items match specific objectives within a standard, the *Balance of Representation* criterion indicates the degree of proportionality with which assessment items are distributed across objectives within a standard. In other words, to what extent are items *evenly distributed* across objectives? Balance of Representation is a calculated index, with 1 indicating a perfect balance and values closer to zero indicating most items were aligned with only a few objectives within the standard. (See Flowers, Browder, Ahlgrim-Delzell, & Spooner, in press, or Tindal, 2005 for the Balance of Representation formula and calculation instructions). In an analysis of four states' mathematics

and science assessments, Webb (1999) used .7 as the criterion for an acceptable balance and found a high percentage of standards (71% to 100% across states and tests) met that criterion. Applying this criterion to one state's alternate assessment, Roach et al. (2005) found that all the standards had acceptable results for Balance of Representation. In contrast, Flowers, Browder, and Ahlgrim-Delzell (in press) found that only one of three states' language arts alternate assessment had any standards that met Webb's criteria for acceptable or weak balance of representation. The other assessments had 0% acceptable across the standards. Resnick et al. (2003) also investigated balance using qualitative judgments about "the relative importance of test items [given] to content and skills" (p. 20) in comparison with the importance stated in the standards. Based on responses to a set of open-ended questions, item reviewers rated the balance of sets of items on a scale (good, appropriate, fair, or poor). Raters found very few standards were rated appropriate or higher on language arts and math tests at the elementary, middle, and high school levels.

Criteria for Evidence Related to Content Coverage

Regardless of the type of assessment or the achievement standards upon which it is based, similar types of information should be included in test specifications to evaluate validity evidence related to content coverage. Standard 3.3 provides a comprehensive list of contents:

> The test specifications should be documented, along with their rationale and the process by which they were developed. The test specifications should define the content of the test, the proposed number of items, the item formats, the desired psychometric properties of the items, and the item and section arrangement. They should also specify the amount of time for testing, directions to the test takers, procedures to be used for test administration and scoring, and other relevant information. (AERA et al., 1999, p. 43)

While the same kinds of information should be provided across assessments, the criteria established to judge content coverage (e.g., enough items per standard) as acceptable may vary. For example, the more partitioned a state's content standards are, the more difficult it may be to have assessments meet Webb's (1999) suggested Range of Knowledge criterion; using only that criterion as a target for "good" coverage may result in tests that are prohibitively long. The level of specificity of the content standards (e.g., global standards vs. specific grade-level objectives) against which items are aligned also may influence the statistics that are obtained on content coverage (see Chapters 7 and 8). It is not until states develop evidence related to content coverage and consistently share that information with the psychometric and special education communities, that we will have a sufficient body of evidence to establish criteria of "acceptable" content coverage across the range of assessment options in large-scale assessment systems.

Evidence Regarding Response Processes

Considering again the range of assessment formats available for students with disabilities, response processes represent the cognitive behaviors required to respond to an item. In selected response and some constructed response items, such as those seen on many general assessments, the student's response process is the focus. However, when assessments require the judgment of teachers (e.g., in assembling a portfolio or completing a rating scale), or raters (e.g., of extended response items, performance assessments, or checklists), an understanding of that person's response process also contributes to validity evidence for test score inferences. When statements about cognitive processes of the examinee or of raters are included in validity arguments, evidence about those processes should be described (Standard 1.8). The sections below are intended to help guide the collection of such evidence.

Item Specification and Review Procedures

While test blueprints described in the previous section are based on content and item type, tables of specifications may also be developed by forming a matrix that includes content and cognitive demand. Each cell in the table shows the number of items used to assess each topic or strand, at each level of cognitive demand. States use various frameworks for describing cognitive demand, each with a different number of levels and accompanying descriptors. For instance, the Minnesota Comprehensive Assessment Series II (MCA-II) test specifications use three levels of cognitive demand based on Bloom's taxonomy for aligning its reading and math assessments with state standards (Minnesota Department of Education, 2005). The blueprint for South Carolina's PACT tests in science describes six levels of cognitive demand (South Carolina Department of Education, 2001). Webb's (1999) alignment criterion for depth of knowledge includes four levels. Haladyna, Downing, and Rodriguez (2002) highlight the need for a taxonomy for classifying items based on cognitive demand that is empirically based and more universally used. While this suggestion was made in the context of classroom assessment, a well-founded taxonomy would also be useful for validity studies of large-scale assessments. Figure 2 illustrates an excerpt from the MCA-II test specifications for third grade reading. These specifications include not only the content and item type, but also the cognitive levels (A, B, or C).

Figure 2. Sample MCA-II Test Specifications for Third Grade Reading. (Minnesota Department of Education, 2005)

Item Types	Cognitive Levels	Strand 1 – Reading and Literature Sub strand C. Comprehension Standard: The student will understand the meaning of texts using a variety of comprehension strategies and will demonstrate literal, interpretive, and evaluative comprehension.	Item Totals
MC or CR	A B	**I.C.4** The student will retell, restate, or summarize information orally, in writing and through graphic organizers.	0-2
MC or CR	B C	**I.C.5** Students will infer and identify main idea and determine relevant details in nonfiction text.	1-3

Response Processes of the Student

Student response processes to most general assessment items (selected response, short constructed response) may be considered through typical item review procedures. Test developers may start with a table of specifications, and external reviewers' judgments about content and cognitive demand for each item may be compared with what was intended by the test developer. As with any other type of expert review process, reviewers should be well-trained, and their selection, expertise, training, and rating procedures should be thoroughly documented (Standard 1.7).

While expert ratings of cognitive demand provide some indication of students' use of lower- and higher-order thinking skills on the assessment, it does not rule out the possibility that the student's response is based on something other than what was intended. Haladyna (1999) describes a range of procedures designed to determine students' cognitive processes when responding to paper and pencil test items. One commonly used method is a think aloud procedure, in which students verbalize everything they think about while working through a problem.

Cognitive processes may be more difficult to assess directly for alternate assessments aligned to alternate achievement standards. Think aloud procedures are not feasible when students cannot communicate verbally, and interview methods -- even with the benefit of assistive devices for students who communicate nonverbally -- may not work when students are still developing metacognitive awareness. In these cases, direct observational data collection methods may be needed to document possible cognitive processes. Administration of performance assessments may be videotaped, for example, and analyzed for behavioral evidence of processes. Videotaped accounts of instructional activities that yield products included in a portfolio also may provide supportive evidence of the student's response processes. Obviously, such resource-intensive data collection methods would be appropriate for empirical studies about the assessment rather than a component of the assessment administration for all students.

One unique approach to standardizing the cognitive demand in alternate assessment is seen in Pennsylvania's Alternate System of Assessment (PASA Project, 2003), in which a series of otherwise standardized performance tasks are administered in one of three ways, depending upon the level of cognitive demand that is determined to be appropriate for a student by the teacher. To encourage high expectations for student performance, the level of cognitive demand is incorporated into the student's score.

Response Processes of Teachers and Raters

When someone other than the student is partially responsible for the responses to an assessment, that person's cognitive behavior also exerts some influence on the student's score. Studies on one state's portfolio-based alternate assessment support this assumption, as teacher training and understanding of the portfolio scoring system was associated with strong assessment scores (Browder, Karvonen, Davis, Fallin, & Courtade-Little, 2005; Karvonen, Flowers,

Browder, Wakeman, & Algozzine, in press). When a teacher administers a performance assessment, completes a checklist, or assembles a portfolio, the teacher's exact responsibilities should be delineated in the assessment's specifications. As discussed in Chapter 8, the influence of judges and raters attenuates measurement reliability, which in turn impacts validity.

When assessment administration procedures are more standardized, there is less opportunity for the teacher's cognitive processes to influence the student's score. Clearly written, easy-to-follow scripts may help teachers adhere to a prescribed sequence of minimally intrusive prompting on performance tasks. Another basic method to maximize standardization is by having a neutral monitor present during the administration of performance tasks (South Carolina Department of Education, 2005).

As they are used with students, think-aloud procedures also may be used with teachers and raters. Teachers may be interviewed about the administration of a performance assessment or asked to think aloud about completing a checklist or rating scale, or about compiling a portfolio-based alternate assessment. Think-aloud procedures may be used to examine how raters follow the prescribed rating process, and possible places where failure to follow the process raises questions about validity of the score interpretation (Heller, Sheingold, & Myford, 1998). Post-hoc interrater agreement indices also may be used to assure consistency in rating a performance or scoring a product (see Chapter 8 for more information on rater agreement).

Criteria Related to Evidence Based on Response Process

Considerations in evaluating criteria related to response processes are like those for content coverage. While well-established methods exist to collect this evidence, there is not yet a sufficient body of empirical evidence to provide clear standards for what is acceptable. Early studies on alignment (Flowers, Browder, & Ahlgrim-Delzell, in press; Resnick et al., 2003;

Roach et al., 2005; Webb, 1999) provide some possible benchmarks but should not be considered gold standards at this point. States should be aware of methodological problems in considering these sources of evidence that might attenuate statistics obtained (e.g., judging the alignment on Webb's depth of knowledge criterion when comparing specific assessment items to vaguely worded content standards).

Evidence Regarding the Internal Structure of the Test

The *Standards* (AERA et al., 1999, pp. 13-15) call for a study on internal structure as part of test validation. For valid test score interpretations and validity generalization, it is expected that (a) the items show some level of internal consistency (Standard 1.11); (b) the internal structure of the test remains stable across major reporting groups (p.15); and (c) the internal structure of the test remains stable across alternate (and hopefully equivalent) forms of the same test (pp. 51-52). This section of the chapter addresses these three topics. (Note, in Chapter 8, these issues are addressed in the context of assessment reliability.)

Inter-Item Consistency

With an assessment consisting of items measuring the same construct or tapping the content standards in the same subject area (like reading or mathematics), it is expected that these items show some level of consistency among themselves. In other words, it is desirable that student responses to various items or parts of the test go together in some reasonable fashion, and do not contradict each other in any substantial way.

Internal consistency can be checked by looking at the inter-item correlation matrix. It is desirable that all inter-item correlations are positive. Internal consistency is also manifested in overall test reliability indices such as KR20 or Cronbach's alpha. Phillips (2000) reported that a coefficient alpha of at least 0.85 was judged as adequate for standardized tests such as Texas'

high school graduation test. It should be pointed out, however, that the variability among students taking an alternate assessment is typically not as large as the variability among students taking the general assessment. Therefore, the coefficient alpha for a modified assessment may be lower than the 0.85 threshold.

Inter-Strand Consistency

An assessment (based on grade level, modified, or alternate achievement standards) may be designed to measure knowledge and skills in several separate content strands. In mathematics, for example, the strands may range from simple computations (operations) to more complex topics such as probability and statistics. In other situations, like assessment of writing, a strand may be one aspect of writing (such as expository or persuasive), and various strand scores are sometimes obtained by scoring the same student response using different rubrics. Since these separate strands are parts of a larger construct, it is desirable that they are somewhat related to each other (but not exceedingly so). Inter-strand correlations are expected to be positive and moderate. High inter-strand correlations are not desirable because the strands may essentially reflect very similar types of skills or abilities. Table 1 provides an illustration of data on between-strand correlations for the South Carolina PACT assessments in mathematics in 2003 (South Carolina Department of Education, 2003).

Table 1. Inter-strand Correlation Matrix for PACT Mathematics Assessments

MATHEMATICS					
Strand	**N**	**Mean**	**Standard Dev.**	**Minimum**	**Maximum**
Number & Operations	50,070	8.901	3.849	0	17
Algebra	50,070	10.832	4.113	0	18
Geometry	50,070	7.238	2.901	0	16
Measurement	50,070	2.909	2.168	0	8
Data Analysis & Probability	50,070	5.158	2.504	0	13
Between-Strand Pearson Correlation Coefficient Matrix					
	Number & Operations	Algebra	Geometry	Measurement	Data Analysis & Probability
Number & Operations	1.000	0.764	0.655	0.665	0.661
Algebra	0.764	1.000	0.654	0.644	0.661
Geometry	0.655	0.654	1.000	0.608	0.607
Measurement	0.665	0.644	0.608	1.000	0.606
Data Analysis & Probability	0.661	0.661	0.607	0.606	1.000

Unidimensionality

It is of considerable importance to check the number of dimensions (constructs) as operationally measured by the assessment and then reflected in the test data. Subject areas such as reading, mathematics, and science (at the lower grade levels) are typically thought of as single constructs; however, student performance on the assessment items may be contingent on other unintended and irrelevant factors. Performance on constructed response items in mathematics, for example, may dependent partially on reading level. Likewise, to solve a science problem, reading and writing skills as well as some knowledge of mathematics may be essential.

Assessments based on alternate and modified achievement standards typically are given to students with a wide range of disabilities and academic training, so the internal structure of the assessment instrument may be more complex than for the case of typical students assessed on grade level achievement standards under standardized conditions. Performance tasks administered with scaffolded assistance may reveal both the ability of the student and the

assessor's judgment regarding the level of prompting needed by the student to get the correct responses. Such judgment may bring an extra dimension to be considered in any interpretation of the test data.

Interpretation of test scores is more straightforward when the assessment taps into one unique dimension (construct) like reading or mathematics. Unidimensionality may be checked via a variety of statistical methods ranging from classical principal component analysis (PCA) or factor analysis (FA) techniques, to more modern methods based on item response theory (IRT). The scree test in PCA and FA has been used quite often in this regard. There are rules of thumb regarding unidimensionality. Reckase (1979), for example, noted that the unidimensionality condition may be met if the first (largest) eigenvalue accounts for more than 40% of the total test variation. Other IRT techniques include the DIMTEST and similar procedure developed by Stout and his associates (Stout, 1987; Nandakumar & Stout, 1993). evidence.

Similarity of Factorial Structure and DIF Across Accommodated and Non-Accommodated Conditions

Many authors, such as Geisinger (1994), have noted major measurement issues when students are assessed by standardized tests that have been administered under non-standard conditions. By adhering to standard procedures, errors in scores can more likely be attributed to random or individual errors, rather than administrative errors, and scores can be interpreted similarly for all participants (Geisinger, 1994). Some students with disabilities require accommodations to allow an assessment to tap into a student's ability on the construct(s) being measured, curtailing the effect of a student's disability on his/her test result. A true accommodation should allow a student to be assessed in such a way that a disability does not misrepresent the student's true performance, hence leveling the playing (or testing) field.

Students should ideally be placed on equal footing and not advantaged or disadvantaged because of a disability, and the interpretations of their scores should be valid for the test's intentions.

In her landmark writing on the balance between the individual rights of the student with a disability and the integrity of the testing program, Phillips (1994, p. 104) indicated the need to address a number of questions including the following two: (1) Does accommodation change the skill being measured? and (2) Does accommodation change the meaning of the resulting scores?

A host of psychometric issues need to be considered in addressing these two questions. Willingham (1989) pointed out that, to achieve score comparability, various forms of the assessment need to display similar factor structures, and no differential item functioning (DIF) should exist across student groups. In a long review of psychometric, legal, and social policy issues about test accommodations for examinees with disabilities, Pitoniak and Royer (2001) concluded that further legal decisions would determine assessment accommodation policies and that more research is needed on test comparability.

Analysis of factor structures may be conducted via the PCA, FA, or IRT techniques previously described. DIF analysis may be performed via processes described in Chapter 7 on the quality of test items. It should be mentioned that evidence regarding similarity in factor structure and DIF are more easily collected when the same assessment is administered under varying conditions, as separate test forms. This is not usually the case where the accommodated form is the same as the regular form. Often there are significantly fewer numbers of accommodated versus non-accommodated students, and still fewer when breaking out analyses by accommodation, limiting the appropriateness of conducting PCA, FA,

IRT, or DIF analyses. In addition, most students receive multiple accommodations in each administration.

However, when accommodated and non-accommodated forms are developed and constructed to reflect the same strands (content standards), PCA or FA may be carried out at the strand level rather than at the item level. This type of analysis was used in two studies based on the South Carolina High School Exit Examination (Huynh, Meyer, & Gallant-Taylor, 2004; Huynh & Barton, in press). An analysis of the similarity of inter-item and inter-strand correlations may also provide evidence on the similarity of the construct assessed by the instruments across various groups.

Similarity of Factorial Structure and DIF Across Major Reporting Groups

When data are available, it may be of interest to determine whether a test's factorial structure is similar across important reporting groups such as students with varying disabilities. An example of this type of analysis may be found in Huynh and Barton (in press), who compared the factor structure of the same test across groups of students with physical disabilities, learning disabilities, and emotional disabilities. It is noted that this type of evidence may be hard to compile with small reporting groups. In this case, perhaps a consensus agreement via "juried" assessment, like one used in Oregon, may be the best choice (Oregon Department of Education, 2004).

Special Considerations for Tests with Cross-Grade Items: Internal Structure and DIF

As noted in earlier chapters of this volume, an assessment based on modified achievement standards may be narrower in content breath and coverage than an assessment based on general achievement standards. Under some major conditions presented in Chapter 7, an alternate assessment based on modified achievement standards may be comprised of several

"*cross-grade*" items that were originally designed for students at other grade levels. As discussed in Chapter 7, cross-grade items may be used under certain conditions to enlarge the item bank for the purpose of test form construction.

In some cases, it may be possible to find a section of the assessment based on grade level achievement standards (GLAS) that is similar to the assessment based on modified achievement standards (MAS) either at the item level (all items are identical) or at the strand level (all strands are the same). In these cases, it may be possible to check the similarity between the GLAS section and the MAS section in terms of factorial structure. When enough data are available, it may be possible to check for DIF between students who took the GLAS section and those who took the test with MAS.

Similarity in Types of Errors Made by Students

Evidence of similarity of constructs across major reporting groups may also be collected by analyzing the major types of errors made by students on various assessments administered to these groups. A study along these lines for the South Carolina High School Exit Examination was reported in Barton and Huynh (2003). Overall, the question is "Are the types of errors made by students taking the general assessment with accommodations or students taking the test based on modified achievement standards similar to those of general education test takers at a similar grade level, level of mastery, etc.?" Similarity in the errors is a good indicator that the assessment instruments tap into similar constructs.

Evidence Based on Relations to External Variables

The *Standards* (AERA et al., 1999, pp. 13-15) also call for validity evidence based on relations to other variables. External evidence for the construct being measured may be found in the relationship between the test and other similar or dissimilar measures. Evidence of this type

are sometimes referred to as "convergent" and "divergent," respectively. For example, a reading assessment should yield scores that are more closely related to other reading scores than to math scores (convergent evidence). Similarly, a math assessment should yield scores that are more closely related to other math scores than to reading score.

The table below, for example, reports correlational data on the relationship between Palmetto Achievement Challenge Tests (PACT) scores and scores on the TerraNova tests for grades three and six. The table is extracted from the 1999 PACT Technical Documentation written by Huynh, Meyer, and Barton (2000). In this table, PACT scores are scale scores, and TerraNova subtest scores are normal curve equivalents. The data indicate that PACT English Language Arts (ELA) assessments relate more strongly to the TerraNova reading and language components than to the TerraNova mathematics component. Similarly, PACT mathematics assessments relate more strongly with the TerraNova mathematics component than with the other two TerraNova components. However, the divergent evidence for validity is stronger for PACT mathematics assessments than for ELA assessments.

Table 2. Correlation Between the PACT and the TerraNova for Grades 3 and 6

Grade/Content	**TerraNova**		
	Reading	Language	Math
Grade 3 ELA	.776	.768	.729
Grade 3 math	.689	.687	.803
Grade 6 ELA	.755	.754	.741
Grade 6 math	.710	.705	.863

Conclusions and Recommendations

Together with the information in Chapter 7 about validity evidence at the item level, this chapter provides the reader with a set of methods for collecting evidence to support the validity of score interpretations in large-scale assessment systems. The good news is that many of the techniques for examining evidence related to content coverage, response processes, internal

structure, and relationships between test scores and other variables, are well established in the psychometric literature. The challenge will come in determining what constitutes "good" validity evidence using each of the techniques described in this chapter. At the time this chapter is being written, there is not a sufficient body of theoretical and empirical evidence to recommend minimally acceptable values of statistical indicators for all the sources of validity evidence. Especially for some assessments that are yet to be designed (e.g., alternate assessments based on modified achievement standards) and for assessments that may be taken by small groups of students, new, small-sample techniques for gathering and assessing validity evidence are needed. Following are some recommendations for collecting validity evidence at the assessment level:

1. States need to collect and document validity evidence for all four general areas described in this chapter: content coverage, response processes, internal structure, and relations to other variables. Even if strong validity evidence is collected in one or two areas, the failure to collect evidence across all four areas will weaken arguments that may be made for the validity of score interpretations.

2. Validity evidence for content coverage and response processes should be collected with the greatest precision possible. Blueprints for assessments should clearly state the strands or sub-strands within each content area, the types of items used to assess those strands, and the levels of cognitive demand required to respond to the items. Evidence of content coverage will be stronger if links can be made between specific achievement standards (grade level, modified, or alternate) and assessment items. Evidence collected through alignment methods that consider breadth and depth of content coverage will provide a richer understanding of content match than item-level judgments alone. States will need to carefully consider methods to assess response processes to rule out possible problems with construct-irrelevant variance. The response

processes of teachers, raters of performance assessments, and students with the most significant cognitive disabilities who complete tasks for performance-based alternate assessments, should be considered as well.

3. Regarding evidence related to the internal structure of the assessment, evidence may include inter-item correlations, inter-strand correlations, summary data from test dimensionality, similarity of the test's internal structure across major reporting groups, and the nature of the errors made on the test by major reporting groups.

4. Validity evidence collected via relationships between scores on the target assessment and external measures should include both relationships with similar measures (convergent evidence) and dissimilar measures (divergent evidence). Correlational analyses of scores on these measures may be used to make judgments about the quality of evidence from relationships with external measures.

5. It will continue to be of great importance to thoroughly document the backgrounds of students who are involved in the field-testing process (Standard 1.5). Disability labels alone may not be good proxies on which to base assumptions about group homogeneity. Careful documentation of testing accommodations will also be needed (Standard 1.13) for subsequent analysis of the potential unintended impact of accommodations (e.g., construct irrelevant variance).

In general, well-established data collection guidelines should be followed (Downing & Haladyna, 1997) and validity evidence should be clearly documented. The challenges associated with collecting new types of evidence should not discourage the continued study and collection of item and test level validity evidence for *all* types of assessments.

References

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

Barton, K., & Huynh, H. (2003). Patterns of errors made by students with disabilities on a reading test with oral reading administration. *Educational and Psychological Measurement, 63*, 602-614.

Browder, D. M., Karvonen, M., Davis, S., Fallin, K., & Courtade-Little, G. (2005). The impact of teacher training on state alternate assessment scores. *Exceptional Children, 71,* 267-282.

Downing, S. M., & Haladyna, T. M. (1997). Test item development: Validity evidence from quality assurance procedures. *Applied Measurement in Education, 10,* 61-82.

Downing, S. M., & Haladyna, T. M. (Eds.). (in press). *Handbook on test development*. Mahwah, NJ: Lawrence Erlbaum Associates.

Flowers, C., Browder, D., & Ahlgrim-Delzell, L. (in press). The alignment of three states' alternate assessments to state standards. *Exceptional Children.*

Flowers, C., Browder, D., Ahlgrim-Delzell, L., & Spooner, F. (in press). Promoting the alignment of curriculum, assessment, and instruction. In D. M. Browder & F. Spooner (Eds.), *Teaching reading, math, and science to students with significant cognitive disabilities*. Baltimore: Brookes.

Geisinger, K. F. (1994). Psychometric issues in testing students with disabilities. *Applied Measurement in Education, 7,* 121-140.

Haladyna, T. M. (1999). *Developing and validating multiple-choice test items* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education, 15,* 309-334.

Heller, J. I., Sheingold, K., & Myford, C. M. (1998). Reasoning about evidence in portfolios: Cognitive foundations for valid and reliable assessment. *Educational Assessment, 5*, 5-40.

Huynh, H., & Barton, K. (in press). Performance of students with disabilities under regular and oral administrations for a high-stakes reading examination. *Applied Measurement in Education*.

Huynh, H., Meyer, P., & Barton, K. (2000). *Technical documentation for the 1999 Palmetto Achievement Challenge Tests of English Language Arts and Mathematics, Grades Three through Eight*. Columbia, SC: South Carolina Department of Education, Office of Assessment. Retrieved July 13, 2005, from http://www.myscschools.com/offices/assessment/Publications/Index_of_Technical_Reports.htm

Huynh, H., Meyer, P., & Gallant-Taylor, D. (2004). Comparability of student performance between regular and oral administration for a mathematics test. *Applied Measurement in Education, 17*, 39-57.

Karvonen, M., Flowers, C. P., Browder, D. M., Wakeman, S., & Algozzine, B. (in press). A case study of the influences on alternate assessment outcomes for students with disabilities. *Education and Training in Developmental Disabilities*.

Massachusetts Department of Education (2004). *2005 educator's manual for MCAS-Alt*. Malden, MA: Author. Retrieved June 25, 2005, from http://www.doe.mass.edu/mcas/alt/05edmanual.pdf

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

Minnesota Department of Education (2005, March). *Minnesota Comprehensive Assessments Series II: Test specifications for reading and mathematics*. Roseville, MN: Author. Retrieved July 5, 2005, from http://education.state.mn.us/content/087562.pdf

Nandakumar, R., & Stout, W. (1993). Refinement of Stout's procedures for assessing unidimensionality. *Journal of Educational Statistics, 18*, 41-68.

Oregon Department of Education (2004). *Juried assessment 2004-2005: Guidelines for using the juried assessment process and guidelines for jurying a modification*. Salem, OR: Author. Retrieved June 25, 2005, from http://www.ode.state.or.us/teachlearn/testing/admin/juried/juriedassmtmanual0405.pdf

PASA Project (2003). *Pennsylvania alternate system of assessment administrator's manual*. Pittsburgh, PA: Author. Retrieved June 25, 2005, from http://www.pattan.net/files/instruction/adminmanual.pdf

Phillips, S. E. (1994). High-stakes testing accommodations: Validity versus disabled rights. *Applied Measurement in Education, 7,* 93-120.

Phillips, S. E. (2000, April). Legal corner: GI Forum v. TEA. *NCME Newsletter, 8*(2), n.p.

Pitoniak, M., & Royer, J. (2001). Testing accommodations for examinees with disabilities: A review of psychometric, legal, and social policy issues. *Review of Educational Research, 71*, 53-104.

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics, 4,* 207-230.

Resnick, L. B., Rothman, R., Slattery, J. B., & Vranek, J. L. (2003). Benchmarking and alignment of standards and testing. *Educational Assessment, 9*, 1-27.

Roach, A. T., Elliott, S. N., & Webb, N. L. (2005). Alignment of an alternate assessment with state academic standards: Evidence for the content validity of the Wisconsin alternate assessment. *The Journal of Special Education, 38*, 218-231.

South Carolina Department of Education (n.d.). *Blueprint construction of PACT in mathematics.* Columbia, SC: Author. Retrieved June 25, 2005, from http://www.myscschools.com/offices/assessment/Publications/PACTblueprints.htm

South Carolina Department of Education (2001). *PACT science assessment: A blueprint for success.* Columbia, SC: Author. Retrieved June 25, 2005, from http://www.myscschools.com/offices/assessment/Publications/PACTblueprints.htm

South Carolina Department of Education (2003). *Technical documentation for the 2003 Palmetto Achievement Challenge Tests of English Language Arts, Mathematics, Science, and Social Studies*. Retrieved July 13, 2005, from http://www.myscschools.com/offices/assessment/Publications/PACT-Tdoc03.doc

South Carolina Department of Education (2005). *High school assessment program alternate assessment (HSAP-Alt) test administration manual*. Columbia, SC: Author. Retrieved June 25, 2005, from http://www.myscschools.com/offices/assessment/Publications/HSAPAltTAM030805.pdf

Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika*, *52*, 589-617.

Texas Education Agency (2004, August). *State-Developed Alternative Assessment II information booklets*. Austin, TX: Author. Retrieved June 25, 2005, from http://www.tea.state.tx.us/student.assessment/resources/guides/sdaa/index.html

Tindal, G. (2005). *Alignment of alternate assessments using the Webb system: An abbreviated version*. Washington, DC: Council of Chief State School Officers.

Webb, N. L. (1999). *Alignment of science and mathematics standards and assessments in four states*. (NISE Research Monograph No. 18). Madison, WI: University of Wisconsin-Madison, National Institute for Science Education. (ERIC Document Reproduction Service No. ED440852)

Willingham, W. W. (1989). Standard testing conditions and standard score meaning for handicapped examinees. *Applied Measurement in Education, 2,* 97-103.

CHAPTER 10: SCORE REPORTING

In this chapter we discuss practices and possibilities in reporting the performance and achievement of students with disabilities when they participate in large-scale assessments. We begin the chapter with an overview of score reporting for general education content area assessments and achievement of students with disabilities when they participate in the array of options for participation in large-scale assessments (as was described in Chapter 4). We highlight both norm- and criterion-referenced interpretations and then consider both the role of achievement level descriptors in score reporting and the centrality of standard setting in this process as well as the issues to consider in conceptualizing achievement standards for alternate assessments. In communicating these results, we address the role of graphicacy in communicating results and use the *National Assessment of Educational Progress* as an example of a score reporting system that integrates many of the issues we raise. We end this chapter by considering several issues and inferences when including students with disabilities in large-scale assessment score using various assessment approaches.

Overview

The goal in score reporting is to communicate student performance and achievement in relation to specified:

- Grade-level content standards that indicate what students know and can do.
- Achievement standards that indicate how well students know the content and well they can demonstrate the skills specified in the content standards.

The ways in which a student's performance and achievement are communicated must be readily interpretable by parents, teachers, and others responsible for the education of students. The most important outcome from any score report is that the audience is informed about the

outcomes and the inferences that can be made. In reporting scores, therefore, the outcomes need to be contextualized: by the standards, the tasks, the student, the decisions, and changes over time in relation to base rates or prior levels. To communicate effectively, score reports must specify:

- *The grade-level content standards that are targeted by the assessment.* This information is necessary so that parents, teachers, and other educators can make valid inferences about what students know and can do in relation to the specified content standards. Because most alternate assessments modify the breadth, depth, and/or complexity of the grade level content standards, score reports must be explicit about the content covered in the assessment.
- *The achievement standards against which student achievement and performance are evaluated***.** This information is necessary so that parents, teachers, and other educators can make valid inferences about how well students know what they know and how well they have learned the skills specified in the content standards. In previous chapters we discussed grade-level content standards that could be judged against grade level achievement standards, modified achievement standards, and alternate achievement standards. The cut scores assigned to form these achievement standards need to not only reflect integrity of thee core levels (basic, proficient, and advanced) within a set, but also the relationships among the various sets. Score reports must define explicitly the meanings of the achievement/performance categories.
- *Test administration conditions under which students are assessed.* This information is necessary so that parents, teachers, and others can make valid inferences given the conditions in which a student was assessed. In previous chapters we discussed test

administration conditions, including assessments that are administered under standardized conditions (i.e., general education content area assessments administered with or without accommodations, rating scales and observational checklists, collections of performance tasks, and performance events), and assessments that are administered under less standardized conditions (i.e., portfolios). Because the performance of students with significant cognitive disabilities is likely to be dependent upon the conditions under which the performance was elicited, score reports also should specify those conditions explicitly. This supplemental information is important in making inferences about what students know and can do.

Score Reporting[1]

Generally, score reports provide information about student performance and achievement at the total test level (i.e., for a content area). They also may provide information for subscales (e.g., for decoding versus reading comprehension) and can include other information about student performance and achievement (e.g., at the task or item level). Score reports provide information about the contents and formats of the assessment and student identifying information. They also may contain information to support either of both of the following interpretative references.

Norm-referenced interpretations about what students know and can do, so that parents and educators can compare student performance and achievement to those of other student groups (e.g., students in the same grade). The important reference in this interpretation is the student's relative standing in a group of comparable (age or grade-based) students. Several metrics are possible for reporting and displaying this relative standing. Typically, a test reports outcomes using some type of scale score that is expressed on a bell-shaped curve with a mean

[1] Adapted from Ferrara & DeMauro (forthcoming)

and standard deviation. In reporting this information, it is critical to include both values as a student's relative standing is not interpretable without knowing the variance in the group. Graphic displays for norm-referenced interpretations may highlight the students' absolute values in the scale or their (percentile) rank order. In either instance, some interpretive guidelines are important (e.g., what the standard deviation means, or how percentile ranks are to be interpreted).

Criterion-referenced interpretations show what students know and can do, so that parents and educators relative to an achievement standard (e.g., Basic, Proficient, Advanced). Probably the most significant difference with this type of reporting (compared to norm-referenced reporting) is that levels of proficiency are implied relative to skill strands. Not only is the overall score not as important (in value) because what counts is presence within a category (basic, proficient, or advanced), but the reporting metric changes (from scale scores to percentage of students within these various categories). The score reports, therefore, need to consider the kinds of items or strands as well as the values of proficiency within these categories. Importantly, both kinds of information are needed as one without the other makes an interpretation difficult. Graphic displays often are much easier to create for criterion referenced interpretations categorical (using bar charts) as the information is primarily nominal (e.g., percent proficient).

The Role of Achievement Level Descriptions in Score Reporting

A crucial step in establishing achievement standards is specifying what is meant by *proficient* performance. Typically, performance, proficiency, or achievement level descriptions – ALDs, for short—describe the knowledge and skills that students at each level are believed to have attained. That belief is an inference based on student scores, which place them in an achievement level on a specific test, administered under specified conditions. The ALD enables

parents and educators to infer that a specific student has displayed the knowledge and skills described in the ALD based on the student's test performance. If additional evidence exists about the generalizability of scores from that test, the ALD may enable parents and educators to make inferences about that student's ability to demonstrate similar knowledge and skill in contexts beyond the testing situation.

In the standard setting methods described below, judges often are trained to consider a borderline individual or group. On occasion, this borderline individual or group is real and at other times the individual or group is hypothetical. In both cases, the most important consideration is to focus on the minimal difference between two categories: for example, between advanced and proficient or between proficient and basic. The extreme in any category often is less critical for judges to consider than the students who are minimally different from each other. Because this process is used for standard setting, it is important for score reports to include some of this information not only in the technical descriptors (documenting the procedural evidence of the process for setting standards) but also in the reports for communicating to the public.

Setting Achievement Standards

Probably the most critical issue in reporting scores is related to the standards being applied. They form the core value for interpretation. If the standard setting process is flawed, it is difficult for any report to meaningfully communicate results. Although some standard setting systems have been present for decades, the finer elements behind their implementation or their application in more nuanced environments (both in terms of their application to items and individuals) is relatively recent.

Generally, standard setting has been framed as either examinee-centered or test-centered (Taylor & Nolen, 2005). Student-centered approaches focus on populations of students while test-centered approaches focus on items or strands. In either case, the goal is to establish cut scores that establish different proficiency classifications.

- The standard setting method should be appropriate for the assessment approach.
- Procedures should be followed rigorously so that the achievement standards are credible.
- Credibility of achievement standards rests on several requirements: Fidelity and rigor of the procedure, quality and credibility of the performance/proficiency level descriptors, and the qualifications of the standard setting panelists.

Cizak (2004) lists the following 'generic' steps in setting performance standards (Table 3; p. 36)

1. Select a large and representative panel.
2. Choose a standard setting method; prepare training materials and standard setting meeting agenda.
3. Prepare descriptions of the performance categories (i.e., ALDs).
4. Train participants to use the standard setting method.
5. Compile item ratings or other judgments from participants and produce descriptive/summary information or other feedback for participants
6. Facilitate discussion among participants of initial descriptive/summary information.
7. Provide an opportunity for participants to generate another round of ratings; compile information and facilitate discussion as in Steps 5 and 6.
8. Provide a final opportunity for participants to review information and arrive at final recommended performance standards.

9. Conduct an evaluation of the standard setting process, including gathering participants' confidence in the process and resulting performance standard(s).

10. Assemble documentation of the standard setting process and other evidence, as appropriate, bearing on the validity of resulting performance standards." (Source: Hambleton, 1998[2]).

In his summary of standard setting methods, Cizak (2004) describes the following methods.

Bookmark method. Initially introduced by Lewis, Mitzel, & Green (1996), and referred to as a process in which participants identify cut scores by placing markers in a booklet that includes items ordered by difficulty (which then becomes the ordered bookmark method). This method builds off the psychophysics notion of 'just noticeable differences' (JND) in which a continuous gradient has a perceptible threshold that becomes translated into cut scores. The items only need represent the breadth and depth of the content of the test and may include the item (for selected responses) or prompt, rubric, and illustrative responses (for constructed responses). In the probability judgments, participants judge the likelihood that an examinee at the borderline of two performance levels can correctly answer the last item or respond to the prompt (or a fixed probability is used of .67 or .50 depending upon the use of a 2-parameter or 1 parameter model with item response theory). These judgments are then translated into a cut score.

Angoff variations. This set of procedures comes from Angoff (1971) and requires participants to make judgments for each item on whether a borderline student (or a hypothetical target group) is likely to answer it correctly (yes) or incorrectly (no). After a thorough review of the standards and the test (and possibly taking it), participants review the items in two phases

[2] Hambleton, R. (1998). Setting performance standards on achievement tests: Meeting the requirements of Title 1: In L. N. Hansche (Ed.) *Handbook for the development of performance standards* (pp 87-114). Washington, DC: Council of Chief State School Officers.

(with feedback from the first phase). The cut score then becomes the sum of items judged as correctly answered.

Holistic methods. In this method describe by Plake and Hambleton (2001), participants are asked to classify carefully selected materials in basic, proficient, and advanced categories. The cut score becomes the average from the borderline group. An example of a holistic approach is the *body of work* method (Kingston, Kahl, Sweeney, & Bay (2001) that applies logistic regression in calculating the cut score.

Student centered methods. Taylor and Nolen (2005) add two more types of standard setting procedures that are based on teachers' judgments of students. Teachers who are familiar with the standards and test are asked to classify a list of students with whom they are familiar into three groups: proficient, not proficient, and borderline. After the students have taken the test, their scores are analyzed in either of two ways. In the *contrasting groups method*, the cut score is the intersection of the range of scores for two extreme groups (high scores for the not proficient students meet the low scores for the proficient students). For the *borderline group method*, the cut score is the mean or median of the borderline students.

Irrespective of the method for setting standards, the process needs to be evaluated. "Typically, two evaluations are conducted during a standard setting meeting. A first evaluation normally occurs after the initial orientation of participants to the process (training in the method), and, where appropriate, administration to participants of an actual test form…a second evaluation is ordinarily conducted at the conclusion of the standard setting meeting. Commonly, both evaluations consist of a series of survey questions" (Cizek, 2004, p. 44). The various kinds of information collected in these evaluations focus on the procedural evidence (Was training explicit and practicable? Were the procedures implemented? What kind of feedback was

provided? What kind of documentation was collected?). Internal evaluations would focus on consistency within the method for both scores and judgments. External evaluations would focus on comparisons of standard setting methods with each other and with other sources of information.

Conceptualizing Achievement Standards for Alternate Assessments[3]

Because of the relatively recent application of standard setting for alternate assessments, it is very difficult now to fully understand or interpret the outcomes. Are standards for alternate assessments too high or too low? What internal or external criteria should be used to evaluate them? How are the various methods to be compared? What procedural evidence is to be collected in establishing their credibility?

Achievement standards make explicit how much students should know and be able to do. "How much" refers to the content area knowledge and skills specified in the grade level content standards that an assessment targets. Achievement standards represent a state education system's policy statement about how much is good enough and about the aspirations for academic performance that the system and public holds for its students.

NCLB requires that performance on general education content area assessments be reported in multiple levels and that one level be designated as *proficient*. Proficient performance is intended to be a rigorous but reasonable level of performance that all students are expected to achieve. The requirements for multiple achievement levels and a proficient level for all students applies to grade level achievement standards, modified achievement standards, and alternate achievement standards. These requirements apply to all statewide content area assessments, whether they are the general assessment or alternate assessments.

[3] Adapted from Swaffield, Siskind, & Ferrara (2005).

The challenge for alternate assessments is to conceptualize a definition for proficient (and other achievement levels[4]) that is both rigorous and appropriate for the diversity of academic levels of students with disabilities. Rigor is required by NCLB and is necessary to ensure that students are instructed on challenging academic content. Appropriateness is necessary to ensure valid inferences about what students with disabilities know and can do. Conceptions of proficient performance for students with severe cognitive disabilities are evolving. We discuss four conceptions.

Grade-level or grade-band achievement standards. It is common practice to establish separate proficient standards at each individual grade level for general content area assessments. Separate grade-level proficient standards make sense for these assessments because academic curriculum and student achievement change with each grade level. The same cannot always be said for students with severe cognitive disabilities. Academic-cognitive functioning distinguishes students with severe cognitive disabilities; age and nominal grade level may provide little reliable information about what a student may know or be able to do in relation to grade level academic content standards, especially in grades 3 and beyond. Setting different proficient standards for each grade level for alternate assessments may not enable appropriate inferences about what students know and can do.

A single proficient standard for all students in all grades. A single standard for proficient performance has been established for some current alternate assessments. The Ohio portfolio assessment for students with severe cognitive disabilities is one example. A single proficient standard on portfolio assessments enables valid inferences about student achievement because students in all grades participate in the same assessment process. A single proficient standard for

[4] Throughout this chapter we refer to Proficient level alone for simplicity. The discussions apply to multiple achievement levels as well.

other assessment approaches, such as performance events like South Carolina's HSAP-Alt (High School Assessment Program alternate assessment), also may support valid inferences about student achievement. Alternate assessments based on collections of performance tasks with a summary scale require other conceptualizations for a proficient standard, since performance of all students is reported on a single score scale that represents a wide range of performance and achievement.

Multiple proficient standards, one for each level of cognitive-academic functioning.[5] A proficient standard for students with severe cognitive disabilities who read and write would not likely be attainable for other students who may have no literacy skills. In contrast, an appropriate attainable standard for proficient performance for these students would not likely be adequately challenging for students with some literacy skills. Assessments containing tasks that target these diverse levels of academic-cognitive functioning and academic achievement require multiple definitions of proficient performance. For example, the Pennsylvania Alternate System of Assessment (PASA) provides three levels of assessment to account for the wide range of achievement of students with severe cognitive disabilities. Each level is accompanied by separate sets of achievement levels that correspond to the differences in academic challenge of each level of the assessment. The new South Carolina alternate assessment (SC-Alt), which will be implemented in 2007, is a collection of performance tasks with a single summary scale. The tasks and summary scale provide access to the lowest and highest functioning students with severe cognitive disabilities. South Carolina is considering establishing three levels of proficient standard that correspond to high, moderate, and low levels on the summary scale.

Growth standards. An alternative to static standards is to establish standards for year-to-year growth in achievement. For example, proficient performance could be defined for any

[5] Or score scale level, region, or range.

alternate assessment in terms of the number of points of increase from one school year to the next. An advantage of this conceptualization for proficient is that one or more standards for proficient can be established for students at different levels of cognitive-academic functioning or for different levels of difficulty and complexity of an assessment. One drawback to growth standards may lie in their conceptual complexity. It may be difficult for some to accept the notion of amount of growth in achievement as defining proficiency in place of attaining a specified level of performance. In addition, it seems likely that growth standards would have to vary across school years; it seems unreasonable to expect that students with severe cognitive disabilities – or general education students – could maintain similar amounts of growth in achievement every school year.

Irrespective of proficiency and depending upon the standard setting, the result of any score report needs to be effective communication. Score reports need to provide all the essential information for adequately informing the public: legislators, parents, and educators. In these reports, information needs to be effectively organized in a hierarchical and orderly fashion so that relationships among the procedures and the outcomes can help establish valid inferences. In this next section, we consider the importance of graphics in displaying results.

Elements, Examples, and Explanations in Score Reporting

Score reports need to effectively combine procedural explanations with graphic displays of outcomes, and in the process leave the consumer with an understanding of individuals and systems. In this next section, we describe the critical role that graphs have in displaying complex information effectively; we then use NAEP as an example of some components of reports (particularly clear definitions of achievement level descriptors and the display of outcomes); finally, we consider the unique issues when students with disabilities are part of the system.

Elements of Effective Reports

Score reporting is likely to combine text and graphics. The text needs to be logically and sequentially organized so that various critical stakeholders can find information easily. Probably the most important information is outcomes, whether they are of a single student on a series of tasks (or reflect various levels of proficiency on achievement standards) or a system report being submitted for peer review. In developing the graphs that support the interpretations and assist in making inferences, several rules have been developed. Following are 12 rules suggested by Wainer (1997).

1. Show as much data as possible. The whole point behind the use of graphs is to depict information in a clear way that also is multivariate. Using a data density index from Tufte (1983), defined as the number of numbers plotted per square inch, graphic displays should be as high as 50 and more and can still retain clarity and communication.

2. Maximize the data/ink ratio. The purpose of a graphic display is to present contrast, and therefore, it is important for elements not to blur into each other, the background, or any descriptors (scales, legends, etc.)

3. Ensure the visual metaphor is explicitly used. Data should be displayed consistent with the values they represent (e.g., nominal data are not to be depicted in a 'continuous' display).

4. Order of values need to be reflected appropriately in the values of the axes (to convey change on some dimension).

5. Graph data in context. Often, important contextual information needs to be concurrently displayed to appropriately convey base rates or to better allow appropriate interpretations.

6. Maintain consistent scales on the same graph. Broken values that are not contiguous may present information that is falsely interpreted.

7. Emphasize important issues in the display of values, graphic elements, and associated text.

8. Appropriately convey the baseline of base rates. Change needs to begin at some logical and honest starting point.

9. Order the data using important principles, not just alphabetically, to highlight appropriate outcomes (ranking of variables)

10. Include appropriate interpretive assistance in the form of labels, keys, markers, etc. All graphs need more than just visual elements, but all information displayed needs to be consistent, clear, and complete.

11. In clarifying the data, precision in values needs to be appropriate to the scale and the decision.

12. Use previous formats that have been successfully deployed. Many very important and well-formatted graphic displays still are useful as a template.

Probably the best resource for graphic integrity comes from the work of Tufte (1983, 1990, 1997). In addition, Cleveland (1985) has written appropriate guidelines for creating effective graphs that communicate reporting of outcomes. In the end, procedural evidence that is explanatory and empirical evidence that is statistical need to be combined so that the major claim being made is supported. In the next section, we describe come of the elements of the *National Assessment of Educational Progress* because we can highlight the meaning of proficiency in what it means about skills, standards, and outcomes.

Example: National Assessment of Educational Progress (NAEP)

In this assessment system, several examples are readily available from the web site[6]. Generally, the information displayed is exemplary; however, like all large-scale assessments, as more and more demands are being made of them (in the complexities and nuances of the items and content areas, as well as in the decisions being made with the outcomes), this assessment also needs to be carefully managed in the reporting of outcomes.

Achievement levels. There are three achievement *levels* (not standards) for each grade assessed by NAEP (4, 8, and 12): *Basic*, *Proficient*, and *Advanced*. The following definitions apply to all subjects and all grades assessed by NAEP: (a) *basic,* which reflects partial mastery of prerequisite knowledge and skills that are fundamental for proficient work at each grade; (b) *proficient,* which reflects solid academic performance for each grade assessed with students reaching this level by demonstrating competency over challenging subject matter, including subject-matter knowledge, application of such knowledge to real-world situations, and analytical skills appropriate to the subject matter; and (c) *advanced,* which reflects *s*uperior performance.

Application to grade four proficiency standards. Descriptors for these three levels are specific to each of the grade levels. Following are descriptors for fourth grade with the actual score values presented in parentheses (using a total scale range from 0 to 500).

1. Basic – (208). Students performing at the *Basic* level should demonstrate an understanding of the overall meaning of what they read. When reading text appropriate for fourth graders, they should be able to make relatively obvious connections between the text and their own experiences and extend the ideas in the text by making simple inferences.

[6] See http://nces.ed.gov/nationsreportcard/reading/achieve.asp (July 30, 2005)

2. Proficient – (238). Students performing at the *Proficient* level should be able to demonstrate an overall understanding of the text, providing inferential as well as literal information. When reading text appropriate to fourth grade, they should be able to extend the ideas in the text by making inferences, drawing conclusions, and making connections to their own experiences. The connection between the text and what the student infers should be clear.

3. Advanced – (268). Students performing at the *Advanced* level should be able to generalize about topics in the reading selection and demonstrate an awareness of how authors compose and use literary devices. When reading text appropriate to fourth grade, they should be able to judge text critically and, in general, to give thorough answers that indicate careful thought.

Reporting outcomes. These scores are displayed in two primary ways: with the actual achievement levels attained and in terms of the percentage of students attaining each categorical level. In the following two graphs, achievement levels are displayed on the left and proficiency levels on the right (for grades 4 and 8) for students taking the reading test from 1992 through 2003[7].

[7] These graphs were retrieved from the following website: from http://nces.ed.gov/nationsreportcard/states/achievement.asp (July 30, 2005).

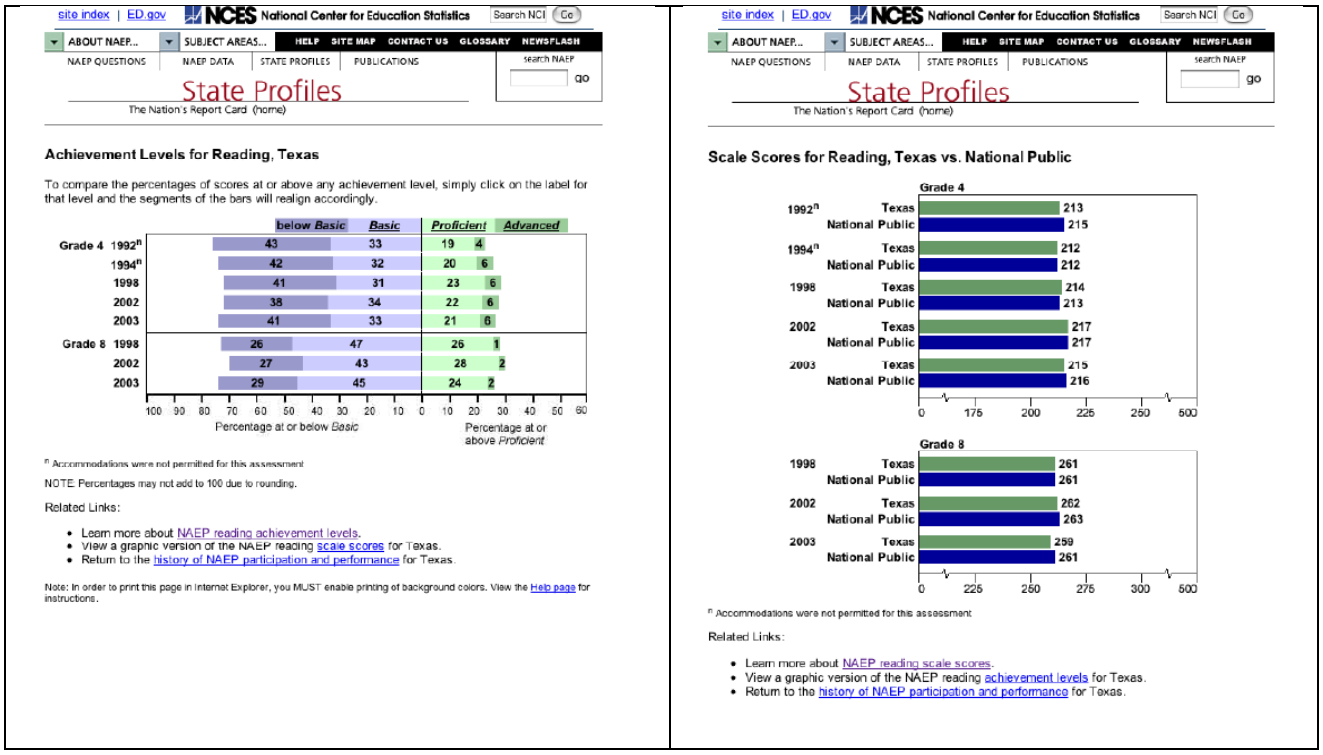

Figure 10.1 NAEP Graphic Displays

While the NAEP reporting system may serve as an example for states to consider in score reports, the inclusion of students with disabilities in NAEP is still problematic. For the most part, the focus has only been on inclusion, with little attention to effects of inclusion, appropriate use of accommodations, or any consideration of applying more universal designs so that ALL students can participate (See National Research Council, 2004). Therefore, in this last section, we turn our attention to issues that arise when reports are generated that either are for students with disabilities as they participate in the large-scale assessment system or are for critical stakeholders and reflect systems-level data.

Explanations: Issues and Inferences

Score reports for students with disabilities should provide the same information in score reports as are provided for students who do not have disabilities. The ways in which students with disabilities participate in large-scale assessments do not alter this requirement. Score reports

for students with disabilities also should contain *additional* information to specify the assessment in which the student has participated (e.g., whether it is a general education assessment or an alternate assessment) and the conditions under which responses were elicited. At the same time, flagging of scores is not an appropriate procedure for students with disabilities. Basically, the interpretation at any level of proficiency needs to fairly report outcomes without drawing attention to questions about the validity of inference. Nevertheless, two important qualifications apply.

- In terms of the logic of drawing inferences about what students know and can do, aggregating scores from general education assessments, alternate assessments with modified achievement standards, and alternate assessments with alternate achievement standards is not defensible. This is true because the inferences about what students know and can do are different.
- In terms of reform policy, it makes good sense to aggregate judgments of performance based on these assessments/standards because of the positive impacts we expect that to have on curriculum, instruction, and expectations for students with disabilities.

Performance of students on general content area assessments and alternate assessments are aggregated together, as required for NCLB reporting. Aggregating performance from such different assessments is supportable if definitions of proficient performance are highly similar. Definitions of proficient for general content area assessments and alternate assessments must support claims of similar levels of performance and achievement. For students who participate in an alternate assessment, this claim is in relation to the content standards targeted by the alternate assessment. The definitions of proficient need not yield **similar** inferences about student performance and achievement. In fact, they cannot support such inferences. But they should

enable **relatively similar** inferences. Relatively similar inferences require that the general content area assessment and alternate assessment both target the same grade-level content standards. The relative, rather than absolute, similarity of inferences arises from adaptations made to the breadth, depth, and complexity of content standards that are targeted by alternate assessments

Score reports should pass the "kitchen-table test." That is, score reports should be adequately clear and complete so that parents of students with disabilities are readily able to understand the assessment in which the student has participated, the content standards on which the student has been assessed, and the achievement standards which are used to report student performance and achievement. The research on score reporting is quite limited. Two comprehensive studies (Goodman & Hambleton, 2004; Ryan, 2003) summarize the general contents and features of score reports for content area achievement tests. They also make clear that little research has been conducted on the comprehensibility and usefulness of score reports for their intended audiences. That research suggests that most score reports often do not communicate effectively and that consumers often misinterpret the information in score reports.

Rather than provide examples of different reports using each of the different approaches to assessment, this section of the chapter highlights critical issues in score reporting with each assessment approach. The critical issue is that the assessment provides an opportunity and a limitation in reporting performance.

General education reports, with and without accommodations. These reports benefit from the use of selection type responses that allow subtest scores to be reported, each with their standard error of measurement with graphic depictions of confidence intervals. Although this allows the state educational agency (SEA) to include a considerable amount of information, it

often is overwhelming. This especially becomes a problem when various levels of aggregation are considered (teacher class rosters, building grade level, and district grade level). Furthermore, the reports often are distributed as fixed files (hard copy or electronically) making the information difficult to sort (e.g. from low to high).

Assessments using rating scales and observational checklists. With the use of rating scales, the various dimensions can be displayed but the process for conducting the observation or the environmental conditions are not easily communicated. Reports are unlikely to separate out judgments as a function of rater set (harshness or leniency) versus performance values. As in the standard setting process, the training and calibration of the raters needs to be described explicitly as part of any student reports.

Portfolios. The critical issues in reporting achievement with portfolios include the following: (a) often the more critical variance is in the text and it is difficult to report on the task without including it; as a consequence, the report either under-represents the assessment by not including the tasks or work samples or it becomes difficult to manage because too much is included; (b) another issue with portfolios is the comparability of evidences within the collection and the degree to which judgments are based on a holistic or analytic evaluation.

Performance events. Reports using performance events are likely to be most like the selection responses of multiple-choice tests; many brief tasks can be assembled to arrive at sub-scores, and considerable information can be reported. In addition, however, the scoring system (if using an objective count) or a rubric (if using a subjective judgment) needs to be part of the report to provide full meaning of individual scores.

Collections of performance tasks with summary score scales. This type of report may benefit from some of the performance event assessment technologies but suffer from the lack of

coherence in concatenating the tasks together into sets that have a summary scale. Considerable information may be available, but the report needs to provide a manner for holding it all together: in how the collections relate to each other as well as the use of the rubric for scoring performance. Furthermore, it is important to distinguish the model being used to report scores: (a) a conjunctive model (performance on all tasks have to be at a certain level of proficiency in an additive fashion), (b) a disjunctive model (all but some performances have to be at a certain level of proficiency), or (c) a compensatory model (performance on various tasks can average out with some high level performance being used to compensate for low level performance in arriving at a proficiency level).

Inferences

Because inferences that can be drawn from score reports may be constrained when assessments are given with accommodations or when alternate assessments change the breadth, depth, or complexity of the content standards being assessed, score reports for students with disabilities should also stipulate that:

- If they participate in the general education assessment without accommodations, it is generally considered a comparable test outcome to all students taking the grade level test. An important caveat is that this interpretation is only about their exhibited performance, not their optimal performance.
- If they participate in a general education assessment with accommodations that may reduce an impediment related to the disability, the report has not flagged the score but reported its value in the same manner as if no accommodation had been used.
- If the same grade level standards apply (and only changes have been made in the types of supports), then that change has been noted in the report. Individual states need to consider

this decision carefully, however, as the same effect may accrue as with flagging accommodations.

- For some students, the grade level content standards have been specifically manipulated in breadth, depth, and complexity. For these students, the inferences about grade level achievement need to be constrained. The problem in reporting is likely to be in the manner that this constraint is communicated. Because the changes are significant enough, it is unlikely that the actual score value can be aggregated with the students who have taken the general education test (with or without accommodations). Nevertheless, because of the standards setting process (see below) judgments of 'proficiency' have been made and this judgment is comparable.

Summary and Recommendations

We began the chapter by considering the type of score reporting as a function of the assessment approach.

1. Both the grade level content standards and the achievement standards need to be presented in a report so that the value of the outcomes can be interpreted relative to the content (breadth, depth, and complexity) of the standards.

2. Score reporting needs to be accompanied by a reference that is fully explained. If a norm-reference is used, then the interpretations need to be explicitly described (using some metric to depict relative standing); if a criterion-reference is used (which is most typical given most assessments are standards-based), the report should include some kind of item or strand map of performance with both value (achievement on the scale) and proficiency (presence in a category as a function of percentage in that category).

3. Achievement level descriptors need to be accompanied with some description of the standard setting dimensions, including the method used and examples of borderline cases (individuals or groups).

4. Explicit reference to some achievement standard needs to be present in the report. Presently, several options are available in which achievement is anchored to a grade level standard, a single standard across all grades versus multiple standards across the grades or across the various participation options. Or a growth standard in which progress (improvement) is valued rather than performance.

5. Irrespective of method for setting standards or model for depicting achievement, the report should include an easy to interpret graphic display that is designed for intuitive interpretation and maximum readability. Graphs need to be self-contained in their ability to convey all the critical information and not require extensive reference in text.

6. Using the National Assessment of Educational Progress as a model for either reporting (or highlighting issues), reports for students with disabilities needs to address several issues to provide appropriate inferences.

7. The assessment approach needs to be considered in addressing unique issues and to provide the best evidence supporting both the score value as well as the judgment of proficiency. As we discussed in Chapter 2, evidence that is convergent needs to be presented along with information that counters alternative explanations.

References

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (pp. 508-600). Washington, DC: American Council on Education.

Cizak, G., Bunch, M. B., & Koons, H. (2004). Setting performance standards: Contemporary methods. *Educational Measurement*, *23*(4), 31-51.

Cizek, G. J. (2001). *Setting performance standards: Concepts, methods, and perspectives.* Mahwah, NJ: Lawrence Erlbaum.

Cleveland, W. S. (1985). *The elements of graphing*. Monterey, CA: Wadsworth Publishing.

Ferrara, S, & DeMauro, G. T. (2006 forthcoming). Standardized Assessment of Individual Achievement in Grades Kindergarten Through 12. In R. L. Brennan (Ed.), *Educational measurement* (4th ed.). Phoenix, AZ: Greenwood Publishing.

Goodman, D. P., & Hambleton, R. K. (2004). Students test score reports and interpretive guides: Review of current practices and suggestions for future research. *Applied Measurement in Education*, *17*(2), 145-220.

Kingston, N. M., Kahl, S. R., Sweeney, K., Bay, L. (2001). Setting performance standards using the body of work method. In C. J. Cizak (Ed). *Setting performance standards: Concepts, methods, and perspectives*. (pp. 219-248). Mahwah, NJ: Lawrence Erlbaum.

Lewis, D. M., Mitzel, H. C., & Green, D. R. (1996, June). Standard setting: a bookmark approach. In D. R. Green (Chair). *IRT-based standard setting procedures utilizing behavioral anchoring*. Symposium at the Council of Chief State School Officers National Conference on Large-Scale Assessments, Phoenix, AZ).

Mitzel, H. C., Lewis, D. M., Patz, R. J., & Green, D. R. (2001). The bookmark procedure: Psychological perspectives. In C. J. Cizak (Ed). *Setting performance standards: Concepts, methods, and perspectives*. (pp. 249-281). Mahwah, NJ: Lawrence Erlbaum.

National Research Council (2004). *Keeping score for all*. Washington, DC: Author.

Plake, B. S., & Hambleton, R. K. (2001). The analytic judgment method for setting standards on complex performance assessments. In C. J. Cizak (Ed). *Setting performance standards: Concepts, methods, and perspective*. (pp. 283-312). Mahwah, NJ: Lawrence Erlbaum.

Ryan, J. M. (2003 October). *An analysis of item mapping and test reporting strategies*. (Available from SERVE, P.O. Box, 5367, Greensboro, NC 27435; www.serve.org.)

Swaffield, S., Siskind, T., & Ferrara, S. (2005). Conceptions of performance standards for alternate assessments. In G. Tindal (Organizer), *Setting Alternate Achievement Standard – Challenges and Solutions*. Presentation at the National Conference on Large Scale Assessment, San Antonio.

Taylor, C. S., & Nolen, S. B. (2005). *Classroom assessment: Supporting teaching and learning in real classrooms*. Upper Saddle River, NJ: Pearson.

Tufte, E. (1983). *The visual display of quantitative information*. Cheshire, CT: Graphics Press.

Tufte, E. (1990). *Envisioning information*. Cheshire, CT: Graphics Press.

Tufte, E. R. (1997). *Visual images*. Cheshire, CT: Graphics Press.

Wainer, H. (1997). *Visual revelations*. Mahwah, NJ: Lawrence Erlbaum.

CHAPTER 11: DOCUMENTING THE VALIDITY OF USING TEST SCORES OF STUDENTS WITH DISABILITIES FOR ACCOUNTABILITY

In Chapter 1 of this volume, a model was presented for including students with disabilities in large-scale assessments. Chapter 2 described the state's responsibility for establishing the validity of the large-scale assessment system. An annual validity evaluation of the testing program was recommended, as were methods for collecting the evidence that makes the annual evaluation possible. Section II of this volume (Chapters 6, 7, 8, and 9) provided detailed technical information on how states might collect validity evidence in various categories.

This chapter gives advice on how to document this evidence to establish the validity of test score interpretations and uses. This chapter not only provides information for technical documentation of any testing program designed to assess the achievement of students with disabilities, but also describes the concepts, principles, and procedures useful in documenting any large-scale assessment program.

Brief Review of the Validation Process

In Chapter 2 of this volume and in other sources (e.g., AERA, et al., 1999), the process of validation is described as a means for establishing the validity of using a test score in a state accountability formula. The process begins with a definition of what we want students to learn and what our tests measure; in most states, a set of content standards have been developed and approved, and they form the basis for the content of the large-scale assessments. The second step is to create an argument about what the test score from the assessment means and why the test results can be validly interpreted as measuring this content. In the validation process, a claim is made that such interpretation and subsequent use of these scores are valid. The final step is to assemble and organize validity evidence in support of this argument and claim, and, at the same

time, consider evidence that may work against the claim. The reason for looking at both supporting and disconfirming evidence is to eliminate or reduce each threat to validity. When assessing the learning of students with disabilities, the threats to validity are considerable. Thus, validation is more challenging.

The process of validation is very much like a court proceeding involving a complaint against the plaintiff (the test sponsor). The complaint is that the test validity is inadequate and the test produces scores that are not useful. The plaintiff claims innocence of this accusation (i.e., claims that the test is not inadequate and that the test scores lead to meaningful inferences about student performance). An argument is made in favor of plaintiff and evidence is assembled and presented to support the argument and the claim of innocence. The "prosecuting attorney" presents negative validity evidence to expose the argument and weaken the claim. The prosecuting attorney shows the argument to be implausible and the claim of innocence false. The "defense attorney" tries to show that the argument is sound, the claim is true, and the evidence supports the use of test scores. The defense attorney also tries to show that the negative validity evidence is inconsequential. The jury (the evaluator of the testing program) evaluates the evidence presented for both claims and determines which is true.

Thus, validation is a rational process that informs the test sponsor, the public, and other interested parties about the reasonableness of using a set of test scores of students with disabilities for accountability purposes such as determining adequate yearly progress (AYP). We should view it as a two-sided argument: Is the claim for test use true or false? Of course, validity exists in degrees, but at some point, we must decide to trust the test data or to abandon them.

What is Technical Documentation?

Technical documentation involves the collecting, processing, and organizing of validity evidence that will exist in an archive but can be reported in different ways to different audiences. According to Becker and Pomplun (in press), "complete and thorough documentation of the test development process can be viewed as the foundation necessary for valid interpretation and use of test scores." That is why Chapter 6 of the *Standards for Educational and Psychological Testing* (AERA, et al., 1999) is devoted entirely to the idea of technical documentation. Sponsors of testing programs have the responsibility to report the validity of their test score use to various constituencies, including the Federal government, state government, local education agencies including the school districts and their boards, and the public, which includes the media, parents, and advocacy groups.

Table 11.1 below summarizes standards that are presented in Chapter 6 of the *Standards* (AERA, et al., 1999). Although these standards apply to all testing programs, especially those with high-stakes purposes, it is important to note that some of these standards have implications for testing students with disabilities. Standard 6.4 calls for the description of the population to be tested. If the set of test specifications is altered to adapt to students with significant cognitive disabilities, this too should be noted. Standard 6.6 is focused on the program of instruction. This standard recognizes the value of providing appropriate opportunities for students to learn and re-learn important content, as reflected in the content standards, especially when that content has been modified in terms of breadth, depth, or complexity. Standard 6.9 mentions validity studies. It has been widely recognized that research on testing students with disabilities is deficient, and validity studies are much needed. There are several different types of validity studies (Haladyna, in press) and later in this chapter, validity studies are discussed in relation to solving problems

that exist in the valid assessment of students with disabilities. Standard 6.11 appears to recognize the need for accommodations in test administration for students with disabilities.

Table 11.1. Standards from Chapter 6 of the Standards Concerning Technical Documentation

6.1. Test documents should be made available to those interested in this information.
6.2. Documents should be complete, accurate, and clearly written for their intended audiences.
6.3. The rationale for the test and the intended interpretations and uses should be clearly stated. Evidence should support these interpretations and uses. When misuse is possible, cautions should be specified.
6.4. Provide test specifications and a description of the population to be tested.
6.5. Technical data including descriptive statistics should be provided in a technical report.
6.6. When a test reflects a course or specific training, a curriculum, textbook or packaged instructions, the supporting documentation should identify and describe these materials.
6.9. Studies should be cited that provide validity evidence.
6.10. Materials should be presented that help interpret test results.
6.11. If administration methods vary, then evidence should be presented for the interchangeability of results.

What Are Other Reasons for Technical Documentation?

Besides being part of the process of validation, technical documentation serves other important purposes. Documentation is a primary tool for communicating with test users (AERA, et al., 1999, p. 67). Publications and archival information can show the public that the test sponsor has taken appropriate steps to assure that students with disabilities are being properly assessed and that accommodations and modifications in test content, test design, and test administration are leading to more valid interpretations of student learning.

Test scores that are used for accountability can have high-stakes implications and has warned that, as the stakes for test score use go up, so does the need for validation. If tests are going to be used for high-stakes decisions about students, teachers, schools, and states, then validity must be our prime objective. Technical documentation of the validity evidence assures the test sponsor that the public is being well served, and that those whose test scores are being used are less likely to suffer negative consequences.

Tests used for accountability can be legally challenged. Litigation is generally

unwelcome. Legal educational testing expert Susan Phillips (1996) observed that the best defense against legal challenge is clear documentation of test development, administration, and scoring procedures.

Validity Evidence

Procedural versus Statistical/Empirical Validity Evidence

As Chapter 2 suggests, all validity evidence can be classified as coming from the use of appropriate test development procedures (summarized in documents), or as coming from statistical analyses or research (summarized in charts, tables, or reports). The convenient dichotomy is procedural or statistical/empirical validity evidence. In some circumstances, it is necessary to integrate the two forms of evidence. In many instances, the two forms of evidence are naturally complementary. For example, Downing and Haladyna (1996) showed that for item development and the assessment of item quality, procedural and statistical/empirical evidence are integrated to effectively argue for the validity of item responses. This point is made by Haladyna (2004) in the context of multiple-choice item development but applies equally to constructed-response testing.

Threats to Validity: Negative Validity Evidence

Messick (1989) recognized that validity evidence might work against the argument and claim for validity. In Chapter 2, it was recommended that threats to validity be investigated with the idea that finding these threats and dispelling or reducing their severity can only increase the strength of the argument and claim for validity.

Construct under-representation refers to situations where the test content may not be exactly what was intended. This threat is likely to be greater in assessing the achievement of students with disabilities, especially when test content is modified. An argument must be made

as to the appropriateness of the modified content for the inference we want to make about student achievement. Content-related validity evidence seems to be the focal point for this threat to validity, but item quality also is a factor.

A more pervasive threat to validity is CIV, which is systematic error. As noted in Chapter 2, there are many types of systematic error. Haladyna and Downing (2004) discussed these sources of systematic error and provide documentation for how these sources of error can pollute test score interpretations and weaken the use of test scores. With the population of students with disabilities, the threat to validity of systematic error is even greater because many extraneous factors related to disabilities are introduced into the assessment process and into the actions taken to make the assessment more valid.

Content-Related Validity Evidence

Table 11.2 identifies testing standards (taken from the *Standards,* 1999) that relate to the content of the test, whether the validity evidence is procedural (P) or statistical/empirical (S/E), and gives a brief précis of the standard. Not all content-related standards are presented here. Instead, only the ones that bear on the unique problems associated with assessing students with disabilities appear in the table[1]. Most of these standards are fundamental to all testing programs but are worth emphasizing in the context of assessing the achievement of students with disabilities.

[1] Readers should consult each actual standard and other standards not listed for a fuller understanding of each standard and how this kind of evidence supports validity.

Table 11.2. Selected Standards for Content-Related Validity Evidence

STD	P	S/E	Brief Description
1.1	√		Develop the interpretive argument for validity.
1.2	√		Define what the test measures.
1.3	√		Identify the population.
1.5		√	Describe the sample taking the test.
1.6	√	√	Describe the process by which the content is identified for testing including any empirical studies of subject matter agreement.
1.7	√	√	Present information about the qualifications and accuracy of subject matter experts who help develop the test.
1.8	√		Document the process used to identify content and how well the subject-matter experts performed their duties.
1.19	√		Where assessment leads to alternative content and tests, the rationale for this should be clear and the data convincing about its validity.
1.24	√	√	Report unintended or negative consequences.
4.1	√		Report the meaning of test scores and intended interpretations in the technical report.
4.16	√		Where test specifications must be changed for an alternate assessment, a new scale should be produced or users of these scores should now this fact.
7.1	√		If research reports differential test score interpretations because of a student's disability, this fact should be noted in future interpretations and uses.
7.2 7.3	√	√	When research shows CIV due to subgroup membership and one of these groups has a disability, the validity of any inference should be questioned and verified. The same principle applies to item level data.
10.1	√	√	The disability should in no way interfere with the accurate assessment of the content being assessed.

Reliability

A primary type of validity evidence is reliability[2]. Table 11.3 identifies relevant standards and other information consistent with the presentation in Table 11.2 for content-related validity evidence. Standards 2.10, 2.11, 2.12, 2.14, 2.18, and 2.19 all seem relevant to the unique issues faced in testing students with disabilities. Technical documentation of reliability should contain both a rationale and descriptions of concepts, principles, and procedures of reliability and the SEM, as well as present data bearing on each issue.

[2] Chapter 2 of the *Standards* is expressly devoted to this important topic.

Table 11. 3. Selected Standards for Reliability-related Validity Evidence

STD	P	S/E	Brief Description
2.1 2.2		√	Reliability estimates and standard errors of measurement (SEM) should be reported.
2.3		√	Reliability estimates and SEM should be reported for scores that show AYP.
2.4	√		The methods should be described and justified, and the sampling should be described.
2.10	√	√	When performance testing is used and performance is rated, the consistency and other performance factors of raters should be documented.
2.11	√	√	When we test different populations, reliability estimates and SEM should be reported for each population, if materially different.
2.12		√	Differences among grade levels and age groups should be recognized with appropriate reliability estimates and SEMs.
2.14		√	Conditional SEMs should be reported for multiple cut scores that are used in AYP.
2.18		√	Changes in test administration (i.e., accommodated administrations) justify separate reliability estimates.
2.19		√	The SEM for a group mean should always be reported, especially for AYP.

Item Quality

The most basic building block of any test is the test item. Consequently, validity evidence bearing on item quality should also be presented. The *Standards* (AERA, et al, 1999) are surprisingly sparse in the number of standards listed and the extensiveness of discussion on item quality. Table 11.4 below provides standards bearing on validity evidence of item quality. Although only four standards are listed, the first one in the list, standard 3.7, is inclusive of many extended and detailed activities in item development that all apply to universal item design and the validation of items for this population of students.

Table 11.4. Selected Standards for Evidence Related to Item Quality

STD	P	S/E	Brief Description
3.7	√	√	The procedures to develop, review, and tryout items should be documented.
3.8	√	√	Tryouts should be particularly well documented for students with disabilities and the populations they represent.
3.9		√	Item analysis should be conducted.
6.4	√		Activities used to develop items should be documented.
7.3		√	Differential item function should be part of the validation process for item responses.

Given the limits of these standards, test sponsors would do well to consider other sources of advice and examples of validity evidence on item development (Abedi, in press; Baranowski, in press; Downing, in press, Downing & Haladyna, 1996; Haladyna, 2004; Welch, in press; Zieky, in press). Some of the concerns in item development that should be discussed in the technical documentation are as follows and as appropriate (no report need address all issues):

1. What do the test specifications say about item formats?

2. Is there an item-writing guide?

3. What are the qualifications of the subject-matter experts (SMEs) who write the items?

4. Was universal item design used?

5. Did SMEs complete item-writing training?

6. How were items developed?

7. How were items reviewed?

8. Were items field-tested using think-aloud procedures?

9. Were items field-tested by administering items to students from the targeted groups?

10. What were the results of item analysis?

11. Were studies of differential item functioning conducted? What were the results?

12. Were studies done of the structure of the content using item responses? Was the analysis appropriate?

One area that is not generally well documented in technical reports is the connection between content to test items. The process used to link content to test items and to classify items by content categories should be carefully documented in a report that is either part of the technical report or offered as a separate citation.

Test Design, Including Modification of Content

Tests can be designed in many ways. For students with disabilities, test administration may have to be adapted, and test content may have to be modified to reduce the breadth, depth, or cognitive complexity. Item formats may have to be selected that are more suitable for students with disabilities. Table 11.5 summarizes standards for validity evidence related to test design.

Table 11.5. Selected Standards for Validity Evidence Related to Test Design

STD	P	S/E	Brief Description
3.15	√		This standard calls for great care and documentation in the design of tests with attendant conditions and materials, such as accommodations and assistive devices.
7.3	√		Where DIF exists, action should be taken to remove this threat to validity in test design.
10.1	√		When designing the test, make certain that a disability does not interfere with obtaining an accurate indication of achievement.
10.2	√	√	Research-guided principles and psychometric advice is necessary for making modifications in content and in the test for an appropriate assessment.
10.3	√		Pilot testing is recommended before a test becomes operational.
10.4 10.5	√	√	Any modification should be well supported and documented.
10.7	√		Modifications of content and tests should be made for each significant disabling condition, if sample size permits this kind of reporting.

Test Administration, Including Accommodation

The *Standards* are very specific about six strategies for modifying test administration to accommodate students with disabilities (AERA, et al., pp. 103-104, 1999). States and school districts should very intentionally study these strategies and determine which one(s) will be implemented and how. The underlying validity principle here is that all students should have the opportunity to perform to their maximum under ideal conditions; if any test administration conditions interfere with students' performance, they create a substantial threat to test score validity. Table 11.6 summarizes the relevant standards.

Strategy 1. Modifying presentation format. For student with vision or hearing impairments, the test may be administered appropriately for the disability of the student.

Strategy 2. Modifying the response format. For students with various physical impairments, the test administrator might allow the student to point to answers or to verbally respond.

Strategy 3. Modifying timing. Some students may require more time or more breaks between parts of the test. Not having enough time to complete a test that is not a speed-oriented test creates a serious threat to validity.

Strategy 4. Modifying test setting. Some students perform better in a setting away from other students. Some students might need better lighting or a location more conducive to maximum performance.

Strategy 5. Using only portions of a test. When tests utilize mixed formats, only certain portions of the test may pose problems with a student's performance. Administering the whole test under these conditions might lead to an incorrect inference about the student's achievement. For instance, a test that requires listening might pose a problem to a hearing-impaired student or a student with attention deficit problems.

Strategy 6. Using alternate assessments. Sometimes, the nature, degree, and extensiveness of the modifications required lead to the development of an alternate assessment instead. Alternate assessment is one of the most challenging aspects of test development for students with disabilities and is treated in more detail elsewhere in this volume.

Table 11.6. Selected Standards for Evidence Related to Test Administration

STD	P	S/E	Brief Description
5.1	√		Follow test administration procedures unless an accommodation is justified.
5.3	√		When accommodations have been approved, the test taker should be notified of this fact.
7.12 10.10	√		All students should be administered a test so that treatment is equitable and administration conditions are comparable in terms of being standardized.
10.6	√		If time limits are altered, fatigue and other factors should be studied and support decisions about time limits.
11.23	√		A mandated test should provide and document recommended accommodations.

As noted above, we have many standards that bear on test administration and affect students with disabilities. Many of these standards relate to documentation of procedures for changes in administration and assurance that the results are fair to the student with respect to their achievement being assessed. Some standards may seem redundant, because they closely relate to other standards cited previously. This happens because the task of establishing validity evidence for participation of students with disabilities in large-scale assessments cannot be isolated to singular standards and testing procedures.

Test Scoring

Although test scoring is mundane, several standards bear on validity evidence for students with disabilities. Students with disabilities are more likely to be taking tests that are scored using subjective judgment. With subjective scoring threats to validity arise from rater effects such as severity/leniency, central tendency, restriction of range, halo error, among others. Table 11.7 contains several standards that call attention to issues in scoring tests. Although these standards apply to all students and all testing programs, extra vigilance is needed with subjective scoring.

Table 11.7. Selected Standards for Evidence Related to Test Scoring

STD	P	S/E	Brief Description
3.14	√	√	Scoring of tests with performance formats should be done with additional scrutiny for factors that may distort results. Results should be documented.
5.1 5.2	√		Care should be taken in scoring, particularly when a student with a disability is involved.
5.8	√	√	Stresses score accuracy and documentation of quality control.
5.9	√	√	As scoring of these students is more likely to involve performance formats and subjective judgment, threats to validity are more common and should be carefully monitored.

Scaling for Comparability

Test score scales are used as a basis for interpreting scores, particularly for accountability purposes where gain, difference, or growth of groups are being observed. If possible, we try to place all students on the same scale for comparability. However, when testing students with disabilities, the problem of developing a comparable test score scale is very complex. Scoring of alternate assessment that relies on subjective judgments of the status of each student seems more suitable to report than creative scaling efforts that do not link to the scales we use for other students. Nevertheless, to comply with the accountability requirements of NCLB, test results for students with disabilities must be combined with comparable data for students without disabilities. Data must be interpretable as students with and without disabilities move from one achievement standard category to an adjacent category. Table 11.8 provides the two standards that apply to accommodated or modified tests. The major issue in these standards is whether flagging of the scores derived from accommodated or alternate assessments of students with disabilities is justified.

Table 11.8. Selected Standards for Evidence of Comparability of Scores for Accommodated or Modified Tests

STD	P	S/E	Brief Description
9.5 10.11	√	√	When there is evidence of score comparability between accommodated and modified assessments, no flag should be reported with a student score.

Standard Setting

Cut scores are set by qualified subject matter experts who study the purpose of the test and review actual test items. They use one of many recommended procedures for standard setting (see Chapter 10) and establish cut scores to identify places on the score scale that distinguish test performance to place students in categories for AYP. The comparability or fairness of such cut scores in alternate assessments is subject to validation and documentation of

these procedures is part of that validation. Only the one set of four standards are listed in Table 11.9 bears on setting cut scores for achievement categories.

Table 11.9. Selected Standards for Validity Evidence Related to Standard Setting

STD	P	S/E	Brief Description
4.9 4.16 4.20 4.21	√	√	The rationale for setting multiple cut scores should be clearly presented in technical documentation. This cluster of standards focuses our attention on how cut scores are developed and interpreted for all students and for students with disabilities.

Reporting

Score reporting has become a science within the field of testing (see Ryan, in press). The language of reporting must be interpretable to laypersons, so technical jargon must be minimized or explained. In dealing with students with disabilities, even greater sensitivity is needed. When reporting scores for a disability group, we should ensure that the inference we make about the group is accurate and that a disability did not interfere with the assessment nor unduly affect (raise or lower) the group's scores. Stigmatizing students by identifying them as members of a group should be avoided if possible. Flagging a student score may create doubt about the validity of our interpretation of that score. As in other instances, we should provide written justification for any decision we make.

Table 11.10. Validity Evidence Regarding the Reporting of Test Scores to Various Constituencies

STD	P	S/E	Brief Description
5.10	√		The language of score reports should be appropriate for intended audiences.
7.8	√	√	When scores are disaggregated by disability, are the scores interpretable?
8.8	√		Score reports should not be stigmatizing to students as function of their disability.
9.5 10.11	√	√	When there is evidence of score comparability between accommodated and modified assessments, no flag should be reported with a student score.

Summary

This section has provided a comprehensive summary of standards that should guide the documentation of validity evidence for assessment of students with disabilities. Although categories of validity evidence are extensive, the standards were selected because they are uniquely appropriate for the target population: students with disabilities.

This chapter only provides a skeletal framework for thinking about technical documentation. State and school district testing programs will typically not have sufficient resources to exhaustively identify, collect and archive all validity evidence, but the goal of any testing program sponsor is to accumulate this validity evidence over the lifetime of the testing program, and, in an evolutionary way, improve assessment of students with disabilities.

The Technical Report

A main source of information about technical documentation is the technical report (manual). This document is usually prepared by someone who is intimately familiar with the testing program and has access to all documents relating to validity. The technical report should be written for a nontechnical reader and handle technical topics in a way that educates the reader about validity. The technical report should ideally address all categories of validity evidence and relevant standards from the *Standards* should be cited and linked to each piece of evidence.

How Should the Technical Report Be Organized?

The technical report is not just a repository of validity evidence but should give the reader a comprehensive account of where this evidence is and how it supports or undermines the validity argument and claim. The document should not only reveal the argument and claim but also show how the evidence leads to a judgment about validity. The technical report should inform policy makers about any weaknesses in the testing program and how to repair these

weaknesses. This is done by offering short-term and long-term recommendations for improving the testing program and describing a research agenda that identifies problems that threaten validity and seeks remedies through research that informs future policy.

Table 11.11 contains an outline of an ideal technical report although most technical reports are not this detailed. The model technical report outline contains many elements that are repeated in each annual report or are modified very slightly year to year. The annual technical reports should trace the evolution of the testing program, showing how threats to validity and weaknesses in the program were handled in the past and what happened the next year as a result.

Table 11.11 Outline for an Ideal Technical Report

Topic	**Comments**
Introduction/Overview	Tells how the report is organized.
Purpose	Tells what is being measured and the purpose of the testing program, including desired interpretations and use of the test score (for AYP).
Historical Perspective	Brief discussion of events leading up to this year's report.
Description of Current Program	Gives information about the testing program, students served, test content, and other facts.
Validity, validation, and validity evidence	Provides overview of what validity is, what validation is, and how evidence is used for validation—for the consumer of this information
Professional test standards	Mentions the *Standards* and how it guides the establishment of validity evident.
Legal issues	If high stakes use is intended, mentions threats from litigation and what is being done to ward off these threats.
Validity evidence	The next sections contain categories of validity evidence as noted. In each section, appropriate standards are cited and evidence is linked to each.
Content-related evidence	As noted in chapter 2 of the Standards and elsewhere in the monograph.
Item development and validation	Includes both procedural and empirical evidence.
Test design and development	How were test designed or modified?
Test administration	Is there a test administrator's manual and how were tests administered? Were accommodations made? Appropriate?
Reliability	Reliability estimates and SEMs for the conditions needed.
Standard setting	Were the cut scores set appropriately? How?
Scaling/comparability	Is there a scale that is used for making inferences about student growth?
Rights and responsibilities	Test preparation and other factors bearing on students' right when taking the test and when scores are reported.
Threats to validity	An honest look that factors that might distort test scores and suggestions about what to do about it.
Validity studies	The research agenda and recent progress
Security	What has been done to ensure security
Recommendations	Suggestions for short-term and long-term efforts to improve the program keeping a focus on validity and threats to validity.

Other Ways to Disseminate Reports of Validity

Thus far, the technical report has been mentioned as a main way to disseminate technical documentation. However, there are other ways to provide technical documentation that have proven effective.

WEB Pages

The meteoric growth of the worldwide web and the development of web pages as a means for dissemination of information have made the public reporting of validity evidence through technical documentation much easier. Virtually all states have education department web pages that provide a comprehensive display of its activities, including its state testing program. Unfortunately, few deal directly with the population of interest in this volume, students with disabilities. Technical documents are difficult to find posted on these web pages, but states and school districts should consider making such information available to the public for all the good reasons supporting technical documentation.

Press Releases

States have a responsibility to communicate validity evidence to the public, to policy makers, and to other interested parties. Press releases of reports of validity convey to these constituencies that validity is important and that the testing program is dedicated to developing test scores for students with disabilities that can be validly interpreted.

Reports to the State Board of Education and the Legislature

Departments of Education and testing contractors routinely present findings to state boards and legislative assemblies on progress, results, and problems in statewide assessment. Having good technical documentation is invaluable in these presentations. Also, teaching nontechnical leaders who serve a state or school district about validity and the technical demands

of any testing program instills an understanding that will carry policy and future support forward in creating the best assessments possible for all our students.

Reports to the Federal Government (AYP)

The Federal government requires reports of student achievement and accountability because of the legislation referred to as *No Child Left Behind*. The guidelines for determining AYP may be changing, but the fact that this type of evaluation is continuing is not in doubt. Each state and school district has a responsibility to its students and to all others, including the Federal Department of Education, to create a defensible accountability system and to present evidence that its system leads to valid assessments of the entire school population including students with disabilities.

Summary

States and school districts have both a responsibility and an opportunity to report technical documentation results to its constituencies. There are many ways to spread information about the quality of a testing program. Test sponsors should use each of these communication options because dissemination of information about validity shows that the state or school district is carrying out its mission and shows that it is working on ways to improve its testing programs by eliminating or reducing threats to validity.

Evaluating the Testing Program

Buckendahl and Plake (in press) state that testing programs should undergo evaluation to improve the testing program and to fulfill the state's responsibilities to the public that the testing program serves. There has been a persistent belief among test developers that this kind of evaluation can be very helpful to test sponsors and to their constituencies (Downing & Haladyna, 1996; Madaus, 1992; Ruch, 1925). As with any consumer product, an independent, third-party

evaluation provides a credible source of information about the validity of test score interpretations and uses. If this third-party evaluation concurs with other evaluation reports including the technical report, then the public can have more confidence that their students are being well served.

Summary and Recommendations

This chapter began with a brief review of the validation process. Then the rationale for technical documentation of validity evidence was presented. The primary reason for technical documentation is to establish validity. Legal defensibility is another important issue. Good technical documentation of a testing program for students with disabilities may discourage costly and unfair legal challenges.

The *Standards for Educational and Psychological Testing* (AERA, et al., 1999) were cited as a basis for identifying categories of validity evidence, and specific standards were cited that bear on specific instances of validity evidence. Throughout the chapter, the standards were linked to categories of validity evidence and specific steps in test development, administration, scoring, and reporting.

As discussed in Chapter 2, and as illustrated further in Chapters 6, 7, 8, and 9, validity evidence can be very complex and exist in two forms, procedural and statistical/empirical. The technical report is a useful device for summarizing the validity evidence communicating it to interested parties. The technical report is also much more, because it also provides context and ends with recommendations for improvement of the testing program. The technical report might also identify a research agenda that addresses problems that may threaten test score validity. The technical report is not the only source of validity evidence. Other documents may exist and should be archived. The technical report will refer to these archived documents. Finally,

evaluation of the testing program should be done by an independent third-party; such evaluations can be invaluable in establishing the credibility of a testing program and in providing insights into how to improve the program. The following recommendations follow naturally and logically from the content of this chapter.

1. Any large-scale assessment program that includes students with disabilities should produce technical documentation that informs policy in that jurisdiction about the validity of using test scores for accountability purposes. The constituencies for the technical documentation are the Federal government, state government, local education agencies including school districts, and the public.

2. An annual technical report should be completed that displays the definition of achievement, the validity argument, the validity claim, and the validity evidence supporting the argument and the claim. Evidence opposing the claim also should be included and a discussion should offer analyses and recommendations for removing or minimizing the threats to validity arising from negative evidence.

3. The testing program sponsors should keep an archive of documents that contain the technical reports, validity studies, important memoranda, research reports, and any other documents that address procedures or empirical findings that bear on the categories of validity evidence (as presented in Chapter 2 and elaborated in Chapters 6, 7, 8, and 9, and in this chapter).

4. As a matter of policy, the results of the annual technical report should be made public. Access should be provided to all documents bearing on validity. Full disclosure is recommended, but only if the test itself is secure from disclosure that would undermine validity.

5. Evaluation by an independent third-party evaluator is highly desirable. Such evaluation may provide valuable insights into how the testing program might be improved. The summative part of the evaluation informs the test sponsors and the public about the validity of the sponsor's efforts to assess this population of students.

References

Abedi, J. (In press). Language issues in item development. In S.M. Downing and T.M. Haladyna (Eds). *Handbook of test development*. Mahwah, N.J.: Lawrence Erlbaum Associates.

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME) (1999). S*tandards for educational and psychological testing*. Washington, DC: Author.

Baranowski, R.A. (In press). Item editing and item review. In S.M. Downing and T.M. Haladyna (Eds). *Handbook of test development*. Mahwah, N.J.: Lawrence Erlbaum Associates.

Becker, D. F., & Pomplun, M. R. (In press). Technical reporting and documentation. In S.M. Downing and T.M. Haladyna (Eds). *Handbook of test development*. Mahwah, N.J.: Lawrence Erlbaum Associates.

Buckendahl, C., & Plake, B. (In press). Evaluating tests. In S.M. Downing and T.M. Haladyna (Eds). *Handbook of test development*. Mahwah, N.J.: Lawrence Erlbaum Associates.

Cronbach, L. J., & Meehl, P. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *51*, 281-302.

Downing, S. M. (in press). Twelve steps for effective test development. In S. M. Downing and T. M. Haladyna (Eds.), *Handbook of test development.* Mahwah, NJ: Lawrence Erlbaum Associates.

Downing, S. M. & Haladyna, T. M. (1996). A model for evaluating high-stakes testing programs: Why the fox should not guard the chicken coop. *Educational Measurement: Issues and Practices*, *15*(1), 5-12.

Haladyna, T.M. (In press). Roles and importance of validity studies in test development. In S.M. Downing and T.M. Haladyna (Eds). *Handbook of test development*. Mahwah, N.J.: Lawrence Erlbaum Associates.

Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice*. 23(1): 17-27.

Haladyna, T. R. (2004). *Developing and validating multiple-choice test items-2nd edition*. Mahwah, NJ: Lawrence Erlbaum.

Haladyna, T. M. (2002). Supporting documentation: Assuring more valid test score interpretations and uses. In G. Tindal & T. M. Haladyna (Eds.), *Large-scale assessment programs for all students: Validity, technical adequacy and implementation*. Mahwah, NJ: Lawrence Erlbaum Associates.

Joint Committee on Testing Practices (1988). *Code of Fair Testing Practices in Education*. Washington, DC: Author.

Kane, M. (In press). Content-related validity evidence in test development. In S.M. Downing and T.M. Haladyna (Eds). *Handbook of test development*. Mahwah, N.J.: Lawrence Erlbaum Associates.

Linn, R.L. (In press). The Standards for Educational and Psychological Testing: Guidance in test development. In S.M. Downing and T.M. Haladyna (Eds). *Handbook of test development*. Mahwah, N.J.: Lawrence Erlbaum Associates.

Madaus, G. F. (1992). An independent auditing mechanism for testing. *Educational Measurement: Issues and Practices*, *11*(1), 26-31.

Messick, S. (1984). The psychology of educational measurement. *Journal of Educational Measurement*, 21, 215–237.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–104). New York: American Council on Education and Macmillan.

Messick, S. (1995a). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*, 741–749.

Messick, S. (1995b). Standards of validity and the validity of standards in performance assessment. *Educational Measurement: Issues and Practice*, *14*(4), 5-8.

No Child Left Behind Act (NCLBA) of 2001, Pub. L. No. 107-110, §_1111, 115 Stat. 1449-1452 (2002).

Phillips, S. E. (1996). Legal defensibility of standards: Issues and policy perspectives. *Educational Measurement: Issues and Practices*, *15*, 5-14.

Ruch, G. M. (1925). Minimum essentials in reporting data on standard tests. *Journal of Educational Research*, *12*, 349-358.

Ryan, J. (In press). Practices, issues and trends in student test score reporting. In S.M. Downing and T.M. Haladyna (Eds). *Handbook of test development*. Mahwah, N.J.: Lawrence Erlbaum Associates.

Thurlow, M.L., Thompson, S.J., & Lazarus, S.S. (In press). Considerations for the administration of tests to special needs students: Accommodations, modifications, and more. In S.M. Downing and T.M. Haladyna (Eds). *Handbook of test development*. Mahwah, N.J.: Lawrence Erlbaum Associates.

Thurlow, M., House, A., Boys, C., Scott, D., & Ysseldyke, J. (2000). *State participation and accommodation policies for students with disabilities: 1999 update* (Synthesis Report 33). Minneapolis: University of Minnesota, National Center on Educational Outcomes.

Webb, N. (In press). Identifying content for assessing student achievement. In S.M. Downing and T.M. Haladyna (Eds). *Handbook of test development*. Mahwah, N.J.: Lawrence Erlbaum Associates.

Welch, C. (In press). Item/prompt development in performance testing. In S.M. Downing and T.M. Haladyna (Eds). *Handbook of test development*. Mahwah, N.J.: Lawrence Erlbaum Associates.

Zieky, M. (In press). Fairness review in assessment. In S.M. Downing and T.M. Haladyna (Eds). *Handbook of test development*. Mahwah, N.J.: Lawrence Erlbaum Associates.

CHAPTER 12: CONSEQUENCES OF ASSESSMENTS

When consequences are attached to testing outcomes, an emotional debate emerges among the pubic, parents, educators, and measurement experts, both praising the use of tests (e.g., Cizek, 2001; Driesler, 2001; Phelps, 1998, 2005) and condemning them (e.g., Kohn, 2001; Thompson, 2001). The purpose of this chapter is to provide a framework for examining the consequences of test score interpretation and use in a high-stake testing program. The basic principles in examining consequences of test interpretation and test use are the same for students with disabilities as those used with students who do not have disabilities.

The authors of the 1997 Amendments to the Individuals with Disabilities Education Act (IDEA) reasoned that the inclusion of students with disabilities in high stakes testing programs would result in positive consequences for students with disabilities (Ysseldyke, Dennison, & Nelson, 2004). The potential benefits included greater access to the general curriculum, improved instructional methods, and higher expectations for student learning (Browder, Fallin, Davis, & Karvonen, 2003; Browder, Spooner, Algozzine, Ahlgrim-Delzell, Flowers, & Karvonen, 2003; Quenemoen, Lehr, Thurlow, & Massanari, 2001; Thurlow & Johnson, 2000; Wehmeyer, Lattin, & Agran, 2001). However, some aspects of testing can have adverse as well as beneficial educational consequences (Messick, 1992). Examples of such consequences include promotion and retention decisions for students, salaries for educators, or rewards, sanctions, and interventions for schools (Forte Fast & Hebbler, 2004). As described in the *Guidelines for High Stakes Testing* (AERA, 2000), test developers and users should make a concerted effort to explain these possible effects to policymakers.

The Role of Consequences in Validity Theory

The role of consequences in validity theory has been debated for decades (Moss, 1998). Some argue that validity, reliability, comparability, and fairness of the results of assessments, are not just measurement principles; they are social values that have meaning and force whenever evaluative judgments and decisions are being made (Messick, 1992). The validity of the evaluation system in its entirety needs to be evaluated in terms of its effects in improving teaching and learning. Consequential aspects of validity must be explored (Messick, 1989, 1992; Shepard, 1997; Linn, 1997). Others take the opposite view (Popham, 1997; Wiley, 1991); they see the study of consequences as a complication to the concept of validity. Regardless of which side of the debate measurement experts embrace, the consensus is that actions tied to test results are a moral and ethical concern that must be examined. The *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999, 13.1) states, "It is the responsibility of those who mandate the use of tests to monitor their impact and to identify and minimize potential negative consequences. Consequences resulting from the uses of the test, both intended and unintended, should also be examined by the test user" (p. 145).

The *Standards* distinguishes between those consequences that are directly relevant to validity and other consequences that may inform decisions about social policy but may not be considered validity evidence (AERA, et al., 1999). While we conceptually separate a test from its consequences, an evaluation of the consequences associated with testing is necessary for validation of a system. Regardless of how accurately and consistently we can identify proficient students and schools; the exercise is meaningless if nothing happens to improve the work of schools (Marion & Gong, 2003).

Current thinking views validity as building on two major lines of evidence: evidential and consequential (Messick, 1989). Evidential validity evidence deals with the interpretability, relevance, and utility of test scores, and is discussed at length in other chapters of this volume. Consequential validity evidence examines the value implications of test scores as a basis for action and the actual and potential social consequences of using these scores. Consequential validity evidence is the focus of this chapter; it addresses the value implications of test scores as a basis for action and the actual as well as potential social consequences of using these scores.

Consequential Aspects of Assessments

The primary measurement concern with respect to adverse consequences is that any negative impact on individuals or groups should not derive from any source of test invalidity such as construct under-representation or construct-irrelevant variance (AERA et al., 1999; Messick, 1989). That is, low scores should not occur because the assessment is missing something relevant to the focal construct that would have permitted the affected students to display their competence[1]. Therefore, it is critically important to design assessments with full construct representation and limited irrelevant variance. By addressing these issues a priori, resulting test scores are a more accurate reflection of what students know and can do.

Construct under-representation. While general assessments, with and without accommodations, reflect standardized content coverage for all students, alternate assessments vary in breadth, depth, and complexity. Typically, IEP teams or teachers determine what is included in portfolios, observations, and performance tasks (see Chapter 3 for approaches to assessment), both in terms of content and substance. If the assessment contains only products/performances from a narrow content area, the construct may be under-represented and

[1] This issue has been discussed in greater detail in Chapter 2 of this volume.

inferences from the assessment results will be severely constrained. All approaches of assessments used in alternate assessments (i.e., ratings, portfolios, or performance tasks and events) can be prone to construct under-representation; they may have limited items/tasks/products because of practical feasibility, or they provide limited opportunity to observe behavior because of the disability. Alternate assessments also may consist of items/tasks/products requiring less cognitive demand (i.e., prompting students to engage in limited analysis, synthesis, or application of information), thereby reducing the full representation of the construct being assessed.

Many alternate assessments are designed to reflect knowledge, skills, and abilities using various items/tasks sampled from a domain. The construct being assessed is operationalized with a limited range of tasks that students complete and from which we make inferences to the larger domain. Performances/tasks traditionally used in special education classes have been based on a functional curriculum and may require skills or attributes that have nothing to do with the grade level academic content domains to be addressed in the large-scale assessment. Although the performance tasks are placed in a realistic context or authentic setting, they do not allow us to make inferences about knowledge, skills, or other attributes that should be assessed and what behaviors or performances reveal those knowledge, skills, or other attributes (Messick, 1994).

Messick (1994) recommends one solution to avoiding construct under-representation that may retain close similarity to full content coverage of grade level expectations: mixed sampling designs with different performance tasks administered to different samples of students. This would provide a picture of how groups of students are performing in content areas. Often, with portfolio assessments, this kind of design is inherent, requiring careful sampling of students when conducting an alignment study (see Tindal, 2005). Of course, this strategy comes at the

expense of reporting individual student's scores. Lane (2005) suggests another solution: have states carefully craft content standards, make the standards coherent across grade levels and across content areas that reflect student progress, make them accessible to teachers, and have teachers share their understanding of the standards and their implementation in instructional and testing practices.

Perhaps, the most direct way to reduce construct under-representation or misrepresentation is for state educational agencies to design assessment tasks based on the academic content standards. This can decrease the construct-under-representation in alternate assessments, especially for those alternate assessments that allow teachers the flexibility to select academic content standards and design activities/tasks to meet each student's individual needs.

Construct irrelevant variance. Any factors that systematically raise or lower a test score *unfairly* suggest construct-irrelevant variance. Student characteristics, cheating, and/or test preparation can affect test scores and lead to inaccurate inferences concerning a student's proficiency (Haladyna & Downing, 2004). A case can be made that there is a higher level of construct-irrelevant variance in alternate assessments for several reasons. Some states give teachers and IEP teams flexibility in selecting assessment content, determining levels of expected performance, administering tests, and scoring the results of the tests. All of these provide many opportunities not only for increasing unsystematic error (affecting reliability) but also for introducing construct-irrelevant variance into assessment results (and thereby affecting validity).

Several methods are available for examining construct-irrelevance in assessments after the assessments have been constructed. Differential item and test functioning (DIF and DTF, respectively) can be used to statistically detect construct-irrelevance (see Chapter 7 for more

discussion). Traditional DIF studies have examined gender and ethnic group effects, but type of disability (interfering behaviors) also should be examined. Limited research exists exploring DIF on large-scale assessments for students with disabilities (see for example, Yovanoff & Tindal, in press). Another procedure for examining differential functioning of items is an alignment process proposed by Webb (1997) and Resnick, Rothman, Slattery, and Vranek (2003-2004), both of whom examine *sources of challenge* whereby experts are asked to evaluate each item/task for construct-irrelevance. A thorough examination of construct irrelevance in assessment combines expert judgment when designing items or tasks, followed by a statistical DIF procedure on the data collected to confirm accuracy of expert judgment (Kopriva, 2000).

Types of Consequences

Evaluation of consequences requires analysis of the full array of intended and unintended impact. Two types of consequences should be considered in any accountability system, imposed or emergent (Forte Fast & Hebbler, 2004). Imposed consequences are those required or mandated by the local educational agency (LEA) or state education agency (SEA). These include the rewards offered for high levels of performance, sanctions for low levels of performance, and interventions to support improvement of performance. Emergent consequences are those activities that occur either in anticipation of imposed consequences or following them chronologically. They include changes in (a) professional development, (b) instructional methods, or (c) resource allocations. Whether imposed or emergent, previous research provides the most useful starting point for gathering information on the consequences of high stakes testing. Gathering evidence of the intended effects is a necessary first step (Lane, Parke, & Stone, 1998).

Results of accountability systems are intended to impact: (a) the implemented curriculum; (b) the instructional content and strategies; (c) the content and format of classroom assessments; (d) student, teacher and administrator motivation and effort; (e) learning experienced by all students; (f) the nature of professional development support; (g) teacher participation in administration, development, and scoring of assessment; (h) student, teacher, administrator, and public awareness and beliefs about the assessment, criteria for judging performance, the use of the assessment results; and (i) the use and nature of test preparation materials (Cizek, 2001; Frederiksen & Collins, 1989; Koretz, Barron, Mitchell, & Stecher, 1996; Koretz, Stecher, Klein, & McCaffrey, 1994; Linn, 1993; Linn, Baker, & Dunbar, 1991; Messick, 1992).

The examination of potential misuses of the assessment and the assessment results (i.e., unintended consequences) may include: (a) the narrowing of curriculum and instruction to focus only on the specific learning outcomes assessed and ignoring of the broader construct reflected in the specified learning outcomes (e.g., Chabran, 2003; Cizek, 2001; Corbett & Wilson, 1991; Mehrens, 1998; Smith & Rottenberg, 1991); (b) use of test preparation materials that are closely linked to the assessment without making changes to the curriculum and instruction (e.g., Firestone, Camilli, Yurecko, Monfils & Mayrowetz, 2000); (c) use of unethical test preparation materials; (d) differential performance on the assessment for subgroups of students; (e) an inappropriate or unfair use of test scores, such as questionable practices in the reassignment of principals and teachers; (f) undesirable effects on teacher morale (e.g., Corbett & Wilson, 1991; Lattimore, 2001; Smith & Rottenberg, 1991); and (g) students' diminished self-esteem and increased stress level (Smith & Rottenberg, 1991).

Consequences of Accommodations

No Child Left Behind legislation mandates that *all* students will be proficient in reading, math, and science by the year 2014. Thus, a positive consequence of high-stakes testing has been an increased focus on grade level academic content for students with disabilities. With this focus on the achievement of students with disabilities has come a widespread interest in the use of test accommodations for students who otherwise might not be able to demonstrate their full knowledge. Cizek (2005) lists this widespread awareness of the use of test accommodations as a *positive, unintended* consequence of high-stakes testing. He also argues that this awareness has spilled over into the creation of classroom assessments. Research and professional development activities being undertaken at CAST (Center for Applied Special Technology) further demonstrate how the concept of accommodations has become central to instructional design (especially when considered as an a priori universal design feature).

Many states allow all students to use accommodations during testing, and do not limit accommodation use to students with disabilities. For example, because of the systemic nature of accommodations being allowed for all students and given the court mandate (see the settlement from the Advocates for Special Kids – ASK), Oregon has initiated a system for formally judging the potential consequences of assessment changes as part of allowing them to become accommodations. In this system, teachers from various disability groups, administrators, and university researchers are impaneled. Throughout the year, requests for changes are collected and quarterly the panel is convened to judge whether the change should be allowed as an accommodation using four criteria. First, the purpose of the change is determined. In this process, the panel attempts to understand how access is being limited and how the change would eliminate the access problem. Second, the function of the change is ascertained by explicating

how testing would look and operate with the changes being implemented. Third, the consequences of making an error are considered, both in denying the request for the individual as well as allowing it. In the former case (considered a false negative), issues are raised about the individual student being unable to complete the test in a standard fashion, while in the latter case (considered a false positive), issues are raised in the mistakes that would be made in perhaps unfairly raising the student's score. Finally, the consequences of the change are considered from a systemic perspective in which the panel deliberates on the potential impact of the change on the entire assessment system. In this analysis costs are considered as well as potential reactions of stakeholders (parents, teachers, administrators) and issues of professional development. In the end, these four criteria are an attempt to consider consequences in a fair and equitable manner and most importantly in a prospective manner to make recommendations to the state educational agency. This process is one of the few known to consider consequences as a deliberative component of the testing system and elevate the judgments to a public forum.

Consequences of Alternate Assessments

Despite over a decade of examining the consequences of general education assessments, disagreement exists over the impact of high stakes testing. Mehrens (1998) found that (a) evidence for a test's influence on curricular content or instruction is not clear; (b) the effect of large-scale assessments on teacher motivation, morale, stress, and ethnical behavior is sketchy; and (c) little empirical evidence exists on how assessments affect students.

In a more recent review by Thomas (2004), examples of damage are presented through hundreds of newspaper articles collected from across the country. At the heart his book, Thomas addresses them from various stakeholders' positions: politicians and their staff, administrators and their staffs, the public and parents, test-makers and test givers, teachers, and students. In his

last chapter (lessons to learn), he demarcates the major problems as (a) the fitness of lawmakers to make decisions about educational reform, (b) the fallacy of developing tests with a 'one size fits all' perspective, (c) the competition of academics on the world stage (with international testing), (d) the violation of a basic physics principle that states "a space that is completely full cannot accommodate anything new until something has been removed" (p. 270), and (e) the application of a static model versus a growth model.

In contrast to the negative examples shared by Thomas (2004) is the publication of a book edited by Richard Phelps (2005) supporting high-stakes assessments: *Defending Standardized Tests*, in which he claims that "Never before has obtaining such an abundance of accurate, rich, and useful information about student learning been possible" (p. xii). This same perspective is echoed by Cizek (2005), who lists evidenced-based positive consequences of high-stakes testing: (a) meaningful and focused professional development, (b) teachers' increased understanding of assessment practices and principles, (c) improved use of assessment data in classrooms, schools, and districts, and (d) improved curriculum alignment across teachers and classrooms. There is definite disagreement about the effects (intended and unintended) of high-stakes testing.

For alternate assessments, implementation has outpaced research on their impact and consequences. Most of what we know about the consequences of alternate assessments has come from teacher and administrator surveys (Ahlgrim-Delzell, & Wakeman, 2005; Flowers, Browder, & Ahlgrim-Delzell, in press; Kampfer, Horvath, Kleinert, & Farmer Kearns, 2001; Kleinert, Kennedy, & Kearns, 1999; Towles, Garrett, Burdette, & Burdge, 2003). Not surprisingly, the findings from the consequences of alternate assessments are like those found in general education assessments. The following sections represent five major categories of intended and

unintended consequences of alternate assessments (modified from the work of Mehrens, 1998): (a) curricular, instructional, and IEP reforms; (b) improvement in student learning; (c) teacher motivation, morale, stress, beliefs about testing process, and ethical behavior; (d) student motivation and self-concept; and (e) communication with parents, constituents, and public. These intended and unintended consequences, both negative and positive, are not meant to be exhaustive; the potential consequences may be endless.

Curricular, Instructional, and IEP Reforms

Earlier works have discussed how alternate assessments may impact curriculum, instruction, and IEP reforms (Browder, Flowers, Ahlgrim-Delzell, Karvonen, Spooner & Algozzine, 2004; Browder, Spooner, et al., 2003; Ford, Davern, & Schnorr, 2001). Alternate assessments were designed to help promote access to the general curriculum; that is, to broaden the curricular focus for this student population. Therefore, they have a more academic focus than traditional life skills curricula for students with significant cognitive disabilities. One consequence of having students with disabilities included in alternate assessments might be that teachers would improve their instruction and focus more on academic grade level expectations. Yet, there have been inconsistent findings about the impact of alternate assessments on curricular and instructional reforms.

On the positive side, survey findings suggested that alternate assessments increased focus on the general curriculum and the use of professional development activities to help teachers align instruction with general curriculum frameworks (Zatta & Pullin, 2004). Another positive consequence of alternate assessment was that teachers reported that the alternate assessment process encouraged the use of age-appropriate materials (Flowers, et al., in press). Teachers also

reported that inclusion of students with disabilities in large-scale assessment programs, at any level, has resulted in a more demanding curriculum (Olson, 2004).

The results of a 2003 survey of 983 special education teachers from five states (Flowers, et al., in press) reflect that most teachers do not believe that alternate assessments increase SWD access to the general curriculum, or that students with disabilities participate in more general education classroom activities. More than half of the teachers report no improvement in the quality or variety of their instruction. About half the teachers report that alternate assessment competed with the individual student's needs and few teachers report that alternate assessment make it easier to provide community-based instruction. Most teachers do not perceive that students with disabilities are receiving an overall better-quality education, and only 24% perceive an improvement in the quality of the IEP objectives. Thirty-nine percent of teachers report they have adequate resources to undertake the alternate assessment process.

The results of a survey of 708 special education administrators (Ahlgrim-Delzell, & Wakeman, 2005) from 49 states and districts were like those of special education teachers. Most administrators do not agree that the implementation of alternate assessments increase special education teachers' knowledge of general curriculum, require curriculum revisions for students, or increase general education teachers' expectations of students with disabilities. Survey participants do not agree that participation in alternate assessments lead to higher expectations for students with disabilities by special education and general education teachers increases in understanding of the needs of students with disabilities by principals, or better instructional strategies by their teachers. In contrast, teachers report benefits such as increased student choice making and use of schedules (Kleinert, Kearns & Kennedy, 1997; Kleinert, Kennedy, et al,

1999). A list of potential curricular, instructional, and IEP consequences is reported in Table 12.1.

Table 12.1. Curricular, Instructional, and IEP Reform Consequences of Alternate Assessments

	Curriculum, Instruction, and IEP
Positive	
	Access to the general curriculum
	Greater participation in general education classroom
	Age/grade appropriate curriculum for students with significant disabilities (SWSD)
	Improved instruction—variety of strategies
	Improved classroom assessments
	Improved progress monitoring of SWSD
	Improved IEP objectives
	Greater student progress on IEP objectives
	Greater resources (assistive technologies, paraprofessionals)
	Increased professional development opportunities
	Improved quality of professional development opportunities
Negative	
	Narrowing the curriculum
	Neglecting important knowledge/skills in functional curriculum
	Competes with individual student's needs

Improvement in Student Learning

Improved scores may represent improvement on the domain tested, but it does not necessarily follow that it is a true improvement in a student's education (Mehrens, 1998). Results from a 28-state descriptive study conducted by Amrein and Berliner (2002) provide little support for a relationship between high-stakes testing and increased student achievement on other validated measures of learning. While in the field of special education, there is evidence that students with severe disabilities are meeting proficient levels on alternate assessment; findings on the impact of alternate assessment on overall student learning are mixed. Survey results suggest that teachers believe that students with severe disabilities should be included in state accountability systems, that states are setting higher expectations for students with disabilities (Flowers, et al., in press), and that students are learning more self-determination skills (Kleinert, Kearns, & Kennedy, 1997; Kleinert, Kennedy, & Kearns, 1999).

But most survey findings do not provide evidence that alternate assessments improve student learning or that alternate assessments are improving the overall quality of education for SWSD (Ahlgrim-Delzell, & Wakeman, 2005; Flowers, et al., in press). In Kentucky, Kleinert et al. (1999) reports that teachers were concerned that the assessment was more a reflection of their performance than the performance of their students. A list of potential positive and negative consequences of participation in large-scale assessment is reported in Table 12.2.

Table 12.2. Improvement in Student Learning

Improvement in Student Learning	
Positive	
	Increased learning (not just increase in test scores)
	Increased expectations
	Overall improvement in the quality of education
	Greater generalizability of skills
	Greater self-determination
	Greater transition to adult living
Negative	
	Differential performance for subgroups
	Test the degree of the student's disability and not ability
	Widen the achievement gap

Teacher Motivation, Morale, Stress, Beliefs About Testing Process and Ethical Behavior

In the larger assessment arena, researchers report instances of teachers engaging in questionable testing behaviors under the pressures of high-stakes accountability (Cizek, 1999). Examples include preparing students for an assessment by practicing actual test items or clones of items (Popham, 1991), inappropriately cueing students during testing, or changing incorrect answers before submitting results (Cizek, 1999).

It is important that any reform of the large-scale assessment process have support from special education teachers. The assessment is unlikely to impact instruction if teachers do not see any connection between the assessment results and their instructional approaches (Mehrens, 1998). A consistent finding on the consequences of alternate assessments is that they create more paperwork for both teachers and administrators (Ahlgrim-Delzell, & Wakeman, 2005; Flowers,

et al, in press) and they compete with teachers' instructional and personal time (Ahlgrim-Delzell, & Wakeman, 2005; Flowers, et al, in press; Kleinert, Kearns, & Kennedy, 1997; Kleinert, Kennedy, & Kearns, 1999). In addition, teachers report that students with severe disabilities could not meet the expectations set by the states, AA scores do not reflect the true abilities of their students, and teachers question the accuracy and consistency of AA scores (Flowers, et al., in press). A list of potential teacher motivations, morale, stress, beliefs about testing process, and ethical behaviors is reported in Table 12.3.

Table 12.3. Teacher Motivation, Morale, Stress, Beliefs About Testing Process and Ethical Behavior

	Teacher
Positive	
	Increased motivation
	Increased effort
	Greater knowledge of general education curriculum
	Belief about usefulness of content standards and assessments
	Belief in technical quality of assessment
	Raise general education teacher's expectations
	More involvement of school administration
Negative	
	Increased stress
	Decreased morale
	Increased paperwork
	Increased attrition
	Unethical behaviors (e.g., cheating)
	Lack of support

Student Motivation and Self Concept

There has been little research on the effects of general or alternate assessments on student motivation. The intended consequence for assessments is to increase students' motivation to learn, but some hypothesize that high-stakes tests cause increased anxiety and stress (Hancock, 2001; Zohar, 1998). Some research suggests the use of performance-based assessments (not specifically alternate assessments) increases students' motivation to learn and increases the amount of engagement with performance tasks and portfolio assignments more than other types

of assessments (Kane & Mitchell, 1997). In one large-scale study of alternate assessments, teachers of students with disabilities report that most students are not aware of the AA process and do not understand their tests results (Flowers, et al., in press).

Table 12.4. Student Motivation and Self-Concept

	Students
Positive	
	Increased motivation
	Improved self-determination
	Increase value of SWD in school climate
	More attentions from school administration
Negative	
	Feelings of inadequacy
	Increased frustration
	Increased stress
	Minimum understanding or involvement in assessment process
	Students misplaced in modified or alternate achievement standards

Communication With Parents, Constituents, and the Public

Reporting test results provides that outside of schools' evidence of how schools and students are performing. States need to examine whether assessment results and changes in schools based on those results are valued by parents, constitutes, and the public. Collaboration between educators and parents has long been considered a best practice indictor for programs for students with severe disabilities (Meyer, Eichinger, & Park-Lee, 1987). Discussing the alternate assessment process in the IEP, collaborating with parents to obtain evidence of achievement of state standards, and communicating the meaning of outcomes are three ways that educators can involve parents more in the alternate assessment process.

There is little research on the consequences of alternate assessment on communications with parents, constituents, and the public. In a teacher survey, few teachers believe that AA increases their ability to communicate with parents or that it increases the involvement of parents in the assessment process (Flowers, et al., in press).

Table 12.5. Communication With Parents, Constitutes, and the Public

Parents, Constitutes, and Community	
Positive	
	Improve communication between schools and parents
	Improve the view of SWD
	Educational changes are valued by those in the natural environment
	Benefit to having SWD in accountability programs
	Restore public confidence (Nothing to hide)
Negative	
	Reassignment of teachers and principals
	Provide data for critics

Recommendations for Collecting Evidence About the Consequences of Assessment Programs

Consequences imply a cause-and-effect relationship, but most of the research methods used to examine the consequences of test use do not allow educators to draw causal inferences (Mehrens, 1998; Reckase, 1997). Despite this limitation, empirical evidence for the relationship of assessment scores and external criteria of success for students with disabilities can be tested, and it is reasonable to evaluate the extent to which empirical evidence suggests positive and/or negative consequences (Lane, 1999).

As a first step in collecting evidence concerning the consequences of test use, it is important to review documents describing the purpose and intended outcomes of the assessment program (Lane, 1997). Improving instruction and student achievement are the ultimate goals of an accountability system, but a review may describe the specific tangible outcomes of these programs, especially for alternate assessments. While the academic content standards are the same for *all* students, alternate assessments judged against alternate achievement standards usually include other outcomes valued for students with the most significant disabilities (e.g., self-determination). An evaluation of the curriculum, standards, or learning outcomes allows an evaluator to reflect on the specific goals of the educational program. Furthermore, evaluation of professional development, resources, and support places the evaluation in the appropriate context.

Evidence needs to be collected from multiple levels inside and outside the educational program (Lane, 1997). Collecting data at the state, district, school, and classroom levels can ensure a comprehensive evaluation of the consequences (Cronbach, 1989). For example, examining the curriculum and interviewing the curriculum specialist can provide evidence of the impact of adopting curriculum. Of course, educational agencies need to consider time, cost, and feasibility and would want to make long term plans for looking across levels.

Much of the consequential evidence is based on surveys of teachers and principals. Surveys are cost effective and increase generalizability because of their typically large sample sizes. Surveys have limitations, however, in that they cannot capture complex instructional practices, may lack shared understanding of terminology, and are susceptible to self-report bias (Lane, 1997). Surveys can only capture the perceptions of the respondents and may not reflect the actual effect.

Lane et al. (1998) argue for more direct evidence in evaluating consequences, such as analyses of classroom instruction and assessment activities. Observations and case studies have the potential of capturing the complexity of instructional practices and can illustrate factors contributing to quality instruction and student learning. The limitations to these methodologies are that they are costly, time consuming, labor intensive, and lack generalizability. Collection of classroom artifacts (e.g., lesson plans, classroom assessments, test preparation materials, instructional materials) provides a direct method for evaluating instructional practices but is also time intensive.

There has been promising work on developing indicators of classroom practice to monitor and support school reform (Aschbacher, 1999; Porter, & Smithson, 2001). The tools being developed can be used to rate important dimensions of classroom practices, instruction,

academic content, and student work. These tools provide special educators with the unique opportunity to develop valid and reliable methods of capturing accurate data on what is happening in the classroom with instruction and content.

Questions for Evaluating the Consequences of Test Use

Although consequences can be documented using these five variable (curricular, instructional, and IEP reforms; improvement in student learning; teacher motivation, morale, stress, beliefs about testing process and ethical behavior; student motivation; and communication with parents, constituents, and the public), it is important to act on the information collected and to use it to improve and validate the entire assessment and accountability system. The *Theory of Action* (Forte Fast & Hebbler, 2004) provides an accountability system validation process framework. The key premises within this framework are: (a) map the system by identifying the goals and specify the actions and logic by which the system is meant to achieve these goals; (b) evaluate the indicators by reviewing the definitions and construction of each indicator; (c) evaluate the decision rules by analyzing the process by which indicators are combined and fed into decisions; and (d) evaluate the consequence by analyzing the full array of them that are or could be applied to schools. Forte Fast and Hebbler (2004) suggest five major questions that should be answered for evaluating consequences:

- How well are rewards, sanctions, and interventions implemented?
- How do school and LEA characteristics, as well as other facets of the context, moderate the implementation of the consequences?
- To what degree are the intended actions occurring in relation to the application of rewards, sanctions, and interventions?

- To what degree are negative, unintended consequences occurring in relation to the application of rewards, sanctions, and interventions?
- To what degree are the reform activities associated with achievement the goal of the systems?

Baker, Linn, Herman, and Koretz (2002) recommend longitudinal studies be planned and implemented, and that they report evaluations of the effects of the accountability program. Minimally, questions should determine the degree to which the system: (a) builds capacity of staff; (b) affects resource allocation; (c) supports high-quality instruction; (d) promotes student equity access to education; (e) minimizes corruption; (f) affects teacher quality, recruitment, and retention; and (g) produces unanticipated outcomes.

Summary and Recommendations

Our efforts to properly assess students with disabilities can have both positive and negative consequences. We encourage states and their educators to continue to document positive consequences when assessing students with disabilities but also note (and minimize) negative consequences that come from the various threats to validity as discussed. Full analysis of all possible consequences related to test interpretation and test use is likely to create an unreasonably large universe of issues to examine. Furthermore, evaluation of the consequences of assessments is an on-going process with no perfect method for fully understanding all the potential positive and negative consequences. Therefore, states need to set priorities and collect evidence that demonstrates progress over time and good faith implementation of assessment programs. In the end, for any accountability system, the positive consequences must outweigh the negative consequences by some factor greater than the costs (Mehrens, 1998). The challenge

for states is to establish a system of data collection that monitors the assessment program and an intervention strategy that will effectively address the quality of their assessment program.

1. Procedural evidence needs to be systematically collected in the development and implementation of large-scale assessment system to document threats to construct irrelevant variance and construct under-representation.

2. Both positive and negative consequences need to be collected and collated in a public forum that allows an appropriate analysis of evidence to argue for or against the central validation claim in all aspects of the assessment system: for tests administered with accommodations and for all forms of alternate assessments.

3. Attention should be given to the consequences in three arenas: (a) in the measurement technologies that get developed and implemented in an assessment approach, (b) in the populations of students who are assessed, and (c) in the content areas that get addressed through various assessment approaches.

4. Given that much of the evidence is of unknown worth at the time of its collection, a system needs to be implemented for cataloging and cross classifying potential relationships that may emerge over time and have influence on accommodations and alternate assessments.

5. Because the content standards at each grade level provide such a broad array of knowledge and skill, any changes in the tests to better include students with disabilities needs to reinforce skill acquisition in the service of learning content while documenting this relationship from multiple perspectives (curriculum, learning, student and teacher motivation, teacher professional development, and policy changes).

References

Agran, M., Alper, S., & Wehmeyer, M. (2002). Access to the general curriculum for students with significant disabilities: What it means to teachers. *Education and Training in Mental Retardation and Developmental Disabilities*, *37*, 123-133.

Ahlgrim-Delzell, L. & Wakeman, S. (2005, April). *School administrators' perceptions of the benefits and consequences of including students with significant disabilities in school.* Paper presented at the annual meeting of the North Carolina Association of Research in Education, Chapel Hill, NC.

American Educational Research Association (AERA), American Psychological

American Educational Research Association. (2000). *Guidelines for high stakes testing.* Washington, DC: Author.

Amrein, A. L., & Berliner, D. C. (2002). *The Impact of High-Stakes Tests on Student Academic Performance: An Analysis of NAEP Results in States with High-Stakes Tests and ACT, SAT, and AP Test Results in States with High School Graduation Exams.* Retrieved on July 1, 2005, from http:wwwgreatlakescenter.org. Educational Policy Studies Laboratory, Arizona State University: Arizona.

Aschbacher, P. (1999). *Developing indicators of classroom practices to monitor and support school reform*. Retrieved March 9, 2005, from CRESST Web site: http://www.cse.ucla.edu/CRESST/Reports/TECH513.PDF

Association (APA), & National Council on Measurement in Education (NCME) (1999). S*tandards for educational and psychological testing*. Washington, DC: Author.

Baker, E. L., Linn, R. L., Herman, J. L., & Koretz, D. (2002). *Standards for educational accountability systems* [Policy Brief 5]. Los Angeles: UCLA National Center for Research on Evaluation, Standards, and Student Testing.

Browder, D. M., Fallin, K., Davis, S., & Karvonen, M. (2003). A consideration of what may influence student outcomes on alternate assessment. *Education and Training in Developmental Disabilities*, *38*, 255-270.

Browder, D. M., Spooner, F., Algozzine, R., Ahlgrim-Delzell, L., Flowers, C., & Karvonen, M. (2003). What we know and need to know about alternate assessment. *Exceptional Children, 70*, 45- 62.

Browder, D., Flowers, C., Ahlgrim-Delzell, L., Karvonen, M., Spooner, F., & Algozzine, R. (2004). The alignment of alternate assessment content with academic and functional curricula. *The Journal of Special Education, 37,* 211-223.

Chabran, M. (2003). Listening to talk from and about students on accountability. In M. Carnoy, R. Elmore, & L. Siskin (Eds.), *The New Accountability High Schools and High-stakes Testing* (pp. 129-145). New York: Routledge Falmer.

Cizek, G. (2001). More intended consequences of high-stakes testing. *Educational Measurement: Issues and Practice, 20(4)*, 19-27.

Cizek, G. J. (1999). *Cheating on tests: How to do it, detect it, and prevent it*. Mahwah, NJ: Lawrence Erlbaum Associates.

Cizek, G. J. (2005). High-stakes testing: Contexts, characteristics, critiques, and consequences. In R. P. Phelps (Ed.), *Defending standardized testing* (pp. 23-54). Mahwah, NJ: Lawrence Erlbaum.

Corbett, H. D., & Wilson, B. L. (1991). *Testing, reform, and rebellion*. Norwood, NJ: Ablex.

Cronbach, L. J. (1989). Construct validation after thirty years. In R. E. Linn (Ed.), *Intelligence: Measurement, theory and public policy* (pp. 147-171). Urbana: University of Illinois Press.

Driesler, S. D. (2001). Whiplash about backlash: The truth about public support for testing. *NCME Newsletter, 9(3)*, 2-5.

Firestone,W. A., Camilli, G., Yurecko, M., Monfils, L., & Mayrowetz, D. (2000). State standards, socio-fiscal context and opportunity to learn in New Jersey. *Education Policy Analysis Archives, 8(35)*.

Flowers, C., Browder, D., & Ahlgrim-Denzell, L. (in press). The alignment of three states alternate assessments to state standards. *Exceptional Children.*

Ford, A., Davern, L., & Schnorr, R. (2001). Learners with significant disabilities. *Remedial and Special Education, 22,* 214-222.

Forte Fast, E. & Hebbler, S. (2004). *A Framework for Examining Validity in State Accountability Systems.* Council of Chief State School Officers, Washington, DC.

Frederiksen, J. R., & Collins, A. (1989). A district's approach to educational testing. *Educational Researcher, 18(9),* 27-42.

Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance: A threat in high-stakes testing. *Educational Measurement: Issues and Practice*, *23(1)*, 17-27.

Hancock, D. R. (2001). Effects of text anxiety and evaluative threat on students' achievement. *Journal of Education Research, 94*(5), 284-290.

Individuals with Disabilities Education Act Amendments of 1997, 20 U.S.C. §1400 et. seq.

Kampfer, S. H., Horvath, L. S., Kleinert, H. L., & Farmer Kearns, J. (2001). Teachers' perceptions of one state's alternate assessment: Implications for practice and preparation. *Exceptional Children, 67*, 361-374.

Kane, M, & Mitchell, R. (1997). *Implementing performance assessment: Promises, problems and challenges*. Mahwah, NJ: Erlbaum.

Kleinert, H. L., Kearns, J. F., & Kennedy, S. (1997). Accountability for all students: Kentucky's alternate portfolio assessment for students with moderate and severe cognitive disabilities. *The Journal of The Association for Persons with Severe Handicaps*, *22*, 88-101.

Kleinert, H. L., Kennedy, S., & Kearns, J. F. (1999). The impact of alternate assessments: A statewide teacher survey. *The Journal of Special Education*, *33*, 93-102.

Kohn, A. (2001). Fighting the tests: A practical guide to rescuing our schools. *Phi Delta Kappan, 82(5),* 349-357.

Kopriva, R. (2000). *Ensuring Accuracy in Testing for English Language Learners*. Council of Chief State School Officers: Washington DC.

Koretz, D. M., Barron, S., Mitchell, K. J., & Stecher, B. M. (1996). *Perceived effects of the Kentucky instructional results information district* (MR-792-PCT/FF). Santa Monica, CA: RAND.

Koretz, D. M., Stecher, B., Klein, S., & McCaffrey, D. (1994). The Vermont portfolio assessment program: Finding and implications. *Educational Measurement: Issues and practice, 13(3),* 5-16.

Lane, S. (1997, March). *Framework for evaluating the consequences of an assessment program*. Paper presented at the annual meeting of the National Council of Measurement in Education, Chicago, IL.

Lane, S. (1999, October). *Validity evidence for assessments*. Lecture presented at the Annual Reidy Interactive Lecture Series, Providence, RI.

Lane, S. (2005, January). *Consequential Validity*. Presentation for CCSSO SCASS-ASR Accountability Systems and Reporting.

Lane, S., and Parke, C. (1996, April). *Consequences of a mathematics performance assessment and the relationship between the consequences and student learning*. Paper presented at the annual meeting of the National Council on Measurement in Education, New York

Lane, S., Parke, C. S., & Stone, C. A. (1998). A framework for evaluating the consequences of assessment programs. *Educational Measurement: Issues and Practice,* 24-28.

Lattimore, R. (2001). The wrath of high-stakes tests. *Urban Review, 33*(1), 57-67.

Linn, R. L. (1993). Educational assessment: Expanded expectations and challenges. *Educational Evaluation and Policy Analysis, 15(1),* 1-16.

Linn, R. L. (1997). Evaluating the validity of assessments. *Educational Measurement: Issues and Practice, 16(2)*, 14-18.

Linn, R. L., Baker, E. L., & Dunbar, s. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher, 20(8),* 15-21.

Linn, R.L., and Herman, J.L. (1997, February). *Standards-led assessment: Technical and policy issues in measuring school and student progress.* CSE Technical Report 426. National Center for Research on Evaluation, Standards, and Student Testing (CRESST) Center for

the Study of Evaluation, Graduate School of Education and Information Studies, University of California, Los Angeles.

Madaus, G. F. (1985). Test scores as administrative mechanisms in educational policy. *Phi Delta Kappan*, 66(9), 611-617.

Madaus, G.F. (1991). The effects of important tests on students: Implications for a National Examination System. *Phi Delta Kappan*, *73*(3), 226-231.

Marion, S. F., & Gong, B. (2003, October). *Evaluating the validity of state accountability systems.* Lecture presented at the Annual Reidy Interactive Lecture Series, Nashua, NH.

Mehrens, W. A. (1998). Consequences of assessment: What is the evidence? *Educational Policy Analysis Archives, 6.* Downloaded on 6/22/2005. http://epaa.asu.edu/epaa/v6n13.html

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

Messick, S. (1992). *The interplay of evidence and consequences in the validation of performance assessments* (ETS RR-92-39). Princeton, NJ: Educational Testing Service.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23,* 13-23.

Messick, S. (1995). Standards of validity and the validity of standards in performance assessment. *Educational Measurement: Issues and Practice, 14(4),* 5-8.

Meyer, L. H., Eichinger, J., & Park-Lee, S. (1987). A validation of program quality indicators in educational services for students with severe disabilities. *The Journal of The Association for Persons with Severe Handicaps*, *12*, 251-263.

Moss, P. A. (1998). The role of consequences in validity theory. *Educational Measurement: Issues and Practice, 17,* 6-12.

Olson, L. (2004, January 8). Enveloping expectations. *Education Week, 23,* 8-20

Phelps, R. P. (1998). The demand for standardized student testing. *Educational Measurement: Issues and Practice, 17*(3), 5-23.

Phelps, R. P. (Ed.). (2005). *Defending standardized testing* Mahwah, NJ: Lawrence Erlbaum.

Popham, W. J. (1997). Consequential validity: Right concern – wrong concept. *Educational Measurement: Issues and Practice*, *16*(2), 9-13

Popham, W. J. (1991). Appropriateness of teachers' test-preparation practices. *Educational Measurement: Issues and Practice, 10* (4), 12-16.

Porter, A. C. & Smithson, J. L. (2001). *Defining, developing, and using curriculum indicators* (CPRE Research Report Series RR-048). University of Pennsylvania: Consortium for Policy Research in Education.

Quenemoen, R. F., Lehr, C. A., Thurlow, M. L., & Massanari, C. B. (2001). *Students with disabilities in standards-based assessment and accountability systems: Emerging issues, strategies, and recommendations* (Synthesis Report 37). Minneapolis, MN: Retrieved September 5, 2001: http://education.umn.edu/NCEO/OnlinePubs/Synthesis37.html

Reckase, M. D. (1997, March). *Consequential validity from the test developers' perspective*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Resnick, L. B., Rothman, R., Slattery, J. B., & Vranek, J. L. (2003-2004). Benchmarking and alignment of standards and testing. *Educational Assessment*, *9*(1 & 2), 1-27).

Shepard, L. A. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice, 16(2),* 5-8, 13, 24.

Smith, M. L., & Rottenberg, C. (1991). Unintended consequences of external testing in elementary schools. *Educational Measurement: Issues and Practice, 10(4),* 7-11.

Thomas, R. M., (2005). *High stakes testing: coping with collateral damage*. Mahwah, NJ: Lawrence Erlbaum.

Thompson, S. (2001). The authentic testing movement and its evil twin. *Phi Delta Kappan, 82(5),* 358-362.

Thurlow, M. L., & Johnson, D. R. (2000). High stakes testing of students with disabilities. *Journal of Teacher Education*, *51*, 305-314.

Towles, E. A., Garrett, B., Burdette, P. & Burdge, M. (2003). *What are the consequences? Validation of large-scale alternate assessment systems and their influence on instruction.* University of Kentucky. (ERIC Document Reproduction Service No. ED478482)

Webb, N. L. (1997). *Research Monograph No. 6: Criteria for alignment of expectations and assessments in mathematics and science education*. Washington, DC: Council of Chief State School Officers.

Wehmeyer, M. L., Lattin, D., & Agran, M. (2001). Achieving access to the general curriculum for students with mental retardation: A curriculum decision-making model. *Education and Training in Mental Retardation and Developmental Disabilities*, *36*, 327-342.

Wiley, D. E. (1991). Test validity and invalidity reconsidered. In R. E. Snow & De. E. Wiley (Eds.), *Improving inquiry in the social sciences: A volume in honor of Lee J. Cronbach* (pp. 75-107). Hillsdale, NJ: Erlbaum.

Yovanoff, P., & Tindal, G. (in press). Scaling early reading alternate assessments with statewide measures. *Exceptional Children*.

Ysseldyke, J., Dennison, A., & Nelson, R. (2004). *Large-scale assessment and accountability systems: Positive consequences for students with disabilities*. NCEO Synthesis Report 51, May 2004. Downloaded on 2/17/2005. http://education.umn.edu/NCEO/OnlinePubs/Synthesis51.html

Zatta, M C. & Pullin, D. C. (2004). Education and alternate assessment for students with significant cognitive disabilities: Implications for educators. *Education Policy Analysis Archives, 12(16)*.

Zohar, D. (1998). An additive model of test anxiety: role of exam-specific expectations. *Journal of Educational Psychology, 90*(2), 330-340.

CHAPTER 13: IMPLICATIONS FOR PROFESSIONAL DEVELOPMENT

By federal regulation and statute, students with disabilities *must* participate in statewide assessments, particularly those developed and administered to meet the accountability requirements of NCLB and IDEA. In this volume, we have presented a model that encompasses all types of assessments used as part of a state accountability system. We have stressed the need for technical adequacy framed as a validity argument for all the assessments adopted within a state, regardless of the approach(es) used. We have applied the *Standards for Educational and Psychological Testing* (1999) and the regulations explicated in the Federal Register (December 9, 2003) to guide our discussions of assessment options for students with disabilities and the evidence needed to assure the validity of any test score interpretation derived from these assessment options. Our approach also is compliant with the peer review guidance (U. S. Department of Education, April 28, 2004).

If states are to implement the ideas, procedures, and practices described in this volume to develop and implement assessment systems with strong technical adequacy, they need to invest considerable resources in professional development focused on at least four groups: measurement experts who need to know more about students with disabilities and the assessments that are appropriate for them; special education professionals who need to become more expert in measurement principles, in general and their applicability in particular to assessments designed for students with disabilities; educational leaders including principals, who oversee the participation of all students in large scale assessments; and individuals likely to serve on IEP teams who have the responsibility for recommending a particular assessment option for each individual

student with disabilities. In this chapter, we outline briefly the focus of professional development for each of these four groups. We begin with a quick overview of some professional development principles, define what each constituent group needs to know, and provide an index of content from the prior chapters that can be used in this training.

Professional Development Principles

Context, process, and content are the three key elements in designing effective professional development. The *context* for professional development needs to be conducive to learning; this may be achieved through the creation of learning communities, the guidance of leadership, and the deployment of resources. The *process* of training should focus on the outcomes of learning, the use of collaborative efforts, and an evaluation of training effectiveness. The *content* of training should be responsive to stakeholders' needs (test coordinators, teachers, related service providers, and administrators). The National Staff Development Council (2001) provides standards for each of these elements to guide staff development personnel in their design of professional development experiences. The Maryland (2001) guide on professional development reflects these three key elements and notes that professional development is most effective when it takes place in vibrant professional learning communities (context), is data-driven – utilizing effective professional development with rigorous analysis of data (process), leading to knowledge, skills, and dispositions that apply research to decision making (content).

Effective professional development also must be a *continuing* process that is successful in creating *lasting* changes in behavior and practice (Jones & Lowe, 1990). To be responsive to changes in policy and to new research on assessing students with

disabilities, states need to develop a system for ongoing information delivery. One indicator of the effectiveness of this ongoing information delivery system will be whether professional practice becomes more consistent with new policy and research over time.

One caveat: Professional development aimed at improving the technical quality of assessment practices is designed to provide more valid inferences about the performance of students with disabilities, but it will not necessarily link directly to improvements in student learning (Sparks & Hirsh, 2000). First, data obtained from an improved assessment system may not be comparable to prior reports on student outcomes. Second, professional development on "who should take what" by itself would not be expected to improve student performance. Instead, educators need resources on how to improve the quality of curriculum and instruction and how to use research-based procedures like understanding key measurement, assessment, and inclusion principles (Assessing One and ALL, Council for Exceptional Children) and applying progress monitoring to improve student outcomes (see www.studentprogress.org).

Content for Professional Development

In the sections that follow, we assume that the design of professional development incorporates appropriate consideration of context and process. We focus here specifically on the content, drawing from the discussion in earlier chapters of this volume.

Primary Resources for Developing the Content of Professional Development

Two primary resources should form the basis of professional development (PD) on participation of students with disabilities in large-scale assessments. The first resource is the *Standards for Educational and Psychological Testing* available from the American Educational Research Association (1999). The second resource is the federal regulations

that specify this participation, such as IDEA 2004 (PL 108-446), NCLB (2002), the regulations on alternate achievement standards (Federal Register, December 9, 2003), and the most current policy statements on the Department of Education's website introducing the new modified achievement standards (www.ed.gov). This volume is offered as a third resource for PD (along with the Tool Kit being developed from the Office of Special Education Programs). States also may want to include their individual state guidelines for the participation of students with disabilities in state assessments. Technical assistance centers offer additional resources that may be useful to PD planners. For example, the IEP team may benefit from information on linking to academic content standards for students with significant cognitive disabilities available through the National Alternate Assessment Center (http://www.naacpartners.org). The National Center on Education Outcomes has numerous reports on the participation of students with disabilities in large-scale assessments (http://education.umn.edu/nceo). The Access Center provides direct assistance, networking, and web-based resources to assist states in building the capacity for all students to access the general curriculum (http://www.k8accesscenter.org). The National Center on Accessing the General Curriculum provides professional development on accessing the general curriculum and universal design (http://www.cast.org/ncac). A complete list of technical assistance centers can be found at http://www.dssc.org/frc/TAGuide/index0013.html

And finally, the web course *Assessing One and All* on the Council for Exceptional Children website (http://www.cec.sped.org/pd/webcourses/cec112.html) uses a case focused approach to understanding key measurement, assessment and inclusion practices

and features links to every state's testing guideline and AYP reports, facilitating a very real, state-individualized feel to the cases.

What Measurement Professionals Need to Know

Measurement professionals working with state assessment systems come to their jobs with expertise in and in-depth knowledge of the *Standards for Educational and Psychological Testing* (1999). They may not be familiar, however, with the need to apply these standards to the entire range of assessment options available for students with disabilities. Given that most will likely have had limited exposure to students across the full range of disabilities, it may be difficult for them to appreciate how the application of these standards may differ for students with disabilities and why these differences are needed.

Their major need, therefore, is an understanding of students with disabilities and the implications of disability for participations and performance on statewide accountability assessments. Professional development might begin with specific information about special education and students who have IEPs. State measurement professionals need to be able to use current terminology in referring to individuals with disabilities and have some understanding of the students' variation in response modes and support needs. They may need a handout that includes the current categories of disabilities in IDEA 2004 and offers both definitions and terminology. Current state categories and definitions would also be relevant. In this context, it may be useful to explain briefly why the term for this population changed in the law from "Handicapped Children" to "Individuals with Disabilities" and to emphasize the importance of people-first language.

State measurement professionals also need to understand that neither individual education plans in general nor assessment participation decisions are based on disability categories. Once a student is eligible for special education services, consideration must be given to the individual's needs in planning a free and appropriate education and in selecting an assessment option.

Measurement professionals need to understand that students with disabilities may demonstrate learning in a wide variety of response modes. They need to allow for this individualization of response and participation in the design and development of assessment options and understand the inferences that can be made from scores derived from different testing formats. They need to apply elements of universal design to assessments to create accessibility for many students who have previously been unable to participate in assessments or demonstrate their knowledge. Finally, they need to appreciate how response variations also are important in considering appropriate accommodations.

State measurement professionals need to know current federal regulations concerning the participation of students with disabilities in statewide assessments and accountability systems. Providing copies of these policies, along with state guidelines, is important. Working from careful knowledge of these regulations is especially important for measurement professionals who develop state-level policies for participation in assessment or who develop alternate assessments.

Measurement professionals need a working knowledge of the concept of "access to the general curriculum." They need to understand that access to the general curriculum is an expectation for ALL students with disabilities. In this context, it may be useful to

clarify why some terms used by measurement specialists can be confusing when applied to the education of students with disabilities, their access to general curriculum, and their achievement of academic content standards. For example, the term "developmental" has a specific meaning in measurement circles; it is used to describe a type of derived or transformed score such as *developmental age* or *developmental quotient.* Early childhood educators use the term "*developmentally* appropriate" to refer to an educational curriculum that is appropriate to students' current abilities. Among those involved in the education of students with severe disabilities, the term "*developmental*" refers to an outmoded practice of continuing to apply an early childhood curriculum throughout the student's lifespan by planning his/her educational program based on the student's mental age[1]. In the same way, the phrase "off grade level" may have different meanings to the measurement and the special education communities. For measurement specialists, the phrase may refer to the scaling of content. To the special educator, this phrase may be used to describe the selection of appropriate teaching materials from a lower grade level to help a student access the general education grade level content (teaching a fifth-grade curriculum using a book written at a second-grade readability level).

In helping measurement professionals understand access to the general curriculum, several resources may be useful. Technical assistance centers that focus on general curriculum access (e.g., www.k8accesscenter.org) provide summary documents that define general curriculum access and provide examples. Applications of this content to students with significant cognitive disabilities may be found through NAAC (http://www.naacpartners.org).

[1] The current focus in educating students with severe disabilities is to use chronologically age-appropriate educational activities and those that provide opportunities to participate in the learning of grade level content.

Of course, professional development for measurement specialists also may have to include specific references to the application of measurement principles to the development and validation of validity of alternate assessment formats. Few states have created or circulated the technical reports on their alternate assessment(s). While typical technical issues, such as item discrimination and differential item functioning, may need to be rethought when applied to alternate assessment formats, documentation and dissemination of technical adequacy must be done. This volume provides useful and practical guidance for gathering validity and reliability evidence.

If the validity arguments presented in Chapter 2 are to be adequately addressed for each testing format used by a state, measurement professionals in the state may need increased knowledge of approved accommodations, the state's inclusion criteria for participation in assessments options, and the reporting of student performance in a standardized manner. They need to address construct misrepresentation and construct irrelevant variance as they relate to assessments for students with disabilities. Further, scoring constructed response assessments require their careful thought and planning. As specified in Chapters 6 and 7, subjective scoring may introduce bias to the judgments made about student performance; scorers must be adequately trained and made aware of potential biases.

With greater knowledge of students with disabilities, test accommodations and modifications, and technical considerations, measurement professionals can make a more substantive contribution to the annual review of technical data related to all the assessment approaches used a state for students with disabilities, and can help to strengthen or increase the inferences made from those assessments.

Table 13.1 summarizes the content that might be addressed in professional development for measurement specialists responsible for including students with disabilities in the statewide assessment and accountability systems.

What Special Educators Need to Know about Measurement

Special educators involved in state planning on how students with disabilities participate in large-scale assessments need to have a thorough knowledge of current federal regulations (Federal Register December 9, 2003, IDEIA, 2004, NCLB, and new guidelines as they appear). Because policies and regulations about this participation have been evolving rapidly, ongoing professional development is especially critical for this professional group. They not only need information about policy and regulation, but they also need copies of new policies and regulations in hand when developing guidelines for participation and/or new alternate assessments for use in individual states. Other helpful resources may be obtained from the National Center on Education Outcomes (education.umn.edu/NCEO) and Research from the Educational Policy Reform Research Institute (www.eprri.org).

Table 13.1. Summary of What Measurement Professionals Need to Know about Special Education

Topic	Examples	Possible Resources
Standards for Testing	-Definition of terms -Standards for validity	*Standards for Educational and Psychological Testing* (1999).
Application of Standards to Students with Disabilities	-Inferences to be made -Alternate assessments -Accommodations	*Including Students with Disabilities in Large-Scale Assessment Systems: A Model for Validating Inferences (*this volume) Tindal, G. & Haladyna, T. M. (Eds.) (2002). *Large-scale assessment for all students: Validity, technical adequacy, and implementation*
Students with Disabilities-Categories, Response Variation, Support	-Categories used -People first language -Why accommodations may be needed -Supports used by students in alternate assessment	IDEA, 2004 State policy http://education.umn.edu/nceo

Table 13.1, continued

Topic	Examples	Possible Resources
Federal Regulations	-All students participate -Not assigned to assessment option based on disability -AYP and students with disabilities	IDEA, 2004 Federal Register, December 9, 2003 www.ed.gov
Access to General Curriculum	-Expectation for access -Applicability of state content standards	www.k8accesscenter.org http://www.naacpartners.org

Access to the general curriculum is a new topic for many special educators, as pointed out by Nolet and McLaughlin (2000). This can create some confusion in how participation of students with disabilities is planned and carried out. Just as special educators rely on measurement experts to be current in their knowledge of standards for assessment, measurement professionals need special educators to have the most current information on access to the general curriculum to plan how students' achievement of state academic content standards can be assessed. Professional development on access to the general curriculum can begin with web site summaries of the issues and preferred practices (www.k8accesscenter.org; http://www.naacpartners.org). The PD should also include more detailed information on how to promote this access (Agran, Alper, & Wehmeyer, 2002; Browder & Spooner, in press; King-Sears, 2001; Wehmeyer, Lattin, & Agran, 2001).

Special educators also need a thorough understanding of accommodations and how they affect participation in large-scale assessments. The *Standards for Educational and Psychological Testing* (1999) define accommodations as any action taken in response to a determination that an individual's disability requires a departure from established testing protocol. Koretz and Barton (2003) acknowledge that while there is limited research about the selection of appropriate accommodations for students with disabilities,

a key to devising those accommodations is an understanding of what biases may be caused by the disability and what alterations of the test might alleviate those biases without producing an unfair advantage.

Special educators who are aware of the needs of the student and range of available accommodations may be more able to perform the difficult task of selecting appropriate accommodations, which should be selected for individual students not applied uniformly to all students with disabilities. But before special educators can do that effectively, they need to be familiar with the extent to which accommodations may influence the construct being measured, the state and local guidelines related to accommodations and assessment, and the ways in which an accommodation may manifest itself within the assessment format (Elliott, McKevitt, & Kettler, 2002; Thurlow, House, Boys, Scott, & Ysseldyke, 2000).

Approved accommodations typically, however, vary by state. Special educators need to be aware of what are approved and non-approved accommodations in their state and what the use of those accommodations mean in terms of test scores and accountability for students. Bielinski, Sheinker, and Ysseldyke (2003) discussed the disaggregation of student test scores by states according to the use of accommodations. States may report all test scores in the aggregate, report accommodated scores separately, or report both. These reporting differences can reflect intended and unintended consequences for students and schools, as discussed in Chapter 10.

To be effective partners in planning, administering, and interpreting appropriate statewide assessments for students with disabilities, special educators need to be much more familiar with acceptable standards for assessment, scoring, and reporting of scores.

Few special educators, whether they are teachers in the field or administrative personnel working in a state department of education, have a solid foundation in measurement, in general, and measurement as it applies to assessment of students with disabilities, in particular. Special educators need more knowledge about each of the major considerations required in making a validity argument: achievement constructs and how they are measured; reliability; test scores and the importance of the validity argument; validity evidence, both procedural and empirical; and finally, construct misrepresentation and construct irrelevant variance. They should have a general understanding of how inferences drawn from an assessment are validated. In the context of statewide assessments, they need to have a thorough understanding of the differences between content and performance standards (or in the language of NCLB, achievement standards). Special educators need information from the *Standards for Educational and Psychological Testing;* Chapter 10 of this guide is devoted exclusively to testing students with disabilities. Additional resources may be also useful to develop this understanding (see Tindal and Haladyna, 2002).

Table 13.2 summarizes the content that might be addressed in ongoing professional development for special educators whose students with disabilities will be included in the statewide assessment and accountability systems.

Table 13.2. Summary of What Special Educators Need to Know About Measurement

Topic	Examples	Resources
Federal Regulations	-All students participate -Not assigned to assessment option based on disability -AYP and students with disabilities	IDEA, 2004 NCLB Federal Register, December 9, 2004 www.ed.gov
Standards for Testing Application to Students with Disabilities	-Definition of terms -Standards for validity -Applications to students with disabilities	*Including Students with Disabilities in Large-Scale Assessment Systems: A Model for Validating Inferences (*this volume) Tindal, G. & Haladyna, T. M. (Eds.) (2002). *Large-scale assessment for all students: Validity, technical adequacy, and implementation*
Access to General Curriculum	-How to promote access -How relates to assessing achievement of state standards	Browder & Spooner (in press for 2006) Kleinert & Kearns (2001) Nolet & McLaughlin (2000) Agran, Alper, & Wehmeyer (2002) King-Sears (2001) Wehmeyer, Lattin, & Agran (2001)
Accommodations-Impact on Inferences that Can be Made	-Research on accommodations -Appropriate inferences	Bilenski, Sheniker, & Ysseldyke (2003) Elliott, McKevitt, & Kettler (2002) Koretz & Barton (2003) Thompson, Blount, & Thurlow (2002)

What Educational Leaders Need to Know

The need for professional development for principals in special education has been well established in the literature (DiPaola & Walther-Thomas, 2003; Goor, Schwenn, & Boyer, 1997; Monteith, 2000; Sage & Burrello, 1994). Specifically, however, educational leaders need professional development to understand the participation of students with disabilities in large-scale assessments. School leaders are keenly aware of AYP and the requirement of reporting separately on the subgroup of students with disabilities. As written in NCLB (2002), schools are held accountable for the Adequate Yearly Progress of all students within their school, including those with disabilities. However, Farkas, Johnson, and Duffett (2003) found that 48% of the principals surveyed in 2001 and 2003 identified the requirement to demonstrate Adequate

Yearly Progress with special education students and English as a Second Language learners as unreasonable. As student annual performance scores are increasingly routinely disaggregated by disability status, the performance of students with disabilities may have serious consequences for students, schools, and administrators.

McLaughlin and Thurlow (2003) have documented the shift in accountability for special education that spans from simple compliance to access to education, to evidence of student learning and performance. To keep pace with this shift, administrators not only need to understand special education law (Davidson & Algozzine, 2002; Davidson & Gooden, 2001), they also must have knowledge of assessments (psychological and educational), of inclusive assessment practices including universal design, and of indicators of best practice instructional strategies for students with disabilities (Monteith, 2000, Patterson, Marshall, & Bowling, 2000). As students with disabilities participate in large-scale assessments with the various participation options described in this volume, administrators need to have specific knowledge of each type of assessment and be familiar with the needs of students who participate in each option.

Administrators may need training related to the selection and use of state adopted accommodations and alternate assessments. Elliot, Braden, & White (2001) described the parameters around which decisions should be based to include all students in large scale assessments. Administrators who are familiar with these parameters can effectively participate in IEP team meetings and facilitate discussion of inclusive instructional practices including the use of classroom accommodations when appropriate. It also is important that principals understand the alternate assessment format(s) used in their state, the content measured on the assessment(s), the requirements of performance by students

using those assessments, and the inferences that can be made when the performance of students with disabilities is judged against modified or alternate achievement standards.

Additionally, administrators must be vigilant in their efforts to ensure that instruction for *all* students is aligned to state grade level content standards. Greene-Bryant (2002) described the need for assessment and instruction to be in direct alignment for all students, including those with disabilities. As all students are included in assessments that are linked to or derived from general education grade level content standards, administrators need to be aware of initiatives that help bridge students' access to and learning of those standards. School administrators need to lead their schools in the alignment of instruction, content, and assessment while at the same time avoiding too great a narrowing of the curriculum or "teaching to the test".

Finally, administrators must be able to help teachers use assessment data to inform their decisions about modifying instruction. The Interstate School Leadership Licensure Consortium (ISLLC) (Council for Chief State School Officers, 1996) defined standards for school leaders that include the use of assessment data in shaping the school vision and instructional program. The American Association for School Administrators (http://www.aasa.org/cas/resources_and_tools.htm) provides a website listing a variety software tool, links to research, and sample plans to assist administrators in making data-based decisions. A summary of what school leaders need to know is shown in Table 13.3.

Table 13.3. Summary of What Educational Leaders Need to Know About Participation of Students with Disabilities in Statewide Assessment

Topic	Examples	Resources
Accommodations	-Guides for use -Support	Elliott, Braden, & White, (2001) Greene-Bryant (2002) McLaughlin & Thurlow (2003) State policy manuals
Alternate assessments	-Guides for use -Build knowledge	Browder, et al. (2003) Greene-Bryant (2002) Hager & Slocum (2005) State policy and assessment administration manuals
Alignment of instruction, content standards, and assessment	-Relationship of educational objectives, content, and test	Elliott, Braden, & White, (2001) Greene-Bryant (2002) Roach, Elliott, & Webb (2005) State technical manuals Tucker & Codding (2002)
Data-based decisions	-Types of data to collect -Use of data	ERS (2003a) ERS (2003b) Jaeger & Tucker (1998) www.aasa.org/cas/resources_and_tools.htm

What IEP Teams Need to Know

The fourth group seriously in need of professional development consists of individuals who are likely to serve on IEP teams. This includes parents, but also general education teachers, special education teachers, special and general education administrative personnel (supervisors, principals, LEA representatives), school psychologists, etc. It is the responsibility of the IEP team to make the critical recommendation about an assessment option for each student with a disability. Team members must be equipped with knowledge about instruction, supports, curriculum, and assessment components to come to an appropriate recommendation.

To begin, potential IEP team members need to understand state and federal legislative efforts and their impact for students with disabilities. This understanding lays the foundation for consensus building necessary for teams' members to agree about a particular student's strengths and needs in accessing the general curriculum and participating in statewide assessment. State and local education agencies are charged with

the responsibility of communicating legislative mandates in user friendly language to support the community's comprehension of education law.

IEP team members must understand the options for participation by students with disabilities in large-scale assessments and the distinguishing elements of each option (see Chapter 4 on the IEP team decision framework). They need to have clear insight into the content of the assessments (i.e., complexity and cognitive demand of each assessment option), the performance requirements (e.g., the inclusion of universal design concepts in grade level assessments) and the participation parameters (i.e., descriptions or checklists created by states related to who the assessments are most appropriate for). States, districts, and schools must work to educate IEP team members including parents and advocates about these critical issues.

Teams must also understand the consequences of their decisions for each student. Teams need to be aware of state and district diploma requirements especially if assigning a student to an alternate assessment judged against alternate achievement standards may cause a student to be placed in something other than a diploma track. In addition, IEP members, including special education teachers, must recognize that assessments are not tied to disability category or placement; students with disabilities assigned to the same classroom may participate in different assessment options.

An important distinction to be made by IEP teams is the curricular basis for the assessment recommendation. An IEP teams that values goals outside of the grade level content may decide to incorporate augmentative curricular elements into a student's IEP. However, it is the grade level content standards that serve as the underpinning of what is addressed in the annual statewide assessment, regardless of the type of assessment

selected. Teams should have knowledge of grade level content to set priorities, if necessary, related to what and how much of the content standards should be covered during the IEP timeframe. Their recommendation to simplify grade level content may lead them to also recommend participation in statewide assessments judged against modified or alternate achievement standards. Teams that have a clear understanding of the content of instruction and the assessment options are better equipped to make suitable recommendations that support student learning.

In states where alternate assessment requires the special education teacher and the IEP team create a portfolio of student accomplishments, members of the IEP team likely need extensive training on how to pinpoint priorities within grade level academic standards and document progress on these priorities. They may need training on selection and alignment of assessment items to general education content standards.

Accommodations for students with disabilities are not necessarily new to IEP team members. What may be new is the form those accommodations may take or the adoption of new accommodations by states. Teams must be acquainted with what are state adopted acceptable instructional and assessment accommodations for students with disabilities.

Because student performance scores “count”, teams must understand what those scores mean in terms of instruction and performance. Teams can use assessment information to make appropriate decisions about what to include in the student’s IEP in terms of the content to be covered and the supports needed. Teams must also be cognizant that assessment content and outcomes should accurately reflect the learning of

the student. As this point is critical for assessments to be considered valid, team members really should be knowledgeable about the uses and meanings of student scores.

Finally, IEP members need practice in using the *Checklist of Educational Needs,* presented as Table 4.1 (Chapter 4) that guides the IEP team decision framework regarding the assessment option recommendation. This practice might include having discussions about simulated students with disabilities for whom different recommendations may be appropriate, or reviews of real case files of students for whom decisions have been made or need to be made.

Table 13.4 summarizes the content that might be addressed in professional development for potential members of IEP teams whose responsibility it is to recommend a participation option for each student with disabilities for whom they plan an appropriate Individualized Education Program (IEP).

Summary and Recommendations

This chapter included an overview of the types of information needed by the various professionals involved in the participation of students with disabilities in large scale assessments. States may find other content is needed to address the unique assessment options and administration guidelines for their context. Our goal in this chapter, and in this volume, was to provide a resource in planning this professional development. Table 13.5 provides an index to information in the chapters in this volume that may be especially useful in planning and carrying out needed professional development.

Table 13.4 What *IEP Teams Need to Know about Participation of Students with Disabilities in Statewide Assessment*

Topic	**Examples**	**Resources**
Federal and state policies	-Who participates -AYP -Accommodations -Alternate Assessments	IDEA, 2004 NCLB State policy manuals
All the participation options	-State's allowable accommodations -State's types of alternate assessments -Alternate and modified achievement standard options if available in state -Decision guide	State policy manuals See Tables 4.1 and 4.2 which were created for IEP team use
Consequences of participation decision	-Impact on graduation -Scope and complexity of curriculum student expected to learn	State policy manuals
Access to general curriculum	-How to align IEP goals to state standards -IEP goals that do and do not relate to achieving academic content standards	Courtade-Little & Browder (2005). Kleinert & Kearns (2001) Nolet & McLauglin (2000) Thompson, Quenemoen, Thurlow, & Ysseldyke (2001)
Accommodations	-Allowable accommodations -How to plan for use of accommodations in instruction	-Additional articles Clapper, Morse, Lazarus, Thompson, & Thurlow (2005) Elliott, McKevitt, & Kettler (2002). Thompson & Thurlow (2003) Thurlow, House, Boys, Scott, & Ysseldyke (2000)

Table 13.5. Index of Content in this Volume for Use in Professional Development

Chapter	Content for Professional Development	Related Professional Decision
1. Overview and Assessment Model	• Concept of validity and its importance in making inferences from assessments • Participation options for students with disabilities • Levels of inference in participation options applied to grade level content standards	• Development and implementation of assessment practices • Which of possible participation options the state utilizes • Understanding concepts of *support, breadth, depth, and complexity* as applied in the development of alternate assessments
2. Validating Assessments	• Examining validity as an argument with claim and evidence • Value of procedural and empirical validity evidence • Validity issues specially construct under-representation or misrepresentation and construct irrelevant variance	• Developing and choosing assessment formats
3. Approaches to Assessment	• Assessment formats used nationwide and validity issues with each	• Developing assessment formats and guidelines to minimize validity concerns
4. IEP Team Decision Framework	• Decisions IEP teams do not make about participation in state assessments • Student characteristics to consider in making participation decisions	• Choosing assessment participation option (for IEP teams) • Creating state guidelines for eligibility in assessment participation options
5. Examples of Content Standards and Assessment Approaches	• Standards: *validity claim* is that the test is developed to adequately reflect the domain of knowledge and skills for the construct and the domain of tasks from which we sample measuring the construct. • Examples of changes in breadth, depth, and complexity in the development of portfolio entries or performance tasks	• Designing state assessments (state level) • Applying state standards to students with differing abilities without construct misrepresentation
6. Test Design, Development, Administration, and Scoring	• Concepts of Universal Design in the development of inclusive assessments • Alignment of assessment content with grade level content standards • Issues in administration and scoring of large-scale assessments	• Teacher bias in administration and scoring assessments for students with disabilities • Alignment when content is reduced in breadth, depth, or complexity
6. Evidence of Item Quality	• Considerations in item quality	• Developing assessment items • Aligning items in assessment to content standards
7. Reliability-Related Evidence	• Repeatability of the behavior elicited by the test and the consistency of the resultant scores • Sources of potential error in various participation option	• Developing assessment guidelines • Applying consistent evaluations of performance
8. Validity Evidence	• Validity in content, measurement, and use of assessment outcomes	• Developing assessment guidelines • Interpreting outcomes

Table 13.5, continued

9. Content Validity	• Documentation of content coverage, response processes, internal structure of the test, and relationship with other acceptable measures	• Role of teachers in the response process for students with more severe disabilities
10. Score Reporting and Standard Setting	• Issues in score reporting • Role of achievement level descriptions • Setting achievement standards	• Meaning of achievement standards in alternate assessment
11. Technical Documentation	• *Standards* related to technical documentation • Organization of the Technical Report • Disseminating reports of the validity of statewide assessments	• Understanding and communicating technical adequacy to various constituents
12. Consequences of Assessments	• Concept of consequential validity • Positive and negative consequence of participation statewide assessment • Intended and unintended consequences	• Recruiting and responding to stakeholder input
13e Professional Development	• What measurement professionals, special educators, educational leaders, and members of IEP teams need to know about measurement concepts, and about participation of students with disabilities in statewide assessments	• What to include in professional development on including students with disabilities in state assessments

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.

Bielinski, J., Sheinker, A., & Ysseldyke, J. (2003). *Varied opinions on how to report accommodated test scores: Findings based on CTB/McGraw-Hill's framework for classifying accommodations* (Synthesis Report 49). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved June 30, 2005, from the World Wide Web: http://education.umn.edu/NCEO/OnlinePubs/Synthesis49.html

Browder, D.M., & Spooner, F.H. (In press). *Teaching reading, math, and science to students with significant cognitive disabilities*. Baltimore: Paul H. Brookes.

Browder, D. M., Spooner, F. Algozzine, R., Ahlgrim-Delzell, L., Flowers, C., & Karvonen, M. (2003). What we know and need to know about alternate assessment. *Exceptional Children*, *70*, 45-61.

Clapper, A.T., Morse, A.B., Lazarus, S.S., Thompson, S.J., & Thurlow, M.L. (2005). *2003 state policies on assessment participation and accommodations for students with disabilities* (Synthesis Report 56). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved June 30, 2005, from the World Wide Web: http://education.umn.edu/NCEO/OnlinePubs/Synthesis56.html

Council of Chief State School Officers. (1996). *Interstate School Leadership Licensure Consortium Standards for School Leaders*. Washington, D.C.: Author.

Courtade-Little, G., & Browder, D. M. (2005). *Aligning IEPs to academic standards for students with moderate and severe disabilities*. Verona, WI: Attainment Company.

Davidson, D. N., & Algozzine, B. (2002). Administrators' perceptions of special education law. *Journal of Special Education Leadership*, *15*, 43-48.

Davidson, D. N., & Gooden, J. S. (2001). Are we preparing beginning principals for the special education challenges they will encounter? *ERS Spectrum*, *19*, 42-49.

DiPaola, M. F., & Walther-Thomas, C. (2003). *Principals and special education: The critical role of school leaders*. (COPPSE Document No. IB-7). Gainesville, FL: University of Florida, Center on Personnel Studies in Special Education.

Educational Research Service. (2003a). *Essentials for principals: Data-based decision making*. Arlington, VA: Author.

Educational Research Service (2003b). Using data for instructional improvement. *ERS focus on*. Retrieved July 13, 2005, from the World Wide Web: http://e-library.ers.org/pdf/premium/ERS%20Focus%20On/FO2003c.pdf

Elliott, S. N., Braden, J. P., & White, J. L. (2001). *Assessing one and all: Educational accountability for students with disabilities*. Arlington, VA: Council for Exceptional Children. (ERIC Document Reproduction Service No. ED458746)

Elliott, S. N., McKevitt, B. C., & Kettler, R. J. (2002). Testing accommodations research and decision-making: The case of "good" scores being highly valued but difficult

to achieve for all students. *Measurement and Evaluation in Counseling and Development*, *35*, 153-166.

Goor, M. B., Schwenn, J. O., and Boyer, L. (1997). Preparing principals for leadership in special education. *Intervention in School and Clinic, 32,* 133-141.

Greene-Bryant, B. (2002). *Including students with disabilities in large-scale assessments: A necessary element of school reform efforts*. Reston, VA: National Association of Secondary School Principals. (ERIC Document Reproduction Service No. ED463603)

Hager, K. D., & Slocum, T. A. (2005). Using alternate assessment to improve educational outcomes. *Rural Special Education Quarterly*, *24*, 54-59.

Individuals with Disabilities Education Improvement Act of 2004, P. L. No. 108-446, 20 U.S.C. section 611-614.

Jaeger, R. M., & Tucker, C. G. (1998). *Analyzing, disaggregating, reporting, and interpreting students' achievement test results: A guide to practice for Title 1 and beyond*. Washington, DC: Council of Chief State School Officers.

Jones, E. V., & Lowe, J. (1990). Changing teacher behavior: Effective staff development. *Adult Learning*, *1*, 8-10.

King-Sears, M. E. (2001). Three steps for gaining access to the general curriculum for learners with disabilities. *Intervention in School and Clinic*, *37*, 67-76.

Kleinert, H. L., & Kearns, J. F. (Eds.). (2001). *Alternate assessment: Measuring outcomes and supports for students with disabilities*. Baltimore: Paul H. Brookes.

Koretz, D. M., & Barton, K. (2003). *Assessing students with disabilities: Issues and evidence* (Technical Report 587). Retrieved June 27, 2005, from the World Wide Web: http://www.cse.ucla.edu/reports/TR587.pdf

Maryland Department of Education. (2001). *Maryland teacher professional development standards*. Annapolis, MD: Author. Retrieved June 27, 2005, from the World Wide Web: www.mdk12.org/instruction/professional_development/teachers_standards.html

McLaughlin, M. J., & Thurlow, M. (2003). Educational accountability and students with disabilities: Issues and challenges. *Educational Policy*, *17*, 431-451.

Monteith, D. S. (2000). Professional development for administrators in special education: Evaluation of a program for underrepresented personnel. *Teacher Education and Special Education*, *23*, 281-289.

National Staff Development Council. (2001). *NSDC standards for staff development*. Retrieved June 27, 2005, from the World Wide Web: www.nsdc.org/standards/

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

Nolet, V., & McLaughlin, M. J. (2000). *Accessing the general curriculum*. Thousand Oaks, CA: Corwin Press.

Patterson, J., Marshall, C., & Bowling, D. (2000). Are principals prepared to manage special education dilemmas? *NASSP Bulletin, 84,* 9-20.

Roach, A. T., Elliott, S. N., & Webb, N. L. Alignment of alternate assessment with state academic standards: Evidence for the content validity of the Wisconsin alternate assessment. *Journal of Special Education*, *38*, 218-231.

Sage, D. D., & Burrello, L. C. (1994). *Leadership in educational reform: An administrator's guide to changes in special education.* Baltimore: Paul H. Brookes.

Sparks, D., & Hirsh, S. (2000). Strengthening professional development. *Education Week*, *19*, 42-43.

Thompson, S. J., Quenemoen, R. F., Thurlow, M. L., & Ysseldyke, J. E. (2001). *Alternate assessments for students with disabilities*. Thousand Oaks, CA: Corwin Press.

Thompson, S., Bount, A., & Thurlow, M. (2002). *A summary of research on the effects of test accommodations: 1999 through 2001* (Technical Report 34). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved July 12, 2005, from the World Wide Web: http://education.umn.edu/NCEO/OnlinePubs/Technical34.htm

Thompson, S., & Thurlow, M. (2003). *2003 State special education outcomes: Marching on*. Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved June 30, 2005, from the World Wide Web: http://education.umn.edu/NCEO/OnlinePubs/2003StateReport.htm

Thurlow, M., House, A., Boys, C., Scott, D., & Ysseldyke, J. (2000). *State participation and accommodation policies for students with disabilities: 1999 update* (Synthesis Report 33). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved July 12, 2005, from the World Wide Web: http://education.umn.edu/nceo/OnlinePubs/Synthesis33.html

Tindal, G. & Haladyna, T. M. (Eds.) (2002). *Large-scale assessment for all students: Validity, technical adequacy, and implementation*. Mahwah, NJ: Lawrence Erlbaum Associates.

Tucker, M. S., & Codding, J. B. (2002). *The principal challenge: Leading and managing schools in an era of accountability*. San Francisco: Jossey-Bass.

U. S. Department of Education (2003). *Title I-Improving the Academic Achievement of the Disadvantaged; Final Rule*, 68 Fed. Reg. 236 (December 9, 2003).

U. S. Department of Education (2004). *Standards and assessment peer review guidance: Information and examples for meeting the requirements of the No Child Left Behind Act of 2001*. Washington, DC: Author.

Wehmeyer, M. L., Lattin, D., & Agran, M. (2001). Achieving access to the general curriculum for students with mental retardation: A curriculum decision-making model. *Education and Training in Mental Retardation and Developmental Disabilities*, *36*, 327-34

Abstract

This White Paper offers a framework for including students with disabilities in large-scale assessment and accountability systems. Responding to IDEA and No Child Left Behind requirements, it distinguishes participation through general assessments, accommodations, and alternate assessments judged against grade-level, modified, or alternate achievement standards. The authors frame validity as an evidence-based argument linking test scores to interpretations and decisions, emphasizing alignment with grade-level content standards and careful attention to breadth, depth, complexity, knowledge, and skills. Guidance addresses assessment design, item quality, scoring, reliability, validity evidence, reporting, accountability uses, consequences, and professional development. The paper stresses that IEP teams should make individualized, annually reconsidered participation decisions based on student needs rather than disability labels or placement. It calls for transparent technical documentation, ongoing evaluation of intended and unintended consequences, and collaboration among measurement specialists, educators, administrators, families, and policymakers to ensure equitable access, defensible score interpretations, and educationally beneficial assessment practices.