DOCUMENT RESUME

ED 473 981 EC 309 762

AUTHOR	Quenemoen, Rachel; Thompson, Sandra; Thurlow, Martha
TITLE	Measuring Academic Achievement of Students with Significant Cognitive Disabilities: Building Understanding of Alternate Assessment Scoring Criteria. Synthesis Report.
INSTITUTION	National Center on Educational Outcomes, Minneapolis, MN.; Council of Chief State School Officers, Washington, DC.; National Association of State Directors of Special Education, Alexandria, VA.
SPONS AGENCY	Special Education Programs (ED/OSERS), Washington, DC.
REPORT NO	NCEO-50
PUB DATE	2003-06-00
NOTE	64p.
CONTRACT	H326G000001
AVAILABLE FROM	National Center on Educational Outcomes, University of Minnesota, 350 Elliott Hall, 75 East River Rd., Minneapolis, MN 55455 ($20). Tel: 612-624-8561; Fax: 612-624-0879; Web site: http://education.umn.edu/NCEO. For full text: http://education.umn.edu/NCEO/OnlinePubs/Synthesis50.html.
PUB TYPE	Reports - Evaluative (142)
EDRS PRICE	EDRS Price MF01/PC03 Plus Postage.
DESCRIPTORS	*Academic Achievement; Elementary Secondary Education; Evaluation Methods; Individualized Education Programs; *Mental Retardation; *Scoring; *Student Evaluation; Test Interpretation
IDENTIFIERS	Arkansas; Kentucky; Louisiana; Oregon; Vermont

ABSTRACT

This report compares the assumptions and values embedded in scoring criteria used in five states (Arkansas, Kentucky, Louisiana, Oregon, and Vermont) for alternate assessments of students with significant cognitive disabilities. The five states use different alternate assessment approaches, including portfolio assessment, performance assessment, Individualized Education Program (IEP) linked body of evidence, and traditional test formats. Analysis finds a surprising degree of commonality in how the states define success for these students. Six criteria are included the states' approaches, either articulated or assumed. They include: (1) content standards linkage; (2) independence; (3) generalization; (4) appropriateness; (5) IEP linkage; and (6) performance. Three scoring criteria are very different across the five states' approaches, including system vs. student emphasis, mastery, and progress. The report offers 10 recommendations for states including identify stated and embedded scoring criteria; clarify whether scoring criteria refer to the student or the system; determine whether underlying assumptions of scoring criteria reflect views of various stakeholders; examine the scoring process in light of the scoring criteria; and examine scores of alternate assessments to ensure that scores reflect original intentions. Appended are an interview guide and student and system criteria definitions and examples. (Contains 19 references and 17 tables.) (DB)

Synthesis Report 50

Measuring Academic Achievement of Students with Significant Cognitive Disabilities: Building Understanding of Alternate Assessment Scoring Criteria

NATIONAL CENTER ON EDUCATIONAL OUTCOMES

In collaboration with:

Council of Chief State School Officers (CCSSO)

National Association of State Directors of Special Education (NASDSE)

Supported by:

U.S. Office of Special Education Programs

Synthesis Report 50

Measuring Academic Achievement of Students with Significant Cognitive Disabilities: Building Understanding of Alternate Assessment Scoring Criteria

Rachel Quenemoen • Sandra Thompson • Martha Thurlow

June 2003

NATIONAL
CENTER ON
EDUCATIONAL
OUTCOMES

The Center is supported through a Cooperative Agreement (#H326G000001) with the Research to Practice Division, Office of Special Education Programs, U.S. Department of Education. The Center is affiliated with the Institute on Community Integration at the College of Education and Human Development, University of Minnesota. Opinions expressed herein do not necessarily reflect those of the U.S. Department of Education or Offices within it.

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

NCEO Core Staff

Deb A. Albus
Ann T. Clapper
Jane L. Krentz
Kristi K. Liu
Jane E. Minnema
Ross Moen
Michael L. Moore
Rachel F. Quenemoen
Dorene L. Scott
Sandra J. Thompson
Martha L. Thurlow, Director

Additional copies of this document may be ordered for $20.00 from:

National Center on Educational Outcomes
University of Minnesota • 350 Elliott Hall
75 East River Road • Minneapolis, MN 55455
Phone 612/624-8561 • Fax 612/624-0879
http://education.umn.edu/NCEO

This document is available in alternative formats upon request.

Executive Summary

In this report, we compare and contrast the assumptions and values embedded in scoring criteria used in five states for their alternate assessments. We discuss how the selected states are addressing the challenge of defining successful outcomes for students with significant disabilities as reflected in state criteria for scoring alternate assessment responses or evidence and how these definitions of successful outcomes have been refined over time. The five states use different alternate assessment approaches, including portfolio assessment, performance assessment, IEP linked body of evidence, and traditional test formats. There is a great deal of overlap across the alternate assessment approaches, and they tend to represent a continuum of approaches as opposed to discrete categories.

On surface examination, the scoring criteria used by the five selected states appear to be different from one another. State responses to the 2001 survey of state directors of special education (Thompson & Thurlow, 2001) suggested significant differences as well. Yet when the scoring elaborations and processes are examined closely, many similarities emerge. After careful analysis of how some assumptions are built into the instrument development or training processes, even more similarities emerge. The definitions and examples and the side by side examination of the criteria, the scoring elaborations, and the assumed criteria in the design of training materials and assessment format yield a surprising degree of commonality in the way these states define success for students with significant cognitive disabilities. Six criteria are included in all of the five states' approaches in some way, either articulated or assumed. They include "content standards linkage," "independence," "generalization," "appropriateness," "IEP linkage," and "performance." Three scoring criteria are very different across the five states' approaches. They include "system vs. student emphasis," "mastery," and "progress." Two states also use additional criteria that no other state uses.

The common criteria and differences in criteria across the five states described here do not reflect "right" or "wrong" approaches. Each of these five states has developed alternate assessment scoring criteria that reflect their best understanding of what successful outcomes are for students with significant cognitive disabilities. Since alternate assessments are a much more recent development than general assessments, the assumptions underlying them are still debated and discussed. These deliberations can sometimes feel frustrating for teachers, policymakers, and other stakeholders, but they are an important activity that helps a state develop an alternate assessment that reflects its educational values for students with significant cognitive disabilities.

We hope that many state personnel will have opportunities to read this report and will be prompted to carefully examine their own criteria, asking what each criterion means and how it is important to each student's success. Alternate assessments are still very new, and taking the time for thoughtful reexamination is critical. We provide questions and recommendations you may find helpful as you work to ensure your alternate assessment reflects your state's best understanding of successful outcomes for students with significant cognitive disabilities.

Table of Contents

Overview

> *To the Reader*: Alternate assessments have evolved dramatically over the past six years since they were first required by the 1997 amendments to the Individuals with Disabilities Education Act. This report focuses on the evolving definitions of appropriate scoring criteria for these assessments. It is targeted to readers who are familiar with the basic approaches and methods of alternate assessment, but who are working toward increased understanding of the unique challenges of defining achievement for students with significant cognitive disabilities, one of the populations for whom alternate assessments may be appropriate. For readers who are not familiar with the recent history and methods of alternate assessments, we recommend as background the Web site http://education.umn.edu/nceo and the specific pages on alternate assessment found on the left side menu on that site. An overview of alternate assessment and frequently asked questions can be found there, as well as links to numerous articles and resources on alternate assessment. The 2001 No Child Left Behind Act (NCLB) has implications for the development of alternate assessments, and it is possible that this law will result in reinterpretations of policies and best practices for alternate assessment. This paper should be viewed as a study of current state practices and not an interpretation of the requirements of NCLB.

Over the past century, the educational community and the public have become familiar with the tools of the educational achievement test and measurement trade. Almost every adult in our society has personal experience with standardized achievement testing, for better or worse. We "know" what public school testing looks like: a group of students sits in the school cafeteria or gymnasium, nervously clutching a supply of number 2 lead pencils, well briefed on techniques to fully and completely mark the bubble sheets, waiting for instructions to turn to the first page of the test booklet. In those test booklets are test items in a variety of formats, for example, multiple-choice, true and false, short answer, or essay items.

What many educators and the public do not know is why or how those particular items–and the blend of item formats–are chosen. By the time the test booklet and bubble sheet are in the hands of the test takers, a great deal of thought and decision-making has taken place, but that work is invisible to the test taker. Typically, for tests currently required by states for the large-scale assessment of student achievement, the test is the end product of a development and field test process managed by a test company and state personnel, working collaboratively with assessment and curriculum specialists. One decision they have made is the test's content specifications. These specifications define what knowledge and skills are tested and how students should demonstrate their ability to use the knowledge and skills. For example, in mathematics, decisions are made about how many items test basic computation skills, how many items test problem-solving abilities, and how many items test other important mathematics constructs. Decisions are made about the depth and breadth of coverage of the content, and the degree of cognitive complexity of the items. Similarly, decisions are made about how student responses are scored. For example, each item may have one correct answer, and that correct answer gets one point. Points may be deducted for incorrect answers, or scoring criteria may allow points for "partially correct" in some way. The "number correct/incorrect" on these items gives the information needed to identify which students have achieved important knowledge and skills used in school and in the future. For short answer or essay items, criteria are developed to score the open-ended student responses, reflecting the knowledge and skills that all children should know and be able to do. All in all, it takes multiple years to develop a test that meets basic standards of test quality.

Thus, once test booklets are in the hands of the students, test items reflect educator and measurement expert understanding of how to measure successful student outcomes, and this understanding ultimately is reflected in achievement scores. This understanding is assumed within the items and the way student responses are scored. Because of the years of experience test developers have had with this kind of achievement testing, these underlying assumptions are rarely discussed or debated publicly, and are invisible to all but those most closely associated with the testing process.

Some adults in our society have not had direct experience with achievement tests like those described above. Until the past decade, many students with disabilities were exempted from participation in large-scale assessments. Although exemption practices varied across the states, students with the most significant cognitive disabilities were almost always exempted. Until recently, very few educators and almost no large-scale measurement experts or test companies had direct experience in defining what successful outcomes are for this very small population of students, at least as these outcomes are defined through measurement of academic achievement. As a result, there is not a century of educator and measurement expert understanding and consensus about how to construct large-scale assessments for students with significant cognitive disabilities, or how this understanding ultimately can be reflected in academic achievement scores.

That situation is changing, however. Over the past decade, state, district, and school staff have become familiar with federal requirements that students with disabilities participate in state and district assessment systems, and that assessment accommodations and alternate assessments be provided for students who need them. The Improving America's Schools Act of 1994 (IASA), the Individuals with Disabilities Education Act reauthorization of 1997 (IDEA), and the No Child Left Behind Act of 2001 (NCLB) all contributed to the trend toward including students with disabilities in assessment. IDEA 1997 first identified alternate assessment as an option for students who cannot participate in general assessments even with accommodations. At the end of the 20th century many states had developed alternate assessments to meet that requirement (Thompson & Thurlow, 2001), although many were still uncertain as to how they would incorporate the results of alternate assessments into accountability formulas (Quenemoen, Rigney, & Thurlow, 2002).

NCLB raises the stakes of these accountability efforts and requires that states must specify annual objectives to measure progress of schools and districts to ensure that all groups of students reach proficiency within 12 years. Provisions included in the proposed regulations for NCLB would allow the use of alternate achievement standards for students with the most significant cognitive disabilities, provided that does not exceed 1.0 percent of all students; and that these standards are aligned with the State's academic content standards and reflect professional judgment of the highest learning standards possible for those students. These provisions were included in the proposed NCLB regulations published in March 2003. In the context of these new accountability requirements, and despite the current uncertainty over the ultimate regulations to be applied to standards for alternate assessments, states are developing scoring criteria that reflect assumptions about what successful outcomes are for students with significant cognitive disabilities in order to ensure that achievement measures reflect these values and assumptions. It is possible that some state alternate assessments, including some described in this report may need to be modified as NCLB requirements are implemented and clarified.

Purpose of This Report

In this report, we discuss how selected states are addressing the challenge of defining successful outcomes for students with significant disabilities as reflected in state criteria for scoring alternate assessment responses or evidence. We describe how these definitions of successful outcomes have been refined over time to reflect growing understanding of the highest learning standards possible for these students. We articulate the underlying values embedded in alternate assessment scoring criteria used by these states, and identify the common ground and the differences that exist across the states as reflected by these values.

Status of Research-based Understanding

States must take several steps in the development of alternate assessment to ensure that achievement standards can be set that truly reflect what the highest learning standards possible are for this small population of students. Numerous researchers are examining the methods used in states to extend or expand the state content standards for the purpose of aligning alternate assessments to the same academic content as the general assessments (Browder, 2001; Browder, Flowers, Ahlgrim-Delzell, Karvonen, Spooner, & Algozzine, 2002; Kleinert & Kearns, 2001; Thompson, Quenemoen, Thurlow, & Ysseldyke, 2001). Others are exploring standard-setting approaches that can be used for alternate assessment in order to define what "proficient" means for accountability purposes (Olson, Mead, & Payne, 2002; Roeber, 2002; Weiner, 2002). In order for extended academic content and achievement standards to make sense for this population, additional work must be done on the criteria used to score alternate assessment responses or evidence. These criteria-setting efforts can build on early investigations that attempted to identify common understanding of what successful outcomes are for students with significant cognitive disabilities (Kleinert & Kearns, 1999; Ysseldyke & Olsen, 1997).

Following the 1997 reauthorization of IDEA, Ysseldyke and Olsen (1997) addressed assumptions that drive alternate assessments and identified four recommendations that shaped early state criteria decisions. They suggested that alternate assessments:

1. Focus on authentic skills and on assessing experiences in community/real life environments.

2. Measure integrated skills across domains.

3. Use continuous documentation methods if at all possible.

4. Include, as critical criteria, the extent to which the system provides the needed supports and adaptations and trains the student to use them.

Kleinert and Kearns (1999) surveyed national experts in the education of students with significant disabilities and found the highest ranked indicators of successful outcomes were integrated environments, functionality, age appropriateness, and choice-making. However, many respondents raised questions about the appropriateness of a focus on functional domains in an era of standards-based reform for all students, and the requirement in the 1997 reauthorization of IDEA that students should have access to, participate in, and make progress in the general curriculum. A discussion ensued among researchers and practitioners about the relationship of general academic content areas and functional outcomes. Thompson and Thurlow (2001) found that by the turn of the century, a shift to standards-based alternate assessment measurement approaches and a focus on the general curriculum was adopted by most states, reinforced by regulations and guidance on federal assessment policy.

Yet, the Thompson and Thurlow 2001 survey of state assessment practices found a continued range of alternate assessment approaches (see Table 1 for a summary description of the alternate assessment approaches we address in this report), and more importantly, no clear consensus on the criteria being used to score alternate assessment evidence. The survey also illustrated the complexities of definitions, both for the approaches themselves, as well as for the criteria used for scoring. These complexities can confuse any discussion attempting to compare and contrast varying approaches in a categorical way. In the interests of sorting out these complexities, the reader may wish to carefully review the discussion of the overlap among approaches below, and refer to this later as we sort through variations in our sample states. Table 1 (Definitions of Alternate Assessment Approaches Discussed in this Paper) and Table 2 (Common Scoring

Criteria Terminology) are provided as a quick glance summary to refer to later in this report, as well as to provide a summary of the discussions below.

Alternate Assessment Approaches: Not Mutually Exclusive Categories

In general, the alternate assessment approaches defined in Table 1 go from a basic methodology of student-by-student individually structured tasks (portfolio assessment) to highly structured common items or tasks completed by every student (traditional tests) as you read down the table. These approaches are not mutually exclusive categories, and as state practices are examined, it is clear that a great deal of overlap in methods occurs.

Portfolio Overlap with IEP Linked Body of Evidence

The "portfolio" approach typically requires the gathering of extensive samples of actual student work or other documentation of student achievement. Very often, this work is in response to teacher-developed tasks with teacher-defined linkage to content standards, thus the evidence varies dramatically from one student to the next. It is the standardized application of scoring criteria to the varied evidence that results in comparability across students. "IEP linked body of evidence" approaches as defined here also may require extensive sampling of work, have similar scoring criteria, and apply them in similar ways to the portfolio approach. However, in this report, the states using a portfolio approach require extensive student products; the state that uses an IEP linked body of evidence has more focused evidence requirements, related specifically to the skills and knowledge defined in the student's IEP, and the documentation of progress in the IEP process. In general, the distinguishing characteristics between "portfolio" approaches versus "body of evidence" approaches tend to be, for the purpose of this report:

(1) the amount of evidence required is more for portfolio, less for body of evidence;

(2) the degree of state provided definition of what specific content is measured is less with portfolios, and there is more state provided definition of specific content for a body of evidence; and

(3) the degree of IEP linkage is less for portfolio and more for a body of evidence.

(A complicating variable is how advanced a state is in implementing standards-based IEP planning, thus the IEP linkage to alternate assessments may be "pushing the envelope" of standards-based reform for students with disabilities. That discussion is beyond the purpose of this report, but will be increasingly important as alternate assessment evolves.)

Table 1. Definitions of Alternate Assessment Approaches Discussed in this Paper

Portfolio: A collection of student work gathered to demonstrate student performance on specific skills and knowledge, generally linked to state content standards. Portfolio contents are individualized, and may include wide ranging samples of student learning, including but not limited to actual student work, observations recorded by multiple persons on multiple occasions, test results, record reviews, or even video or audio records of student performance. The portfolio contents are scored according to predefined scoring criteria, usually through application of a scoring rubric to the varying samples of student work.

IEP Linked Body of Evidence: Similar to a portfolio approach, this is a collection of student work demonstrating student achievement on standards-based IEP goals and objectives measured against predetermined scoring criteria. This approach is similar to portfolio assessment, but may contain more focused or fewer pieces of evidence given there is generally additional IEP documentation to support scoring processes. This evidence may meet dual purposes of documentation of IEP progress and the purpose of assessment.

Performance Assessment: Direct measures of student skills or knowledge, usually in a one-on-one assessment. These can be highly structured, requiring a teacher or test administrator to give students specific items or tasks similar to pencil/paper traditional tests, or it can be a more flexible item or task that can be adjusted based on student needs. For example, the teacher and the student may work through an assessment that uses manipulatives, and the teacher observes whether the student is able to perform the assigned tasks. Generally the performance assessments used with students with significant cognitive disabilities are scored on the level of independence the student requires to respond and on the student's ability to generalize the skills, and not simply on accuracy of response. Thus, a scoring rubric is generally used to score responses similar to portfolio or body of evidence scoring.

Checklist: Lists of skills, reviewed by persons familiar with a student who observe or recall whether students are able to perform the skills and to what level. Scores reported are usually the number of skills that the student is able to successfully perform, and settings and purposes where the skill was observed.

Traditional (pencil/paper or computer) test: Traditionally constructed items requiring student responses, typically with a correct and incorrect forced-choice answer format. These can be completed independently by groups of students with teacher supervision, or they can be administered in one-on-one assessments with teacher recording of answers.

Adapted from Roeber, 2002.

Body of Evidence Overlap with Performance Assessment

The "body of evidence" tendency toward more focused and specific evidence in turn reflects a similarity with the least structured specific "performance assessment" approaches in other states. That is, some performance assessment approaches define the skills and knowledge that must be assessed for each student, but they still allow the test administrator to structure a task that the student will perform to demonstrate the skills and knowledge. The most structured body of evidence approaches tend to be very similar to the least structured performance assessments. In other words, a state may require in a performance assessment OR a body of evidence that a student demonstrate his or her reading process skills by identifying facts, events, or people involved in a story. How the student demonstrates those skills will vary, and the task could involve, for example:

- requiring that a student use a switch to provide different sound effects corresponding to characters in a story, whether read by the student or teacher;
- having a student look at pictures to identify favorite and least favorite parts of a story that was read aloud; or
- a student reading a simple story and then making predictions of what will happen next using clues identified in the text.

As a further source of individualized tailoring in either a highly structured body of evidence or a loosely structured performance assessment, each of these tasks could allow for varying levels of teacher prompting, and thus scoring criteria could include the criterion of the level of prompting/degree of independence. Where the approaches differ is that a body of evidence approach generally requires submission of the student evidence for scoring; a performance assessment approach typically involves the test administrator or teacher scoring student work as it occurs.

Other states that define their approach as a performance assessment provide a high degree of structure and specifically define the task (e.g., having a student look at pictures to identify favorite and least favorite parts of a story that was read aloud, with provided story cards and materials). Yet they typically allow variation in the degree of prompting (ranging from physical prompts to fully independent responses), or in the methods of student responses (from use of picture cards vs. verbal response for example). Even states that use common performance assessment tasks for their required alternate assessment for students with significant disabilities tend to use multiple scoring criteria more similar to portfolio or body of evidence approaches, as compared to simple recording of correct or incorrect responses used in checklist or traditional test formats.

Performance Assessments Overlap with Checklist and with Traditional Test Formats

Most "checklist" approaches ask the reviewer to record whether a student has demonstrated the skill or knowledge. These may include a judgment on degree of independence or generalization as well as accuracy of skill performance, but the judgments may simply reflect accuracy. The difference between checklists and performance assessment approaches where the test administrator scores the performance is that the checklist approach relies on recall and not on actual on-demand student performance. By contrast, "traditional test" formats require the on-demand performance of skills and knowledge, on a specified item, with built in connection to content standards and with accuracy (or "right/wrong") as the primary criterion. The test administrator records student performance as right or wrong, and no further scoring is necessary. This approach is the most similar to the testing approaches most adults have experienced described in the opening section of this report.

The Complexities of Criteria Definitions Across the States

As shown in the descriptions above, there is a great deal of overlap across the alternate assessment approaches, and they tend to represent a continuum of approaches as opposed to discrete categories. Regardless of what approach a state has chosen, or what they have called it, states have begun to revisit the basic assumptions and values underlying their approach to alternate assessment. Central to this effort is the challenge of clarifying what the criteria are on which evidence is judged. These criteria should reflect how state stakeholders define successful outcomes for students with significant cognitive disabilities. For example, a description of a successful outcome for these students may include the ability to do particular skills, with the highest degree of independence and self-choice possible, in a variety of settings or for different purposes, with an ability to get along with others and maintain relationships. Scoring criteria typically are constructed to be sensitive to measuring learning toward some or all of these desired outcomes.

An in-depth study that compares and contrasts approaches across a variety of states and people holding a variety of perspectives naturally runs into terminology challenges; therefore, we define several terms that we use throughout the report. In Table 2 we provide common scoring criteria terminology. The clarification of terms used in describing scoring criteria is even more complex than the clarification of the terms used to describe the alternate assessment approaches, if that is possible.

A critical set of definitions involves the distinctions among student criteria, system criteria, and combination student and system criteria.

Student Criteria vs. System Criteria vs. Combination Criteria

Scoring criteria typically are developed by stakeholders in each state based on professional and research-derived understanding of desired and valued outcomes for this population of students. These can be a direct measure of student achievement (student criteria); they may reflect necessary system conditions essential for student success (system criteria); or they can be a combination of student achievement seen within the context of system provided supports (combination). The distinctions among student, system, and combination criteria are important

Table 2. Common Scoring Criteria Terminology

Scoring Rubrics: Commonly used in scoring alternate assessment evidence or responses, scoring rubrics list the criteria that are desired in student work, along with definitions of quality expectations, generally along a 3-5 point scale.

Scoring Criteria: These are specific definitions of what a score means and how specific student responses are to be evaluated. Generally based on stakeholder and research derived knowledge of what is considered best practice in teaching and learning for these students, states identify scoring criteria to encourage these best practices. These criteria can be articulated in scoring processes, or built into test specifications. Criteria for alternate assessment typically fall into one of three categories: student criteria, system criteria, and a student/system blend.

> **Student Criteria:** Some scoring criteria reflect actual student performance, and may include a quantitative measure of student accuracy on items reflecting desired knowledge or skills, or a qualitative judgment about the overall level of performance of the student in terms of skills and competence, degree of progress, independence, or ability to generalize.
>
> **System Criteria:** Other scoring criteria include system performance, such as whether students are provided instruction in multiple settings, whether they are provided opportunities to plan, monitor, and evaluate their work, whether they work with non-disabled peers, and whether they are provided with appropriate human and technological supports.
>
> **Student/System Blend Criteria:** These system criteria can also be expressed as student criteria: whether the student can demonstrate the skill in multiple settings; plan, monitor, and evaluate his or her work; work with non-disabled peers; and work independently (using appropriate human and technological supports).

Adapted from Roeber, 2002.

to this study; the reader may wish to continue to refer to Table 2 for clarification as we examine criteria in each of the sample states.

Scoring Process

After defining the criteria, the scoring process is designed to reflect and capture the identified outcomes as evidenced in a sample of student work or in student responses to test items. For example, a state may choose to score student work using the two criteria they have identified as central to successful outcomes for students with significant cognitive disabilities: demonstration of specific knowledge or skills, and generalization of knowledge and skills to more than one setting or for more than one purpose. They can then decide to apply the criteria – to actually score the evidence – choosing from several scoring processes. For example, a state could apply the skill and generalization criteria through one of these processes:

1. A process that requires that student work be submitted to scoring centers, where evidence is scored by trained double blind scorers and a third tie-breaker scorer.

2. Scoring of student performance by trained test administrators observing as the student performs the skill or responds to the performance task in more than one setting or for more than one purpose.

3. A two-step process where the skill level is determined by "number correct" on a pencil/paper test, accompanied by a checklist form completed by teacher, parent, and related service provider reflecting their judgment of degree of skill generalization.

Thus, defining scoring criteria and defining the scoring process are different, yet interdependent. Setting scoring criteria has more importance in defining and detecting learning toward successful outcomes than does the scoring process; you can change scoring processes easily as you understand how processes can be improved (e.g., changing from one scorer to two scorers). Scoring criteria generally change only after serious discussion and open debate about the assumptions and values underlying each criterion.

The Relationship Between Assessment Approach and Scoring Criteria

The assumptions and values underlying scoring criteria can remain the same across alternate assessment approaches. For example, assumptions and values for scoring criteria can be the same for a body of evidence or for a performance assessment or for a pencil and paper assessment, as assumed in the example above. In traditional large-scale assessments, which generally use pencil and paper forced-choice formats, scoring criteria often are embedded in the test development processes and are not articulated in the scoring process. That is, criteria such as alignment to content standards, degree of difficulty, and appropriateness for grade level, are addressed in the test's item development, item specifications, form development, and in field testing. For performance assessments, and sometimes for the constructed response items in traditional tests, scoring criteria generally are articulated, often in the form of rubrics. Scoring rubrics are the most common method of applying scoring criteria to portfolios and bodies of evidence.

In this study, we compare and contrast the assumptions and values embedded in scoring criteria used in five states for their alternate assessments. Given the wide range of approaches included in the states we studied, we attempted also to identify the "assumed" criteria underlying more traditional test formats, as well as to identify similar kinds of assumptions underlying more performance based approaches.

Scoring Criteria vs. Achievement Level Descriptors

One final distinction should be made between the scoring criteria that are the focus of this report and another common term used in large-scale assessment, the term "achievement level descriptors." Achievement level descriptors are generally the terms used to communicate how a state has defined "proficiency" based on test results. In the scoring process, scoring criteria are applied to responses or evidence to produce a score. This score is not in itself directly translated into an achievement level. A standard-setting process must be defined to identify what scores mean. Usually this involves identifying "cut scores" that separate achievement levels. The first cut scores identified are those based on an understanding of *proficient* student work. Different cuts determine gradations of quality, and result in defining a range of scores reflecting varying success. It is through this standard setting process that a state defines what scores represent "proficiency." The scoring criteria, of course, are related to the achievement levels and descriptors, but they are not the same. In this study we specifically focus on the way states have defined and are refining their scoring criteria.

Methods

Selecting States

In order to select a small number of states to study, we went through several steps. Our goal was to select states that represented different approaches and scoring criteria. Given the definitional overlaps we have described in the preceding sections, we found that grouping states into categories was not a simple process. First, we examined the results of a 2001 survey of state directors of special education that included several questions about alternate assessments (Thompson & Thurlow, 2001). All states were sorted by their responses to questions about:

a. Alternate assessment approach

b. Scoring criteria (called "performance measures" in that survey)

We identified patterns of responses on these two items, and identified "clusters" of responses to each of the two questions. From these clusters, we selected a small number of states that represented variations in the type of approach used for alternate assessment and variations in the criteria used to score alternate assessment responses or evidence.

The distribution in variations in approach identified in the 2001 survey (Thompson & Thurlow, 2001) is shown in Table 3. Because nearly half of the states reported using a portfolio or body of evidence approach (as defined in the 2001 survey these two approaches were grouped), we decided to select two of these states for the study, but ones that differed on scoring criteria. Then we planned to select one each from the other categories that had fewer states. However, we realized on closer examination that some states that had reported in portfolio/body of evidence category could also be considered as part of the IEP approaches, if the IEP linkage defined the body of evidence. That realization required a revisiting the original clusters, and regrouping based on information we had available in our files from Web sites and from the states. Upon review, we realized that the approach reported by states on the 2001 survey was not always the most accurate label, and we adjusted our clusters of state approaches to match what we found in state descriptions provided in public forums, not simply what was reported on the survey.

The distribution in variations in scoring criteria identified in the 2001 survey fell into two categories: (1) criteria focused on student performance, and (2) criteria focused on system performance. These are shown in Table 4.

Table 3. Alternate Assessment Approaches Across States

Body of evidence/ portfolio[a]	24 states
Checklist	9 states
IEP team determines strategy	4 states
IEP analysis[b]	3 states
Combination of strategies	4 states
Specific performance assessment	4 states
No decision	2 states

Based on the 2001 survey of state directors of special education, conducted by the National Center on Educational Outcomes (Thompson & Thurlow, 2001).

[a] Defined by the submission of student work or other documentation.

[b] Included IEP linked body of evidence, checklists, or IEP team determination of progress.

Table 4. Student and System Scoring Criteria Across States

Student Scoring Criteria	
Skill/competence	40 states
Independence	32 states
Progress	24 states
Ability to generalize	18 states
Other	7 states
System Scoring Criteria	
Variety of settings	21 states
Staff support	20 states
Appropriateness (e.g., age, challenge)	20 states
Gen. ed. participation	12 states
Parent satisfaction	9 states
No system measures	8 states

Based on the 2001 survey of state directors of special education, conducted by the National Center on Educational Outcomes (Thompson & Thurlow, 2001).

Again, several factors contributed to reorganizing the states we considered for our sample. From correspondence with states as part of NCEO's technical assistance activities, we were aware that several states were in the process of changing their assessment approach or their scoring criteria. For example, several states that had reported IEP based approaches in 2001 have since indicated to us that they are in the process of developing a different approach. In addition, several other states had indicated to us that they anticipated changes in their scoring criteria after their first year of implementation and were undecided as to how they would proceed. In order to ensure that we would be analyzing thoughtfully conceived and fairly stable state approaches, we used a purposive sampling technique informed by two indicators. One indicator of stability was that the state had presented its approach and scoring criteria in public forums such as the annual alternate assessment pre-sessions to the CCSSO large-scale conferences, or at the Assessing Special Education Students (ASES) State Collaborative on Assessment and Student Standards (SCASS) meetings. The second indicator was that the state had at least two years of statewide experience with the alternate assessment in the current form, with no substantial changes.

Based on review of materials provided by the states at these public forums, we found that within the portfolio group the approaches to scoring criteria varied depending in part on what test company had supported development of the assessment. For that reason, we decided to select the two states representing portfolio approaches to ensure that two different test company approaches would be represented. We also looked for states that generally conformed to the current understanding of quality assessment approaches as reflected in the NCEO *Principles and Characteristics of Inclusive Assessment and Accountability Systems* (Thurlow, Quenemoen, Thompson, & Lehr, 2001).

This purposive and somewhat subjective sampling process resulted in a much smaller group of states from which to choose. From this group, we chose two portfolio approach states that differed in scoring criteria (and testing company), and one state each representing performance assessment, combination and checklist approaches, and an IEP approach, which represented the IEP linked body of evidence subcluster. The final decision to include particular states was made in part to ensure that the widest range of criteria definitions was represented, and that the states chosen represented the widest possible range of test companies/developers, while still ensuring the basic quality of the approach. We requested and received agreement from the five states selected to participate in the study. The states were Arkansas (portfolio assessment), Kentucky (portfolio assessment), Louisiana (performance assessment), Oregon (combination/checklist), and Vermont (IEP linked body of evidence).

Data Collection Process

The study included both a document review and interview process. The purpose of the document review was to describe each state's approach to alternate assessment, and to identify and report the scoring criteria used to score assessment responses or evidence in each state. The specific terms used by states for this process varied (e.g., scoring rubrics, scoring criteria, scoring domains, scoring rules, or simply number correct for pencil/paper tests). The focus of the interviews was on how each state developed its scoring criteria, definitions, and assumptions. We also asked states how and why these criteria have changed over time, and finally, how states see the criteria being used to improve the education of students with significant disabilities. The interview guide is included in Appendix A.

After an NCEO staff member conducted the document review, written summaries were sent to each state contact person for review, correction, and verification. The same staff member conducted interviews by phone with all of the state contacts. A second staff member reviewed interview recordings and supporting documentation, and developed written anecdotal case studies. The final draft materials were subject to review and comment by the states for accuracy and additional insight.

A final step was taken to compare and contrast the documented and anecdotal understanding of the assumptions each state reflected in its scoring criteria. Through a "side by side" view of the states' scoring criteria (obtained from both written documentation and phone interviews), we proposed common themes that represent the underlying values embedded in each state's scoring criteria and processes. We also attempted to delineate the common ground that exists across the states.

Scoring Criteria From the Selected Sample of States

The approaches and criteria reported in the 2001 state survey for the five states included in this study are shown in Table 5.

After document reviews and interviews, we found the actual scoring criteria used by each state varied somewhat from the categories provided in the 2001 survey. We present here the criteria as presented in state documentation, followed by a more detailed description of each state's articulated and assumed scoring criteria based on documentation and interviews.

Since we are focusing on the complexities of scoring criteria in these descriptions, we provide very few details on the general methods of the state's overall approach. This assumes that the reader has a basic understanding of varied alternate assessment approaches, including portfolios, bodies of evidence, performance assessments, and checklists or traditional test formats. We recommend reviewing the alternate assessment pages of the NCEO Web site at http://education.umn.edu/nceo or visiting each of our sample state's Web sites for more information on the basic approaches. We have made a judgment and have confirmed through our

Table 5. Approaches and Criteria Reported on NCEO's 2001 Survey by the Five Selected States

State	Approach	Student Scoring Criteria	System Scoring Criteria reported on 2001 Survey
Arkansas	Portfolio	a, c	b, c
Kentucky	Portfolio	a, b, c, d	a, b, c, d, e
Louisiana	Performance Assessment	a, c, d	None
Oregon	Combination: Checklist, performance assessment, pencil/paper test	e (number correct on pencil/paper exam)	b, c (on performance assessment)
Vermont	Evidence/IEP linkage	a, b	Other (not described)

NOTE: The approaches and criteria in this table are AS REPORTED in 2001 and differ from descriptions in current analyses in Tables 15-19.

Student Scoring Criteria: **a**=skill/competence level; **b**=degree of progress; **c**=level of independence; **d**=ability to generalize; **e**=other.

System Scoring Criteria: **a**=staff support; **b**=variety of settings; **c**=appropriateness (age appropriate, challenging, authentic); **d**=parent satisfaction; **e**=participation in general education. **None** = No system criteria.

Based on the 2001 survey of state directors of special education, conducted by the National Center on Educational Outcomes (Thompson & Thurlow, 2001).

document review and interviews that these five states share three basic quality indicators. First, each state developed its alternate assessment through an open process involving varied state stakeholders, including teachers, parents, researchers, and technical advisors. Thus, each reflects professional and research based understanding of the best possible outcomes for students with significant cognitive disabilities. Second, each of these states continues to work to understand and document the technical adequacy of its approach, including documenting and improving the reliability and validity of their approach. Third, each of these states can articulate a coherent alignment of its basic assumptions underlying teaching and learning for these students, and the approach and criteria it has chosen to assess that learning. Although the states differ from one another, each reflects a thoughtful approach. This does not mean that these sample states have "perfect" or even "approvable" alternate assessments, or that they will continue far into the future with the methods described here. What it does mean is that we have confidence in the internal consistency and integrity of these state approaches given current understanding of best practices in alternate assessment for students with significant disabilities.

Scoring Criteria in Arkansas

Arkansas uses a portfolio assessment approach, and has a scoring rubric, domain definitions, and scoring weighting processes to define its scoring criteria. Table 6 shows Arkansas's four scoring domains and the criteria used to assign a score from 1 to 4 on each domain to each entry in a portfolio. Only the first three domains are scored for each entry; the last (settings) is scored for the entire content area. The domains are weighted differently.

Table 6. Arkansas Scoring Criteria

DOMAIN* ***see definitions below***	**Score 1**	**Score 2**	**Score 3**	**Score 4**
PERFORMANCE	Student does not perform the task with any evidence of skill	Student attempts the task, but there is only minimal evidence of skill	Student performs the task with reasonable skill	Student performs the task with mastery as demonstrated in multiple settings or on multiple occasions
APPROPRIATENESS	Task does not meet any of these criteria: age-appropriate, challenging or authentic	Task meets only 1 of these criteria: age-appropriate, challenging or authentic	Task meets 2 of these criteria: age-appropriate, challenging or authentic	Task meets all 3 of these criteria: age-appropriate, challenging and authentic
LEVEL OF ASSISTANCE	Following the introduction of the lesson or activity, student performs only with maximum physical assistance (such as hand-over-hand support)	Following the introduction of the lesson or activity, student performs with direct verbal prompting, modeling, gesturing or some physical support	Following the introduction of the lesson or activity, student performs in response to teacher-planned instructional/ social supports (e.g., peers, technology, materials supports)	Following the introduction of the lesson or activity, student performs independently OR student initiates the activity with the use of natural environmental or social supports
SETTINGS	Student performs all tasks in 1 physical setting (e.g., classroom)	Student performs tasks in 2 different settings	Student performs tasks in 3 different settings	Student performs tasks in 4 or more different settings

*** ARKANSAS DOMAIN DEFINITIONS:**

PERFORMANCE – The student's demonstration of skill while attempting a given task. Each individual task (portfolio entry) is scored for Performance.

APPROPRIATENESS – The degree to which the tasks: 1) reflect meaningful, real-world activities with age-appropriate materials, 2) provide a challenge for the student, 3) promote increased independence, and 4) are linked to the Content Standards.

LEVEL OF ASSISTANCE – Determined after the introduction of the lesson activity. The observed accommodations, adaptations, and/or assistance provided to a student during performance of tasks. Each individual task (portfolio entry) is scored for Level of Assistance.

SETTINGS – The observed settings or environments in which tasks are administered/performed. The entire portfolio (not individual entries) is scored for Settings.

Table 7 shows how the number of entries, domain weights, and points possible per entry produce the total points possible for each portfolio. Note that two scorers score each of the first three domains, and the cumulative score is used to determine the total points. It also shows the percentage of a score that is due to each domain. Again, the setting score is given collectively across all math entries and all English Language Arts entries, and not to each entry.

Table 7. Arkansas Weighting of Mathematics and English Language Arts Entries by Domain

Mathematics Entries

5 strands with 3 entries each

DOMAIN	Scorers	No. Entries	Domain Weight	Points Possible	TOTAL POINTS	Percent
Performance	2	15	4	4	480	53 1/3
Appropriateness	2	15	2	4	240	26 2/3
Level of Assistance	2	15	1	4	120	13 1/3
Settings	1	15	1	4	60	6 2/3
					900	**100%**

English Language Arts Entries

3 strands with 3 entries each

DOMAIN	Scorers	No. Entries	Domain Weight	Points Possible	TOTAL POINTS	Percent
Performance	2	9	4	4	288	53 1/3
Appropriateness	2	9	2	4	144	26 2/3
Level of Assistance	2	9	1	4	72	13 1/3
Settings	1	9	1	4	36	6 2/3
					540	**100%**

Scoring Criteria in Kentucky

Kentucky uses a portfolio assessment approach. Table 8 shows the six scoring dimensions and the scoring criteria used to assign a score of novice, apprentice, proficient, or distinguished to each dimension. When all dimensions have been scored, a holistic proficiency level is assigned to the entire portfolio.

Because Kentucky has had an alternate assessment in place several years longer than any other state, it has a system that is the result of many years of continuous improvement. In the process of implementing the Kentucky Alternate Portfolio across almost a decade, Kentucky has learned important lessons about its scoring criteria, and has revised them as a result. For example, in the early years of implementation, self-determination was not a separate dimension. First, it was part of performance; then it was part of context. Because self-determination was moved out of "performance," and then out of "context" to its own dimension, scores could not be compared across years. For the past three years no changes have been made in the criteria. Table 9 describes the lengthy process of continuous improvement that Kentucky has experienced.

Table 8. Kentucky Scoring Criteria

	NOVICE	APPRENTICE	PROFICIENT	DISTINGUISHED
STANDARDS	Portfolio shows little or no linkage to academic expectations	Portfolio shows some linkage to academic expectations	Portfolio shows linkage to most academic expectations	Portfolio shows linkage to all or nearly all academic expectations
PERFORMANCE	Student portfolio participation is passive, no clear evidence of performance of target IEP goals, products are not age-appropriate	Student performs target IEP goals meaningful to current and future environments, products are age-appropriate	Student work shows progress on target IEP goals meaningful to current and future environments in most entries, products are age-appropriate	Student work shows progress on target IEP goals meaningful to current and future environments in all entries, products are age-appropriate
SETTINGS	Student participates in limited number of settings	Student performs target IEP goals in a variety of integrated settings	Student performs target IEP goals in a wide variety of integrated settings within and across most entries	Student performance occurs in an extensive variety of integrated settings within and across all entries
SUPPORT	No clear evidence of peer supports or needed A/M/AT	Support is limited to peer tutoring, limited use of A/M/AT	Support is natural, appropriate use of A/M/AT	Support is natural, use of A/M/AT shows progress toward independence
SOCIAL RELATIONSHIPS	Student has appropriate but limited social interactions	Student has frequent, appropriate social interactions with a diverse range of peers	Student has diverse, sustained, appropriate social interactions that are reciprocal within the context of established social contacts	Student has sustained social relationships and is clearly a member of a social network of peers who choose to spend time together
SELF-DETERMINATION	Student makes limited choices in portfolio products; P/M/E of own performance is limited	Student makes choices that have minimal impact on student learning in a variety of portfolio products; P/M/E of own performance is inconsistent	Student consistently makes choices with significant impact on student learning; P/M/E of own performance is consistent	Student makes choices with significant impact on student learning within and across all entries; P/M/E of own performance is clearly evident; E is used to improve performance and/or set goals

A/M/AT = adaptation, modification and/or assistive technology
P/M/E = planning, monitoring, evaluating

Table 9. A Decade of Continuous Improvement in Kentucky

In 1990 Kentucky passed the Kentucky Education Reform Act (KERA). Part of the Act looked at assessment. Staff in the Exceptional Children Division of the Department of Education wanted to make sure that all students were accounted for. The system was designed for program, not student accountability. Many thought this would be a great opportunity for students with disabilities to take part in accountability systems where they really had not been included before.

A small group consisting of state department personnel, university personnel, teachers, parents, and local education administrators met to begin looking at the best ways to assess students with the most significant disabilities. They sent out surveys across the state and acquired a lot of information. They decided that a portfolio approach would be a good format for compiling assessment information. Once the alternate assessment was developed, this group continued to meet every summer to make revisions and clarifications based upon current literature and research findings in special education, and to design better ways to train teachers.

In order to use the portfolio for program accountability, the group reviewed research literature for components of best practice, research-based instruction. They listed five areas found critical to the education of students with the most significant disabilities. These included: settings, support, social relationships, performance, and contexts.

Kentucky's stakeholder group designed a rubric that reflected those five dimensions. They decided that each portfolio should include a collection of 7 to 10 activity-based entries selected by the IEP team's choice (however, the teacher's voice was strongest in selection decisions). These were generally not based on the general curriculum or state standards in the early years of implementation.

When IDEA was passed in 1997, Kentucky was in the process of revising its general assessment. At that time, the stakeholder group also reviewed the alternate assessment, as they had been doing annually since its initial development. In light of IDEA 97 and new revisions in the general assessment, the group considered whether changes needed to be made in the alternate assessment. IDEA focused on access to the general curriculum and, after careful consideration, the group decided that the alternate assessment portfolios were not really reflective of the general curriculum. A decision was made to change the requirements for entries from 7 to 10 activity-based entries to 5 content-based entries. This decision was made by an advisory group of teachers, with the assistance of university personnel. The entries were different at each grade level (students at grades 4, 8, and 12 were assessed). Language arts was assessed at each grade level.

The criteria and scoring rubric did not change at that point. It still contained the same five dimensions. However, some recombining took place. For example, in the performance area there were concepts that seemed like self-determination, so they were moved to the context area. Later, the context area was changed to self-determination.

Linkage to standards was addressed in the performance area, but in 1998 the stakeholder group decided it was not confident in the way adherence to standards was scored. The portfolio required standards to be addressed, and scorers looked for them, but addressing standards was not really scored. There was a connection because of the content area and focus on the general curriculum. As Kentucky worked with other states, it started seeing a tighter connection between standards and scoring criteria in other states. The stakeholder group decided that addressing standards on the portfolios needed to be scored, so this was added as a sixth scoring dimension. The best of what was found in other states was brought back to Kentucky.

Scoring Criteria in Louisiana

The complete Louisiana assessment program is called Louisiana Educational Assessment Program, or LEAP. The alternate assessment is titled LEAP Alternate Assessment (LAA). The LAA is a performance-based student assessment that evaluates student knowledge and skills on twenty target indicators for state selected Louisiana Content Standards. It is an "on-demand" assessment. The test administrator (teacher or other school staff trained in LAA) organizes activities so that the student can show what he or she knows and can do (target indicators), and then uses a rubric to score the student's performance of state specified standards-based skills. Skills for each target indicator are identified on three levels of difficulty, called Participation Levels. The levels are called Introductory, Fundamental, and Comprehensive. The participation level for each student is chosen by the test administrator, with IEP team input, based on appropriate yet high expectations for each student. Participation levels are defined in Table 10.

Table 10. Participation Levels in Louisiana

Introductory: Skills that require basic processing of information to address real-world situations that are related to the content standards, regardless of the age or grade level of the student (e.g., indicates choice when presented with two items).
Fundamental: Skills that require simple decision making to address real-world situations that are related to the content standard, regardless of the age or grade level of the student. (e.g., expresses a preference to the question "what do you want?").
Comprehensive: Skills that require higher-order thinking and complex information-processing skills that are related to the content standards, regardless of the age or grade level of the student (e.g., communicates detailed information about preferences, such as, describes activity with information about who, what, when, where, why. . .).

Table 11 shows the Louisiana alternate assessment scoring rubric, which includes two criteria, level of independence and evidence of generalizability to multiple purposes and settings. That is, the rubric measures progression from dependence to independence, and from particular skill to generalized skill.

Table 11. Louisiana Scoring Criteria

Given student performance of the standards-based skill at the appropriate performance level, the test administrator uses this rubric for scoring:

0: No performance (at introductory level only).
1: Tolerates engagement or attempts engagement.
2: Performs skill in response to a prompt.
3: Performs skill independently without a prompt.
4: Performs skill independently without prompts for different purposes OR in multiple settings.
5: Performs skill independently without prompts for different purposes AND in multiple settings.

Scoring Criteria in Oregon

Oregon uses four individually administered performance assessments: one of these is the Extended Career and Life Role Assessment System (CLRAS), the other three are Extended Reading, Extended Mathematics, and Extended Writing, three separate measures aligned to the Oregon Standards in reading/literature, mathematics, and writing. The Extended CLRAS assesses related skills in the context of "real" life routines and is administered individually as a performance assessment by a qualified assessor, usually the student's teacher. The selection of routines is informed by the results of a teacher-completed checklist while related skills are selected from the students IEP. The Extended CLRAS is aligned to the Career Related Learning Standards adopted by Oregon's State Board for all students. Extended Reading, Extended Mathematics, and Extended Writing employ curriculum based measures of emerging academics and are aligned to the general Benchmark 1 standards in the same subjects. Students may take one or more of these extended assessments; some students take only the Extended CLRAS. For extended measures in reading, mathematics, and writing, the scores are derived from the number of correct or partially correct items. For the Extended CLRAS, qualified assessors evaluate performance on related skills in the context of core steps of daily life routines according to a rating scale. There are some items, like symbol recognition, at the very lowest end of the extended assessments in reading, mathematics, and writing that seem compatible with Extended CLRAS, and there are some items at the upper end that are sensitive to growth similar to items on the general assessment.

Scoring criteria for the extended assessments in reading, mathematics, and writing are shown in Table 12. Oregon hopes to use results of the alternate assessment for predicting performance on regular assessments whenever that is appropriate. For the general assessment there are four benchmark standards: Benchmark 1, 2, and 3, and the CIM – the Certificate of Initial Mastery benchmark. The state is considering the use of a benchmark "P" for predictive or preliminary that would apply to the scores on the extended academic assessments. A "P" would indicate that sometime in the future the student may be moving to Benchmark 1, and thus into the general assessment.

Some students take only the individually administered Extended CLRAS, a performance assessment and checklist; others take the Extended CLRAS along with one or more extended assessments in reading, mathematics, and writing. A six-point scale is used to score student performance of defined "routines" (routines are selected by the qualified assessor and informed by the checklist results) on the Extended CLRAS (see Table 13). Scores of 1 to 4 reflect increasing levels of independence on the individually selected routines and skills. Two of the scores indicate the task was not done, either because it was not applicable (N) or the student could not do it at all (0). Although the scale does not include a criterion related to generalization, state staff reported in the interviews that generalization across settings is imbedded into the design

Table 12. Scoring Criteria for Oregon's Extended Reading, Extended Mathematics, and Extended Writing Assessments

Accuracy is the articulated criterion. Each individual task item is scored correct or partially correct. The scores relate to level of accuracy on letters, numbers, words, reading rate, reading accuracy, reading and listening comprehension, number concepts, computation, dictation, plus writing words, sentences, and a story.

Although accuracy is the only articulated criterion on these assessments,

- standards-based academic knowledge and skills and
- defined levels of challenge and appropriateness

are embedded into the items themselves, just as they are for the general assessment.

Given the growth model, student progress is also an assumed criterion in the Extended Reading, Extended Mathematics, and Extended Writing assessments.

of the Extended CLRAS. It is instructionally based, so students have continued instruction and are measured in areas where they are acquiring independence. As students become more independent and master environments, additional routines are selected. A high value is given to having students learn and demonstrate their related skills in natural environments, and this is reflected in scoring. Scoring of the Extended CLRAS puts emphasis on prompt fading, but if the student does seven routines, and "social greetings" is one of the related skills performed, then there are seven appropriate environments defined in which to exhibit that related skill. Thus, generalization is embedded in the scoring process.

Table 13. Scoring Criteria for Oregon's Extended CLRAS

Independence Measurement Scale

4 = completes independently
3 = completes with visual, verbal or gesture prompting
2 = completes with partial physical prompting (requires at least one physical prompt, but not continuous physical prompts)
1 = completes with full physical prompting (requires continuous physical prompts)
0 = does not complete even with physical prompting
N = not applicable (due to student's medical needs, the school environment does not provide an op portunity to perform, or the IEP team deems the routine/activity inappropriate for the student)

Scoring Criteria in Vermont

There are three alternate assessments in Vermont: an adapted form of the general assessment, a modified form of the general assessment, and a "lifeskills" alternate assessment for students with significant cognitive disabilities. The modified and adapted assessments are for students who are able to take the general assessment with specific adaptations or modifications. The students who participate in adapted assessments are generally students with sensory impairments who cannot take the regular test because of lack of Braille versions or other lack of appropriate accommodations. The adapted assessment measures the same standards, at the same proficiency levels, as the regular assessment. Modified assessments result in changes to the construct being measured, generally for students who are performing below grade level. These options do not apply to students with significant cognitive disabilities; thus only the lifeskills assessment is described in this study.

Vermont's lifeskills assessment consists of a complex interplay of IEP goals and objectives, Vermont standards, and a body of evidence. The description of the lifeskills alternate assessment reflects this interplay:

- it is based on a portfolio of individually-designed assessments, typically those used to measure IEP progress (body of evidence)
- it measures progress toward mastery of key learning outcomes
- learning outcomes are validated through research
- it is referenced to Vermont standards
- it parallels regular assessment through grades when assessment occurs, assessment content, scoring criteria
- it measures student performance and program components

Four criteria are applied in the scoring process for the lifeskills assessment. These are shown in Table 14. As indicated in the table, movement from one criterion to the next is dependent on meeting the first criterion.

Lessons Learned about Scoring Criteria in the Five States

Early in our review of documents and in state interviews, as we developed an understanding of each state approach, we began to see overlaps, new differences, and shades of meaning underlying the scoring criteria in different states.

Articulated, Embedded, or Assumed Criteria

Some states, for example, score on the quality of a teacher-designed linkage to state academic content standards, while other states embed this linkage into the assessment instruments themselves, so students are directly assessed on their performance on standards-based items. One

Table 14. Vermont Scoring Criteria

Outcome or Related Standard is Referenced in the IEP: The learning outcome can be quoted directly, paraphrased, or can be referenced to the corresponding Vermont Standards listed in the rubric. The outcome might be embedded in a larger goal or outcome. Quality gradations: 0 or 1. If 1, can go to next criterion.

Learning Outcome or Related Standard is Assessed in the IEP: This is evidenced if there are IEP goals, objectives or references to progress measures that will show whether the student is making progress toward mastering the designated learning outcomes. Quality gradations: 0 or 2. If 2, can go to next criterion.

Progress Toward Mastery of Learning Outcome or Related Standard: Progress is defined as development, improvement, or positive changes in the student's performance in relation to designated learning outcomes, including incremental skill development, fading of supports, or generalization to more natural settings. In addition, progress must be measured at least two points in time (i.e., pre-test/post-test), but preferably at multiple times across the 9 - 12 month period. Amount or rate of progress is not measured, only that it has or has not occurred. This can occur with incremental achievement of skills, with incremental fading of supports, or through generalization of skills to less restrictive environments. Quality gradations: 0 or 2. If 2, can go to next criterion.

Mastery: For the purpose of scoring the Lifeskills Portfolio, mastery is defined as the ability to perform a targeted skill or ability independently, or using natural supports and assistive technology that permit independence. If the student needs a parent, teacher, or instructional assistant to perform the skill, independence has not been demonstrated and a "0" should be given.

For some students with degenerative disabilities or health impairments, they may demonstrate progress by "standing still" or regressing as slowly as possible. In these rare cases, maintenance related to quality of life indicators, may be the most appropriate progress measure. This kind of "progress" might be documented through logs, team notes, or other anecdotal records. Quality gradations: 0 or 1.

state scores directly on the IEP linkages, but all four of the other states address IEP linkages through training or in the design of the assessment. We were able to ferret out this variation through the interview process and through written materials sent to us by the five states. This analysis is described here, accompanied by a side-by-side analysis in which we include the actual criteria scored, but also note when a criterion is assumed within the design of the system, although not articulated.

Overlapping Purpose

The distinction between student and system criteria highlighted in the past (e.g., Thompson & Thurlow, 2001) was not as clear in our analysis; we saw overlap in the purpose of the criteria. For example, how do the student measures of generalization and the system measures of multiple settings differ? Some states look for evidence that the student can perform a task in multiple settings as an indicator of generalization of the skill. Other states look for multiple settings offered to students as they are instructed as a measure of system support for high program quality. Is not the purpose of requiring that the system provide multiple settings reflected in the increased generalization of skills shown by the students within the settings? Yet these approaches do differ. With the system approach, the multiple settings are rewarded even if the student has not shown mastery or progress on skills in all areas. In states with the system approach, it seems to be assumed that for these students, whose progress is slow and sometimes hard to document, the offering of the settings is in itself a value to reinforce, regardless of student performance within all settings.

The states selected for this study have, in most cases, grappled with these shades of meaning, and have refined their criteria to reflect precisely their intent. For example, Kentucky staff clarified student vs. system measures in this way:

> Within the 6 rubric dimensions, 5 are strictly system measures. We do need to see that the student is the focus/recipient/participant in those measures but we do not assess the extent to which the actual student performance is demonstrated. It's got to be there but we don't look at how much. The 6th dimension, Performance, requires that the student actually demonstrates progress within entries so it is kind of a student measure - although the assessment itself is only used for system accountability, not student (this is the same as for all students in Kentucky, not just students with disabilities - Kentucky has no student accountability in place). The progress is not a measure of how much, but instead a measure of whether progress is documented at all. We don't ask for a certain amount of progress to move scoring from one level to another. We instead require that the evidence shows progress within certain numbers of entries. (Burdge e-mail correspondence, 2002).

Same Terms with Different Meaning and Multiple Terms Having Similar Meaning

At times we found that scoring definitions or descriptions of gradations of quality for each criterion clarified use of terms in different ways. For example, two states, Arkansas and Kentucky, score on the criterion "performance." By analyzing the rubric gradations of quality, it is apparent that Arkansas is defining performance as evidence of skill to mastery, as demonstrated in multiple settings or on multiple occasions. Kentucky defines performance within the rubric as performance and progress on IEP goals meaningful to current and future environments with products that are age-appropriate. It is not accurate to take just the headings of the rubric criteria and stop there in analyzing what "counts." Thus, in our written and side-by-side analyses, we include all criteria that were headings on the rubric or embedded in the definitions or descriptions of quality gradations and applied in the scoring process. The converse of this is that although Louisiana, Oregon, and Vermont had no criterion labeled "performance," their criteria clearly address similar values to the Arkansas and Kentucky performance criterion. The similar criteria were reflected in terms like skill level, mastery, progress, accuracy, and generalization.

Results: State Side-by-Side Analysis and Criteria Definitions

Given the lessons learned (e.g., articulated, embedded or assumed criteria, overlapping purpose, differences in meaning and multiple terms with the same meaning), we provide a description of scoring criteria used in the five states. These descriptions include actual criteria listed in rubrics, criteria implied in the design of assessment items or format, and additional criteria found in rubric gradations of quality descriptions or described in interviews as having been designed into the alternate assessment items or provided in training. Since we find that the distinction between student and system criteria is blurred at best, we have in some cases defined the same criteria in both ways, and have shown where states use one or both definitions of the criteria. For example, in student criteria we include "Settings when defined as student performing the skills in multiple settings" and in system criteria we include "Settings (multiple) when defined as system offering multiple settings for student." It is not always clear how the scoring process distinguishes between the two, but we have tried to clarify the distinction.

Student criteria identified in at least one state include:

- Accuracy (quantitative, number correct)
- Content standards performance (student performance on standards-based skills or tasks provided by the state)
- Level of assistance (prompts, showing degree of independence)
- Multiple purposes (generalization)
- Mastery
- Performance (qualitative)
- Progress
- Settings (student performance evidence of generalization)

System criteria identified in at least one state include:

- Appropriateness (age, challenge, authenticity, meaningfulness)
- Content standards linkage (teacher developed skills and tasks)
- IEP linkages
- Self-determination (evidence of opportunities for student choice)
- Settings (system provided)
- Social relationships (system provided opportunities)
- Support (access to system provided appropriate supports and technology)

Student Criteria Used Across the Five Sample States

Student criteria used across the five sample states are shown in Table 15. Student scoring criteria reflect actual student performance, and may include a quantitative measure of student accuracy on items reflecting desired knowledge and skills, or a qualitative judgment about the overall level of performance of the student in terms of skills and competence, degree of progress, independence, or ability to generalize. States that included each criterion, and a general definition for each follow Table 15. State by state definitions or justifications for assumed or embedded student criteria are in Appendix B.

General Statements On Use of Student Criteria

All five states have specific student criteria, as defined as either quantitative or qualitative measurement of student performance. None of the states suggested that student criteria were inappropriate for a large-scale assessment. That is in contrast to the discussions of system criteria: systems criteria remain controversial among these states and in the larger testing community. In the traditional pencil and paper tests that are most familiar to us described in the opening section of this report, some states argue that the only criteria measured is student skill and knowledge. Others argue that we can use results from a well-designed traditional test to make some assumptions about what has or has not been taught to a group of students, thus we are measuring the system. For example, if almost all 4th grade students in one school were able to answer test items related to number sense correctly, while almost no 4th grade students in another school could do so, then we can make some assumptions about the opportunities to learn that content in the two schools. The arguments in electing to have system criteria or not generally relate to the discussions states have had with stakeholders on how to define the results of the

Table 15. Student Criteria Used Across the Five Sample States

	Arkansas	Kentucky	Louisiana	Oregon	Vermont
Accuracy (quantitative)				X (Extended Academics)	
Content Standards (student performance on standards-based skills or tasks provided by the state)			X	X (Extended Academics – math, reading, writing; Extended CLRAS – career and life skills standards only)	
Level of assistance (prompts, degree of independence)	X		X	X (CLRAS)	X
Multiple purposes			X		
Mastery	X (part of performance criterion)				X
Performance (qualitative)	X	X			
Progress		X (part of performance and support criteria)		X (growth model, Extended Academics)	X
Settings – student performance evidenced	X (part of performance criterion)		X	X (CLRAS)	X

best possible teaching for students with significant cognitive disabilities, or in other words, how successful outcomes have been defined. The specific student criteria in each state reflect those assumptions and values as well.

Definitions of Student Criteria, Across States

Accuracy (quantitative, number correct)

Only one of the sample states, Oregon, specifically includes accuracy as a criterion. For the Oregon extended assessments in reading, mathematics, and writing, the definition of accuracy is quantitative, and is reflected in number correct (or partially correct), reading rate, and reading accuracy. However, one could argue that all of the sample states had same form of accuracy criteria, including Oregon on its Extended CLRAS, with accuracy defined qualitatively using measures such as independent performance of skills, generalization of skills to multiple settings or for different purposes, and appropriateness criterion such as challenge or complexity.

Content Standards Performance (student performance on standards-based skills or tasks provided by the state)

Louisiana and Oregon have provided items or state specified knowledge and skills that are aligned to state content standards. Oregon's extended assessments in reading, mathematics, and writing are aligned to the math, reading, and writing content; the Extended CLRAS is aligned to the career related learning standards content, but not to math, reading, and writing. The other states score the relationship to standards as a system measure, designed or chosen by the teacher.

Level of assistance (prompts, showing degree of independence)

All the states except for Kentucky look at independence using level of prompting as a student measure. They may allow or encourage appropriate natural supports, adaptations, modifications, and assistive technology as tools to assist in increasing independence. Oregon scores for level of assistance only on the Extended CLRAS, and not on the extended assessments in reading, mathematics, and writing. Kentucky views level of support as a system criterion, encouraging natural supports, accommodations, and assistive technology as provided to the student to increase independence.

Multiple purposes (generalization)

Louisiana looks at both multiple settings and multiple purposes to measure student generalization.

Mastery

Mastery is one indication of increased quality on the Arkansas "performance" criterion; Vermont defines mastery as ability to perform a skill independently.

Performance (qualitative)

Arkansas and Kentucky each define broadly the term "performance," but with very different definitions. The term is almost generic, and the varying definitions probably reflect the fact that the term "performance" can mean whatever a given set of stakeholders defines as the essential outcomes of instruction. Thus, it may be appropriate to suggest that all of the scoring criteria used by states attempt to define "performance" for this population of students.

Progress

Both Kentucky and Vermont link progress to IEP goals; Kentucky also mentions progress in relationship to independence. Oregon's approach is very different, using a growth model for evaluation for the whole assessment system.

Settings (student performance evidence of generalization)

All five states include some form of a generalization criterion, as evidenced by settings. Oregon uses this only in the Extended CLRAS and not in the extended assessments in reading, mathematics, and writing. Kentucky uses this as a systems criterion, measuring whether the system is providing opportunities for instruction, practice, application, or performance of the specified skills in multiple settings.

System Criteria Used Across the Five Sample States

System criteria used across the five sample states are shown in Table 16. System scoring criteria include measures of system performance, such as whether students are provided instruction in multiple settings, whether they are provided opportunities to plan, monitor, and evaluate their work, whether they work with non-disabled peers, and whether they are provided with appropriate human and technological supports. As previously discussed, the distinction between student and system criteria is nebulous. Our sample states addressed this directly or indirectly as part of the interview process, and their comments follow.

General Statements on Use of Systems Criteria

Arkansas stakeholders decided to weight "performance" at the highest level since they believed the emphasis should be on measuring actual student evidence, and not the system. However, determining how to measure quality of programs for these students without looking at appropriateness, level of assistance, or settings is difficult, and their criteria and scoring process allow for that interplay.

For Kentucky, the issues of system measurement were central to decisions about scoring criteria, and about how Kentucky's approach has changed over the last decade. Kentucky believes that for this population of students, looking at *how students are taught* is really important. Assessment may show that students are not achieving well, but if students are not being taught well they cannot be blamed for not learning. Kentucky includes multiple systems criteria in order to "shine a light" on teaching practices, practices that in the past too often took place in isolated classrooms behind closed doors.

In contrast, the Louisiana task force considered adding questions to the assessment that did not measure what a student was doing, but measured what supports were available, and the amount of time students spent in inclusive settings. A decision was made to view the alternate as an

Table 16. System Criteria used Across the Five Sample States

	Arkansas	Kentucky	Louisiana	Oregon	Vermont
Appropriateness	X (age, challenge, authenticity)	X (age and meaningfulness as part of performance criterion)	*	*	*
Content Standards Linkage (teacher developed skills and tasks)	X	X			X
IEP linkages	*	*	*	*	X
Self-determination		X			*
Settings (system provided)	X	X			
Social Relationships		X			
Support (access to appropriate supports)		X			

* Assumed in design of training materials, assessment format, or in criteria.

assessment of a student's performance. The task force decided not to include system measures because it believed that those measures were not included in the general statewide assessment. Some task force members wanted to include those measures as one way to obtain valuable information, but the determination was made that there were other ways to get that information (e.g., monitoring). They decided to separate monitoring from evaluation, even though the data would be used to look at programs. They wanted to see specifically what was happening with these children, thus focused their alternate assessment on student criteria.

Oregon discussed concepts related to the underlying learning theory that drives its assessment system. Oregon personnel believe their assessments are grounded in learning theory in their focus on student criteria, not systems criteria. They focus training on how the alternate assessment relates to instruction and planning, and how teachers might use these measures on their own periodically throughout the year to assess progress. Training personnel stress the importance of using the alternate assessment as an integral part of the instructional process, and not as an add-on task. That is the intent of Oregon's standard assessments; they are to change the way teachers teach.

Vermont's stakeholder group started, and then abandoned, a process of designing its learning outcomes specifically for the alternate assessment. They were committed to using validated learning outcomes, and were reluctant to, as they state, "reinvent the wheel." Instead, they elected to adapt existing materials that were already validated. The authors of the COACH (*Choosing Outcomes and Accommodations for Children: A Guide to Educational Planning for Students with Disabilities*, Giangreco, Cloninger, & Iverson, 1998) from the University of Vermont allowed the group to use their learning outcomes as the criteria for alternate assessment. They are not using the COACH as the alternate assessment; instead, Vermont "borrowed" the learning outcomes and measures student progress against those learning outcomes. One of the COACH authors worked with state staff to cross-link the selected outcomes to Vermont content standards. The stakeholder group decided to ask teachers to show that students had made progress across three of the learning outcomes, but they could pick from anywhere in the developmental spectrum to match the students' needs. As part of their system criteria, Vermont scores the linkage of the evidence to Vermont standards that are cross-linked to the COACH outcomes, as the learning outcomes are referenced and assessed in the IEP.

States that included each system criterion, along with general definitions are found in Table 16. State by state definitions or justifications for assumed or embedded student criteria are in Appendix C.

Definitions of System Criteria, Across States

Appropriateness (age, challenge, authenticity, meaningfulness)

Each of the states has struggled with determining "how high is high enough" to raise expectations and outcomes for this population of students, whether scored as part of the rubric, or embedded in test items or state specified skills. Age appropriateness is a common shared criterion, but level of challenge, authenticity of tasks, and meaningfulness for current and future life roles are also included.

Content Standards Linkage (teacher developed skills and tasks)

Arkansas, Kentucky, and Vermont score on degree of linkage to academic content standards through teacher developed or identified skills or tasks. Two of the states, Louisiana and Oregon, instead provide state developed tasks or specified skills to ensure standards linkages, and do not score for them.

IEP Linkages

Only Vermont scores the nature of linkages to the IEP. Kentucky references IEP goals in its progress definition. Yet all five states emphasize the challenging training issues that are emerging as teachers and other IEP team members, including parents, make the shift to standards-based thinking and planning for IEP goals and objectives. Thus, in the four other states we have marked this criterion as "assumed."

Self-determination (evidence of opportunities for student choice)

Kentucky defines "self-determination" as evidence of degree of student choice with an impact on student learning across and within entries, and evidence of planning, monitoring, and evaluating of the self-performance to improve or set new goals. It is the only state to score for this, noting that there is a significant body of research, including a recent meta-analysis of research (Algozzine, Browder, Karvonen, Test, & Wood, 2001) to document self-determination as a teachable, essential skill for future success for this population of students. However, Vermont state staff point out that although they do not have specific scoring criterion on self-determination, the ability to make choices is embedded throughout their learning outcomes. For example, one of the communication learning outcomes is, "Uses communication to indicate preferences and make choices."

Settings (system provided)

Arkansas scores settings across all content entries, and gives it the lowest weight of all its criteria. Kentucky sees this as an important measure of system quality, giving students opportunities to generalize knowledge and skills. Kentucky links this criterion to that of "Social Relationships," requiring that settings include age appropriate peers.

Social relationships (system provided opportunities)

Kentucky scores social relationships on appropriateness of interaction, frequency, diversity of range in peers, sustainability, and reciprocity of relationships.

Support (access to system provided appropriate supports and technology)

Kentucky defines support as degree of peer support, tutoring, or natural supports; use of adaptations, modifications, or assistive technology that demonstrates progress toward independence.

Common Ground and Differences

On surface examination, the scoring criteria used by the five selected states appear to be different. State responses to the 2001 survey of state directors of special education (Thompson & Thurlow, 2001) suggested significant differences as well. Yet when the scoring elaborations and processes are examined closely, many similarities emerge. After careful analysis of how some assumptions are built in to the instrument development or training processes, even more similarities emerge. The definitions and examples and the side by side examination of the criteria, the scoring elaborations, and the assumed criteria in the design of training materials and assessment format yield a surprising degree of commonality in the way these states define success for students with significant cognitive disabilities. These commonalities are highlighted in Table 17.

Table 17. Student and System Criteria, by State

Arkansas	Kentucky	Louisiana	Oregon	Vermont
STUDENT	**STUDENT**	**STUDENT**	**STUDENT**	**STUDENT**
---	---	---	Accuracy	---
---	---	Content Standards (student performance)	Content Standards (student performance)	---
Level of assistance (prompts, independence)	---	Level of assistance (prompts, independence)	Level of assistance (prompts, independence)	Level of assistance (prompts, independence)
Mastery	---	---	---	Mastery
---	---	Multiple purposes	---	---
Performance (skill to mastery; settings, occasions)	Performance (progress, appropriateness)	---	---	---
---	Progress	---	Progress (growth model)	Progress
Settings (generalization)	---	Settings (generalization)	Settings (generalization)	Settings (generalization)
SYSTEM	**SYSTEM**	**SYSTEM**	**SYSTEM**	**SYSTEM**
Appropriateness (scored)	Appropriateness (scored)	Appropriateness (assumed)	Appropriateness (assumed)	Appropriateness (assumed)
Content Standards Linkage (teacher)	Content Standards Linkage (teacher)	---	---	Content Standards Linkage (teacher)
IEP linkages (assumed)	IEP linkages (assumed)	IEP linkages (assumed)	IEP linkages (assumed)	IEP linkages (scored)
---	Self –determination	---	---	Self-determination (assumed in selected learning outcomes)
Settings (offered)	Settings (offered)	---	---	---
---	Social Relationships	---	---	---
---	Supports (access)	---	---	---

Common Criteria Across the Five Sample States

Six criteria are included in all of the five states' approaches in some way, either articulated or assumed. They include "content standards linkage," "independence," "generalization," "appropriateness," "IEP linkage," and "performance." How the states address these criteria varies however, as shown in Table 18.

Table 18. Common Criteria Across the Five Sample States

1. **Content Standards Linkage.** All states have either scored teacher linkage to content standards or built in linkage to content standards on student skills developed by state.
2. **Independence.** All score for independence in terms of level of prompt, although Oregon has these on the Extended CLRAS only, and not on the extended academics assessments.
3. **Generalization.** All score for generalization in terms of performance in varied settings, although Oregon has these on the Extended CLRAS only, and not on the extended academics assessments.
4. **Appropriateness.** All address "appropriateness" either scored or assumed in design, which incorporates age appropriateness, and level of challenge. Oregon, Louisiana, and Vermont allow and do not question teacher choice of level of challenge, but do provide guidance.
5. **IEP Linkage.** All grapple with IEP linkage, but only Vermont actually scores for it - the others are assumed in design or training.
6. **Performance.** Only two states score for the actual term "performance." Arkansas defines performance in part as level of skill or mastery and multiple settings; Kentucky scores on progress and appropriateness. But all states have some related criterion, e.g., accuracy, mastery, progress, independence, multiple settings, multiple occasions, or multiple purpose.

These six criteria have emerged from a long history of practice, and each is reinforced by research on teaching and learning for students with significant cognitive disabilities. Questions remain about how best to incorporate these commonly accepted indicators of success for these students. Here are some questions about each of the criteria, followed by a discussion of possible options and responses, based in part by discussions with selected alternate assessment researchers and training leaders (NCEO Alternate Assessment Retreat, 2002).

1. Content Standards Linkage. *Is it better to have standardized tasks or common items with linkage to standards provided by the state, or to allow the flexibility and the variability of teacher defined linkages to the content standards?*

Content standards linkages are an essential part of alternate assessments if they are going to meet federal assessment and accountability requirements. It appears from our sample states that either state-provided or teacher-developed linkages to content standards can work, but the training issues related to teacher-defined linkages are very great. The shift to standards-based teaching and learning has been difficult for all teachers, and although training on content standards has occurred in every state, teachers who provide special education services and supports often have been overlooked in the training. Teachers will not be able to align instruction or assessment to content areas without training or access to research-based models. The standardization of tasks or items eliminates the need for teachers to have significant amounts of training for the assessment part of the puzzle. That is because the standards linkage can be embedded in the assessment item or task, and is not dependent on teacher understanding and skill. The standardization of assessment tasks does not address the need for teacher and IEP team member training and support so that instruction becomes standards-based, providing students with opportunities to be successful on the standards-based assessment.

Some states that have selected teacher-defined linkages have done so because of the extreme variability in the student population participating in alternate assessment. They believe that even with the training challenges posed by teacher-defined linkages, flexibility of task and participation level are important characteristics of the alternate assessment. One state interviewee suggested that teacher-defined linkages allow for greater flexibility in documenting a larger number of content standards through evidence of the interdisciplinary use of standards-based skills. For example, student performance on a math standard of computation also might show number sense, data, or probability skills as well as reading and writing and may occur within other content area standards such as science or social studies. Another option that is being discussed in some of these states is to go to "flexible structuring" of performance tasks. For example, the Louisiana approach defines state-specified knowledge and skills (target indicators), but teachers are allowed to design the actual activity in which to demonstrate the knowledge and skills, and the IEP team can choose the one of three "participation levels," reflecting three levels of difficulty, that is most appropriate for the student.

2. Independence. *We have a long history of best practice and a solid research base to support the emphasis on independence for students with significant cognitive disabilities. Is "level of prompting" the best way to judge independence? How do the system-provided opportunities for accommodations, assistive technology, and natural supports fit into the measurement of whether students are maximizing their independence?*
There has been a focus on prompting strategies in preservice education for teachers of students with significant disabilities. Perhaps as a result, Oregon state staff reported that many teachers were unaware of the level of prompting they were using, and as the Extended CLRAS was implemented, many teachers reported difficulty in removing prompts, since they automatically provided them without thinking. Other states also have reported overuse of teacher "automatic" prompting. Some states have found that they had to define differences between what an accommodation, assistive technology, or natural support is versus what a prompt is. The dilemma is that increased use of accommodations, assistive technology, or natural supports can increase independence; increased use of prompting will decrease independence. Clarifying the distinction is important for instruction and for assessment, and scoring criteria and processes need to clearly differentiate among them as well.

3. Generalization. *Is offering rich and varied settings in which students learn enough to indicate program quality, or must we see the student performing skills to a specified degree of success in those settings? Are different purposes the same as different settings in the ability to show generalization?*
This criterion is more complex than it appears at first glance. There is an argument that taking

the same activity and moving it to some other "place" is not necessarily generalization. This approach may work better for life skills, where the focus has been on generalization of functional life skills across natural settings, than for academics. When considering generalization for academic knowledge and skills the focus may need to be on different tasks or materials on which the knowledge and skills, are used. For example, if the student can understand the main idea of à story, can he or she also do that with a picture or a news story? Considering instructional contexts – media and materials, formats of materials, response methods or modes – may provide a more important context for generalization of academic skills than physical locations. The specific focus of generalization may vary by type of disability. For example, for students with autism, the focus may need to be on getting behavior to generalize across different people or to different settings.

State staff from one of the sample states that has a clear commitment to settings as a system measure, suggests that adhering to specific skills applied in academic contexts overlooks the value of the shared culture of participating in multiple settings where all other same age peers are participating. In another of the sample states, state staff commented that inclusive settings are encouraged by scoring for multiple settings. Thus, there appears to be another purpose to including "settings" as a system criterion, that of promoting inclusive practices. Clarifying these distinctions and ensuring that teachers understand them will be important if the policies are to drive desired improvements. A few states, although none of the sample states, score and report on system opportunities for multiple settings, but do not include the scores in accountability calculations. They use the scores to provide feedback to teachers, schools, and districts, and to guide monitoring decisions, but they do not include the scores in system accountability indices.

4. Appropriateness. *If you allow teacher/IEP team choice on the tasks used or the skills to be targeted, how do you determine the appropriate level of challenge, authenticity, or age appropriateness? How do you know if a high challenge is high enough?*
Appropriateness is either stated or assumed across all states, reflecting another value of special education since the early 1980s. Since that time there has been an effort to move away from the "developmental or mental age" approach (even though still used in practice) to a focus on what is appropriate for a student's chronological age. Most states have avoided defining alternate assessments as measurement of a developmental sequence, for example, by rejecting infant or preschool definitions of knowledge and skills as an appropriate linkage to content standards. The value of age-appropriateness is emphasized by references to "grade level standards" in No Child Left Behind, and to the general curriculum in IDEA.

"Appropriateness" as a criterion is typically not just about age. Raising expectations was a centerpiece of the 1997 reauthorization of IDEA, and "challenge" as a sub criterion is a common definition of "appropriateness." The appropriate degree of challenge is difficult to score for a widely differing group of students, although some states are turning to baseline measurements such as pre and post tests to do so. In contrast, the teacher developed "criteria for mastery" normally included in IEPs does not appear to be effective in setting high expectations. "Authenticity" is another common indicator of appropriateness. For example, knowledge and skills instruction and assessment embedded in grade level content or contexts is considered to reflect authenticity. Age appropriateness is easier to see and score – and certainly has been a high value in the push to move away from developmental models that prevailed years ago. However, are the values of appropriateness or level of challenge or authentic learning experiences worth striving toward given the difficulty in scoring these? Clearly many of the states in our sample believe these values are worth the difficulty in scoring, but only Kentucky and Arkansas score these values directly. Louisiana, Oregon, and Vermont provide guidance to teams on choices of participation levels (Louisiana), whether the student takes any of the extended assessments in reading, mathematics, and writing or not as well as the specific routines and embedded skills in the Extended CLRAS (Oregon), and IEP goals and objectives (Vermont), but they do not score on the choice made.

5. IEP Linkage. *Each of the sample states expressed concerns about the quality of IEPs, and each has begun to address the problem through statewide training. How do we address the universally poor quality or poor implementation of standards-based IEPs, and how does this relate to current approaches?*
All five of the states mention this as an issue, and all have devoted training resources to address the problem. Even in Vermont, where alternate assessments are scored on the linkage, the existing quality of the IEP was found to be poor in the first year of implementation of their alternate assessment system. In a study by Sands, Adams, and Stout (1995) over half (55%) of the 341 elementary and secondary special educators surveyed believed that each student with disabilities should have his or her own curriculum based on needs as documented on the IEP. The researchers concluded, "In the absence of a curriculum base that provides direction for special education programs, instructional decision making and procedures are often haphazard and widely divergent" (p. 69). Even with IDEA requirements to provide access to the general education curriculum, Tindal and Fuchs (2000) found it disturbing that the IEP "typically does not conform to the substantive spirit reflecting federal legislation. Rather, IEPs have served primarily as a tool for procedural compliance monitoring, whereby federal auditors make sure that a complete IEP exists for each student receiving special education services and that IEPs document how (i.e., where, when, and by whom) those services are delivered" (p. 4).

There is evidence that implementation of thoughtful alternate assessment approaches can directly improve quality of IEPs. During the first year of implementation in Vermont, only about a third of the portfolios were rated as meeting or exceeding program expectations. Operationally, that meant that the portfolio included an IEP that did not reference Vermont standards, and did not include an adequate progress measure. In the second year, after concerted effort for IEP related professional development, more than two-thirds met or exceeded expectations. This may be an area where thoughtful training connected to alternate assessment implementation can have a rapid positive effect on practice, if the assessments are designed specifically to meet the unique needs of this population.

At a recent discussion at a Technical Advisory Committee (TAC) meeting for a state not in this study, most TAC members wondered why the state's alternate assessment could not simply be a report of how many IEP goals and objectives were met by the students in a given year. They expressed amazement at the research base that indicates the quality of IEPs is not sufficient for the measurement purpose, nor do they generally reflect standards-based instructional goals. These TAC members are typical of the individuals on most state TACs: all have outstanding expertise and experience in measurement of academic achievement for the general population of students. Most have never worked with students who have significant cognitive disabilities, nor have they seen what typical evidence of their learning looks like. Another opinion expressed by the TAC members was that if performance tasks are not standardized for all students, then it is not really measurement. There is a huge gap between the experiences of technical and measurement experts and the experiences of the students who participate in alternate assessments. State alternate assessment stakeholder committee members and state special education personnel who understand these issues need to work closely with technical advisors and state assessment personnel as they develop and refine their alternate assessments.

6. Performance. *Students doing something – is this the common definition of "performance" as defined in varying ways in the state criteria? Is "performance" the bottom line in measurement of academic achievement for students in the alternate assessment?*
That none of the states used the same language to define "performance" is telling – perhaps all of the criteria together is what defines performance! Appropriateness, accuracy, mastery, progress, independence, multiple settings, multiple occasions, or multiple purpose all seem to contribute in varying ways to the criterion of "performance." "Performance" generally reflects that a student has actually learned a standards-based skill or at least demonstrates something that shows progress toward a defined outcome. There is a need for continuing discussion among state stakeholders to ensure that the scoring criteria as applied reflect the assumptions and values that can be defended by research and best practice for the highest possible outcomes for students with significant cognitive disabilities.

Differences in Criteria Across the Five Sample States

Three scoring criteria are very different across the five states' approaches. They include "system vs. student emphasis," "mastery," and "progress." Two states, Kentucky and Oregon, also use additional criteria that no other state uses. A brief description of these differing criteria is shown in Table 19.

Table 19. Differences in Criteria Across the Five Sample States

1. **System vs. student:** Louisiana and Oregon have no directly scored system measures; all the other states do. Both Louisiana and Oregon have built in some systems controls into their approach.
2. **Mastery:** Vermont and Arkansas are the only states that emphasize mastery. However, Oregon's criterion of "accuracy" appears to be addressing the concept of mastery.
3. **Progress:** Louisiana and Arkansas do not score on progress, while Kentucky and Vermont do. Oregon works a slightly different meaning of "progress" into their accountability model.
4. **Single state criteria:** Three states have criteria that no other state has included. Kentucky measures the system for access to supports, social relationships, and self-determination. (Vermont has some learner outcomes specifically related to self-determination.) Oregon is the only state using accuracy as a primary criterion, that is, using number correct, or accuracy with a quantitative basis, on its extended assessments in reading, mathematics, and writing, but not on the Extended CLRAS.

1. System vs. Student Criteria

The system versus student criteria differences among the five states have been examined throughout this paper. It is important for states to articulate why they support system or student criteria. All of the five states have attempted to do so.

2. Mastery

Vermont defines mastery as the ability to perform a targeted skill or ability independently, or using natural supports and assistive technology that permit independence. Thus, "mastery" for Vermont relates to "independence." Arkansas includes "mastery" as part of the definition of "performance;" it is defined as performing a skill in multiple settings or on multiple occasions, and is thus interrelated to generalization. These two examples illustrate the murky nature of clarifying criteria: Is mastery as defined in these states a different criterion, or are these simply different names for independence and generalization? Oregon does not use the term mastery, but the criterion of "accuracy" may be addressing the concept of mastery.

3. Progress

Progress is identified in three states, but the three states have very different approaches to it. Kentucky scores for "progress" in two places: (1) progress on independence as part of the "support" criterion, and (2) progress on IEP goals for the "performance" criterion.

Vermont broadly defines progress as "development, improvement, or positive changes in the student's performance in relation to designated learning outcomes, including incremental skill development, fading of supports, or generalization to more natural settings. . . measured at least two points in time . . . amount or rate of progress is not measured, only that it has or has not occurred." But Vermont also notes that "for some students with degenerative disabilities or health impairments, they may demonstrate progress by 'standing still' or regressing as slowly as possible. In these rare cases, maintenance related to quality of life indicators, may be the most appropriate progress measure. This kind of 'progress' might be documented through logs, team notes, or other anecdotal records."

Oregon has a very different approach to progress. It has committed to a growth model of accountability and looks for progress along a continuum of learning that encompasses all students. Oregon is working to integrate its extended assessments in reading, mathematics, and writing into the continuum, as described earlier, but the Extended CLRAS appears to be outside the continuum.

4. Single State Criteria

Kentucky has carefully studied the research base on teaching and learning for students with significant disabilities, and it has carefully articulated state stakeholder values. In the end Kentucky comes down strongly on the side of extensive system measures, such as access to supports, and also student indicators some states see as "removed" from academics, such as social relationships and self-determination. Kentucky points to a strong research base that connects self determination skills in the context of instruction to achievement of learning goals, suggesting that these measures are strongly related to academics. These decisions also reflect values grown out of inclusion and transition systems change efforts that have been tied to the development of Kentucky's alternate assessment approach.

Similarly, Oregon has gone to its more behavioral research base and stakeholders, and come down strongly as well, but in a very different direction from Kentucky for at least for one part of its alternate assessment system, the extended assessments in reading, mathematics, and writing. Oregon is the only state to use a quantitative measure of "accuracy," on its extended assessments in reading, mathematics, and writing but not on its Extended CLRAS. The Extended CLRAS is very similar to other state approaches, with performance events scored on generalization and level of prompting. Not all students with significant cognitive disabilities take the extended assessments in reading, mathematics, and writing, or even one of them; some take only the Extended CLRAS.

Recommendations

The common criteria and differences in criteria across the five states described in this report do not reflect "right" or "wrong" approaches. Each of these five states has developed alternate assessment scoring criteria that reflect their best understanding of what successful outcomes are for students with significant cognitive disabilities. We hope that many state personnel will have opportunities to read this report and will be prompted to carefully examine their own criteria, asking what each criterion means and how it is important to each student's success. Alternate assessments are still very new, and taking the time for thoughtful reexamination is critical. Here are some questions and recommendations you may find helpful as you work to ensure your alternate assessment reflects your state's best understanding of successful outcomes for students with significant cognitive disabilities.

1. What are your stated and embedded scoring criteria? What practice or research base was consulted as you developed these criteria? In addition to the actual criteria listed in rubrics, be sure to articulate the criteria that are implied in the design of assessment items or format, and additional criteria that may be embedded in definitions or in rubric gradations of quality descriptions.

2. Clarify exactly what the scoring criteria (i.e., develop definitions) are for your state's alternate assessment and assign them to either the student or the system, or both. This will assist in identifying the underlying assumptions reflected in scoring.

3. Determine whether the underlying assumptions reflected in scoring criteria are consistent with the views of various stakeholders, including educators, parents, and assessment personnel.

4. Are the values embedded in scoring criteria clearly articulated in written materials and training on alternate assessment?

5. Could every teacher who has students participating in alternate assessment in your state tell you what the stated and embedded scoring criteria are? Could other IEP team members, e.g., parents, general educators, related service providers?

6. What do you see as the potential that these criteria, and any changes made, will affect quality of programs and outcomes for students with significant disabilities? How will that occur? Do you have data that reflect any changes so far, either anecdotal or formal?

7. Carefully examine the scoring process in light of the scoring criteria to see if the process truly measures the criteria.

8. Examine scores after conducting an alternate assessment in light of scoring criteria – do the scores reflect original intentions? Is it possible to see areas of strength and areas that could be improved in the education of these students based on these scores?

9. Examine changes in scores across years – do these changes reflect progress toward the successful outcomes intended for these students?

10. Consider alternate assessment criteria in light of stated and assumed scoring criteria for

the general assessment – how are they the same and different? What discussions have you had about the need for "the same" or "analogous" components in your alternate assessment and in your general assessment?

As states work to ensure that achievement standards are set that truly reflect the highest learning standards possible for the small population of students with significant cognitive disabilities, the alternate assessment process must be carefully and thoughtfully designed. Since alternate assessments are a much more recent development than general assessments, the assumptions underlying them are still debated and discussed. These deliberations can sometimes feel frustrating for teachers, policymakers, and other stakeholders, but they are an important activity that helps a state develop an alternate assessment that reflects its educational values for students with significant cognitive disabilities.

References

Algozzine, R., Browder, D., Karvonen, M., Test, D.W., & Wood, W.M. (2001). Effects of interventions to promote self-determination for individuals with disabilities. *Review of Educational Research, 71* (2), 219-77.

Browder, D.M. (2001). *Curriculum and assessment for students with moderate and severe disabilities.* New York: Guilford Press.

Browder, D., Flowers, C., Ahlgrim-Delzell, L., Karvonen, M., Spooner, F., & Algozzine, R. (2002). *Curricular implications of alternate assessments.* Paper presented at the National Council of Measurement in Education Annual Conference, New Orleans.

Burdge, Michael. E-mail correspondence to Rachel Quenemoen, November 2002.

Giangreco, M.F., Cloninger, C.J., & Iverson, V.S. (1998). *Choosing outcomes and accommodations for children: A guide to educational planning for students with disabilities* (2nd edition). Baltimore, MD: Paul H Brookes.

Kleinert, H. L., & Kearns, J. F. (1999). A validation study of the performance indicators and learner outcomes of Kentucky's alternate assessment for students with significant disabilities. *Journal of the Association for Persons with Severe Handicaps*, *24*, 100-110.

Kleinert, H. L., & Kearns, J. F. (2001). *Alternate assessment: Measuring outcomes and supports for students with disabilities.* Baltimore, Maryland: Brookes Publishing.

NCEO Alternate Assessment Retreat (2002). Notes from "think tank" retreat of selected national researchers and state training leaders. November 20-22, 2002, Monticello, MN. *

Olson, B., Mead, R., & Payne, D. (2002). *A report of a standard setting method for alternate assessments for students with significant disabilities* (Synthesis Report 47). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Quenemoen, R., Rigney, S., & Thurlow, M. (2002). *Use of alternate assessment results in reporting and accountability systems: Conditions for use based on research and practice* (Synthesis Report 43). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Roeber, E. (2002). *Setting standards on alternate assessments* (Synthesis Report 42). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Sands, D. J., Adams, L., & Stout, D. M. (1995). A statewide exploration of the nature and use of curriculum in special education. *Exceptional Children, 62* (1), 68-83.

Smith, S.W. (1990). Individualized education programs (IEPs) in special education – from intent to acquiescence. *Exceptional Children, 57*, 6-14.

Thompson, S., & Thurlow, M. (2001). *2001 State special education outcomes: A report on state activities at the beginning of a new decade*. Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Thompson, S., Quenemoen, R., Thurlow, M., & Ysseldyke, J. (2001). *Alternate assessments for students with disabilities*. Thousand Oaks, CA: Corwin Press.

Thurlow, M., Quenemoen, R., Thompson, S., & Lehr, C. (2001). *Principles and characteristics of inclusive assessment and accountability systems* (Synthesis Report 40). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Tindal, G., & Fuchs, L. (2000). *A summary of research on test changes: An empirical basis for defining accommodations*. Lexington, KY: Mid-South Regional Resource Center.

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.

Ysseldyke, J. E., & Olsen, K. R. (1997). *Putting alternate assessments into practice: What to measure and possible sources of data* (Synthesis Report No. 28). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

State Materials Used as References

Several state personnel participated in the development of the content included in this report. Each person interviewed shared many wise and thoughtful insights into the development of state alternate assessment systems. It is through the commitment and dedication of each of the people interviewed that alternate assessments continue to improve, providing important information and enabling states to truly leave no child behind. NCEO gratefully acknowledges the following participants in this interview process.
Arkansas: Marcia Harding and Tom Hicks
Kentucky: Mike Burdge
Louisiana: Jeanne Johnson and Leslie Lightbourne
Oregon: Pat Almond
Vermont: Michael Hock

Arkansas

Interview: Marcia Harding and Tom Hicks, October 21, 2002

Arkansas Domain Scoring Rubric for Students with Disabilities, with Scoring Domain Definitions, and Glossary

Arkansas handouts at Alternate Assessment pre-session to CCSSO Large-Scale Assessment Conference, Desert Springs, CA, June 2002

Arkansas Alternate Assessment Implementation Guide for Students With Disabilities, 2000-2001

Kentucky

Interview: Mike Burdge, September 17, 2002

Kentucky Alternate Portfolio Assessment: Teacher's Guide 2001-2002

Kentucky Alternate Portfolio Assessment: Frequently Asked Questions

2001-2002 Training PowerPoint Presentation

Louisiana

Interview: Leslie Lightbourne and Jeanne Johnson, September 20, 2002

http://www.doe.state.la.us/DOE/Assessment/PDFs/LAAtamManual.pdf

http://www.louisianaschools.net/DOE/Assessment/LAA.asp

General Education Access Guide: Section IV Curriculum Issues for Students in Alternate Assessment

August 2000

2002-03 LAA PowerPoint Presentation

Louisiana Alternate Assessment Test Administration Manual, Spring 2002

Oregon

Interview: Patricia Almond, September 19, 2002

Oregon Extended Career and Life Role Assessment System Based on the Career Related Learning Standards Identified with the CAM

Extended Career and Life Role Assessment Systems: Administration Manual 2001-2002

http://www.ode.state.or.us/asmt/administartion/extasmts

http://www.ode.state.or.us/asmt/administration/comprehensiveasmt.pdf

Vermont

Interview: Michael Hock, September 19, 2002

Preliminary Report on 2002 Lifeskills Portfolio Scoring

Summary of Alternate Assessment Validation Study

Participation Guidelines

Documentation of Eligibility form

Schedule for the Collections, Scoring, and Reporting of Lifeskills Portfolios

Scoring Procedures, Guidelines and Rubrics
Portfolio table of contents sheets
Training materials: (a) Building a Better Portfolio, (b) Criteria for Evaluating Lifeskills Portfolio Measurement Strategies, (c) Effective Measurement Strategies and Techniques for Students with Severe Disabilities, and (d) Exemplary Portfolio Pieces form the 2001 Portfolio Scoring Institute

* NCEO extends its sincere thanks to the following individuals who provided us with wise and thoughtful assistance in the development of the analysis and conclusions we present here. They were participants at the NCEO Alternate Assessment Retreat in November, 2002.
Mike Burdge, Jean Clayton, and Mark Hurte – Kentucky
Lynn Ahlgrim-Delzell and Diane Browder – North Carolina
Bryan Ayres and Lynn Springfield – Arkansas
Pam Green and Dan Weiner – Massachusetts
Pat Almond and Joel Arick – Oregon

Appendix A

Interview Guide

General Outline for Information on Calls

The methods used to gather evidence for alternate assessment vary around the country.

The performance measures used as criteria in states to score or rate the evidence also varies.

The primary focus of our interview is the performance measures being used by states. Our interests include a description of any scoring rules, scoring domains, or rubrics that you use. We specifically hope to develop a set of case studies that describe:
a. how each state developed or decided on the measures or criteria used;
b. if, how, and why they have changed the criteria over time; and
c. how the criteria are used for improvement of programs and outcomes for students with significant disabilities.

How was this approach developed?
Can you describe the process?
What practice or research base was consulted as you developed these criteria?
Were the decisions controversial?
Were there additional criteria discussed but not chosen?
What were they and why were they not used?

Have the _________ (rubric, rules, scoring domains…) changed in any way from earlier years? How has it changed? Why has it changed? What occurred that made you consider a change.

What do you see as the potential that these criteria, and any changes made, will affect quality of programs and outcomes for students with significant disabilities? How will that occur? Do you have data that reflect any changes so far, either anecdotal or formal? Do you have this information in written form, or who can I contact to be sure we have an understanding of that information?

What advice would you give other states as they reflect on the criteria they use for their alternate assessment?

What are the key "lessons learned" from your experience?

Appendix B

Student Criteria Definitions and Examples

(from state documentation AND from interview summaries)

Accuracy (number correct/incorrect)

Oregon: Extended assessments in reading, mathematics, and writing

On the Extended assessments in reading, mathematics, and writing, the number correct (or partially correct) defines accuracy. That definition includes a reading rate per minute component and in oral reading there is reading accuracy. According to state staff, it is possible that extended reading, math and writing are actually not alternate assessments but instead are a downward extension of the general assessment.

Student performance on embedded content standards (not dependent on teacher developed linkages)

Louisiana

The main focus of the design of Louisiana's alternate assessment has been on linkage to the content standards and access to the general curriculum. The state provides the linkage through skills in three participation levels. The driving forces for scoring criteria, however, remain in the areas of generalizability and independence rather than on student performance on embedded standards.

Oregon: Extended assessments in reading, mathematics, and writing and Extended CLRAS

Oregon builds the linkages to content standards into its assessment instruments; it does not score the standards linkage. It was assumed in the development of Oregon's alternate assessment that students could move out of the alternate and into the general assessment. With this assumption and the assumption that all students would work toward the same standards, there was no worry about whether standards would be included as part of the alternate assessment measurement system. Right after the IDEA reauthorization, Oregon's state board had just finished adopting standards with the cut points. They essentially said, "We're done with standards. We're not going to reexamine these until they've been in place for a while. We're going to stay the course and see what standards-based reform is all about." So, the group working on the alternate assessment was committed to not developing any new standards, but using with the ones in place for the state. The state Board of Education had already adopted the Career Related Learning Standards, so those could also be used in the alternate assessment (in the CLRAS), along with the math, reading, and writing standards assessed in the Extended assessments in reading, mathematics, and writing. It is unclear if students who participate only in the Extended CLRAS have standards-based evidence in math, reading, and writing.

Oregon: Extended CLRAS

The Extended CLRAS has one area that Oregon's career standards do not have, motor skills.

That was added because families and teachers said mobility was a key element of career skills for the population.

Level of assistance when defined as student independence, actual performance with defined degree of prompting

Arkansas

Level of assistance is defined as the observed accommodations, adaptations, or assistance used by the student during the performance of tasks. On the definitions within the rubric, there are added natural supports and assistive technology use as well. Each individual task (portfolio entry) is scored for Level of Assistance. It is weighted at one-fourth the weight for performance, one-half the weight for appropriateness, but twice the weight for settings.

Kentucky

Support is defined as degree of peer support, tutoring, or natural supports; use of adaptations, modifications, or assistive technology that demonstrates progress toward independence. This could be identified as a student measure or a system measure, depending on how the scoring process looks at actual performance or evidence of provided supports. Kentucky staff clarify that they are looking at this as a system measure, and not a student measure.

Louisiana

The criteria address whether a child can generalize a skill and do it independently without a prompt: As state staff suggest, "We're really measuring two different things with that rubric." As scores increase, level of prompting is joined by degree of generalizability. A student who performs a skill over and over without a prompt but only in one setting and only for one purpose could not get above a "3." The teacher has a period of five weeks to observe the skill. If the student performs the skill for the first time without prompting, the observer would score a 3. If the observer then takes the student to a different setting and structures a task to show the same skill but for a different purpose, and the student performs it, the observer would score a 5. This could happen all on the same day, or it may take many observations before the student achieves the 5. In other words, once the skill has been performed in a second place and for a second reason, the student does not have to perform it again during the five week period.

Oregon: Extended CLRAS

The Extended CLRAS is instructionally based, so students have continued instruction and are measured in areas where they are low in independence. As students become more independent, more environments become available to them. A high value is given to having students learn in natural environments, and this is reflected in scoring. Scoring of the Extended CLRAS puts emphasis on prompt fading. Scoring processes do look for generalization, and that interacts with prompting. If the student does seven routines, and "social greetings" is one of the related skills performed, then all those seven environments where that would be an appropriate place to exhibit that skill. One example is eating lunch in the cafeteria, including going through the cafeteria line. If the student says, "hi" to the cafeteria lady, that's independent. If the teacher raises her eyebrows and tips her head toward the cafeteria lady, that might be a visual or verbal or gesture prompt. The guide includes instructions to the scorers about what to watch for. Generalization is embedded in the scoring process.

Vermont

Incremental fading of supports is one of the categories demonstrating progress and mastery. The student shows progress on a single skill by needing less and less support from adults or classmates. Mastery occurs when the student can perform the skill alone. An example from money skills: the teacher pays for the student's lunch, then the student pays for lunch with the teacher's help, then with a classmate's help, then all alone.

Multiple purposes when defined as student performing skill for multiple purposes

Louisiana

Multiple settings and multiple purposes are both included in the rubric in order to measure generalization. A student who performs a skill over and over without a prompt but only in one setting and only for one purpose could not get above a "3."

Mastery

Arkansas

Mastery is included on the gradations of quality on the performance criterion.

Vermont

Mastery is defined as the ability to perform a targeted skill or ability independently, or using natural supports and assistive technology that permit independence.

Performance

Arkansas

Performance is defined as the student's demonstration of skill while attempting a given task, and at the highest quality gradations, includes mastery, multiple settings, and multiple occasions. Each individual task (portfolio entry) is scored for Performance. The performance domain weight is twice the weight given to appropriateness, four times the weight given to level of assistance, eight times that of settings.

Kentucky

Performance is defined as student progress on specifically targeted IEP goals/objectives that are meaningful in present and future environments, with age-appropriate products.

Progress

Kentucky

Progress is included in the dimension "Performance" which is defined as student progress on specifically targeted IEP goals/objectives. Progress is embedded; it is not a separate scoring criterion. Progress is also part of the "Support" dimension, with use of adaptations, modifications, or assistive technology showing progress toward independence.

Oregon: Extended Academic Assessments and Extended CLRAS

Oregon is committed to a growth model of achievement throughout the whole assessment system. Every assessment director has been committed to using a growth model for both standards and progress. They believe that students who are low on the performance scale but are making great progress should contribute positively to a school's performance rating. They believe that a growth model is good for each student's self esteem, for schools wanting to include them, and fair for the teachers who worked hard to get those students to make progress. So, despite the importance of the growth model in Oregon, the concept of measurement of "progress" is different in Oregon than in the other sample states.

Vermont

"Progress" is defined as development, improvement, or positive changes in the student's performance in relation to designated learning outcomes, including incremental skill development, fading of supports, or generalization to more natural settings.

Settings when defined as student performing the skills in multiple settings

Arkansas

Multiple settings and occasions are included in the quality gradations in the performance criterion. It is included as a system measure in the single score for settings in the scoring rubric.

Louisiana

The criteria address whether a child can generalize a skill and do it independently without a prompt. As state staff suggest, "We're really measuring two different things with that rubric." As scores increase, level of prompting is joined by degree of generalizability. A student who performs a skill over and over without a prompt but only in one setting and only for one purpose could not get above a "3." The teacher has a period of five weeks to observe the skill.

Oregon: Extended CLRAS

Scoring processes do look for generalization. If the student does seven routines, and "social greetings" is one of the related skills performed, then it is scored in all seven environments where that would be an appropriate place to exhibit that skill. One example is eating lunch in the cafeteria, including going through the cafeteria line.

Vermont

Generalization to less restrictive environments is one of the categories for determining progress and mastery–in this situation the student demonstrates progress by using a skill in "real world" situations.

Appendix C

System Criteria Definitions and Examples

(from state documentation AND from interview summaries)

Appropriateness (age, challenge, authenticity)

Arkansas

Appropriateness is defined as the degree to which the tasks: (1) reflect meaningful, real-world activities with age-appropriate materials, (2) provide a challenge for the student, (3) promote increased independence, and (4) are linked to the Content Standards. The domain of appropriateness is weighted at half the weight of performance, twice the weight of level of assistance, and four times that for settings.

Kentucky

Performance is defined as student progress on specifically targeted IEP goals/objectives that are meaningful in present and future environments, with age-appropriate products.

Louisiana

Louisiana has provided guidance to teams on specific skills at three participation levels. Thus, appropriateness is embedded in the state specified skills; they do not score appropriateness. The participation level is chosen by the test administrator, with IEP team input, based on appropriate yet high expectations for each student. The test administrator may determine that the student be assessed at any one of the three levels. However, if the test administrator chooses skills at the Comprehensive or Fundamental level and the student cannot perform at the selected level; then the less difficult level is attempted. "No Performance" can be scored at only the Introductory Level. In each content area, the state provides two state specified skills for each of two target indicators at each of the three levels. Guided by IEP team decisions, teachers determine skills for the remaining target indicators.

Oregon: Extended CLRAS

Oregon has built into the selection of routines a process for IEP teams to grapple with appropriateness, and to avoid any arbitrary "developmental sequence" of routines. Oregon does not score on appropriateness. Oregon developed the list of routines within the categories of the Career Related Learning Standards adopted by the State Board of Education for all students. The individual routines were validated for students with significant disabilities through an in depth analysis of life skills curricula collected from Oregon and the professional literature. After this cross-validation the routines were ranked by teachers and families in terms of what was most important. There is a suggested order in the assessment that is based on, "if you don't know what to select next, here's what the field has said is the most important thing next." For example, disembarking from the bus and arriving at school and eating independently ranked high because those were personal management pieces that the teachers and the parents said were really important. Some teachers did not want to have a prescribed order, and some did not want to be left with no idea if no one identified a preference.

Vermont

Vermont provides some guidance about the appropriate learning outcomes, but it does not score on appropriateness. According to state staff, in the determination of progress the "Incremental Achievement of Skills" probably rewards evidence of increasing challenge.

Content Standards linkages

Arkansas

Entries may be considered nonscoreable if the rules or guidelines in the Implementation Guide are not followed. Entries/portfolios that are considered nonscoreable will receive a score of ZERO; one of the rules that results in a ZERO is that it is determined as "Not to Standard: The entry does not reflect a standard from the curriculum framework."

Kentucky

The Kentucky Academic Expectations are the driving force behind instruction, entry evidence, and portfolio products. Scoring processes look for the degree of linkage to the 54 Academic Expectations identified for all students and assessed by the general assessment, with 28 being prioritized for Alternate Portfolio.

Vermont

The first thing scorers look for is whether the IEP that has been sent in is linked to state standards or the set of learning outcomes from the COACH (Giancreco, et al, 1998). Even in the lifeskills portfolios students are working toward state standards, but they are a different set than their classmates are working toward. If the IEP is linked, then the scorers look to see if there are one or more progress measures in the portfolio that demonstrate a student's progress. One of the things the COACH authors did was a cross-match between the learning outcomes and Vermont's state standards – so that is not something the teachers have to figure out on their own, but they do have to create the evidence of learning, and embed the standards-based learning in the IEP.

IEP linkages

Arkansas

Arkansas has implemented a major training effort to improve the quality of IEPs for student with significant disabilities, but does not include IEP linkages as a separate criterion for scoring its alternate assessment. The training was designed to be compatible with alternate assessment processes, but a decision was made not to directly link it to the alternate assessment. Arkansas is finding that as a result of the training and support offered to teachers on how to write standards-based IEPs with high expectations, the quality of portfolios improves dramatically.

Kentucky

As part of the definitions in the performance dimension, they include progress on specifically targeted IEP goals/objectives. However, Kentucky does not include IEP linkages as a separate criterion, noting that there are IEP goals for individual students that are not reflected in the alternate assessment because the assessment deals solely with standards. Some IEP goals are not connected with general curriculum standards, nor do they have to be, but it is expected that all students have some goals that are clearly standards-based. Kentucky includes methods for targeting IEP goals/objectives for assessment in their *Teacher's Guide* and online training materials for KY teachers.

Louisiana

In their documentation and in their interview, Louisiana emphasizes the shift to standards-based IEPs that raise the bar for all students. However, Louisiana does NOT include IEP linkages as a separate criterion. The state suggests that IEP teams need to target the level at which a student is functioning. Some start very high, some start too low, and some are just right. Teachers are not used to doing assessment in this formal way, so it is taking a while to get teachers to learn how to find the appropriate level of challenge. Trainers have heard teachers say, "I'm thinking about how to make students more independent and I'm working with my aides on ways to fade prompts and have students use more natural cues." A note from a regional director said, "We clearly see how this is showing independence and increases student ability to generalize."

They report that educators are now really looking at IEPs and making sure they are linked with the content standards. This leads to the link back to instruction. Trainers got a sense from some of the folks that the alternate assessment validated what they were doing. Teachers are beginning to understand, for example, that there are math skills involved in "going out for lunch." Teachers are also seeing this as a way to "join the rest of the schooling world." Students are beginning to have access to more environments. Because of the rubrics that say, "Can they do this skill in a different place?" The teacher who wants to be able to take students off campus, now gives further weight to obtaining all those permissions that teachers sometimes need from their administrators. They tell their principals they will get a higher score on their state assessment if the students go out and do some of these things. The trick is to help educators to stay focused on the standards-based knowledge and skills while they do those rich activities that allow students to learn and to show what they have learned. The requirement that the efforts be documented is helping educators to stay focused.

Oregon

Oregon: Extended Academic Assessments and Extended CLRAS:

A goal in Oregon is to help teachers administer the Extended Reading, Extended Mathematics, And Extended Writing assessment and the Extended CLRAS in conjunction with preparing for an IEP meeting – to use the tests to get information for instructional planning. One of the benefits of the Extended CLRAS is that it gives teachers an "ah ha" very quickly about what they might include in a student's IEP, what the instruction might look like, and how they might do the teaching. It ends up being a teacher training tool indirectly. The extended assessments in reading, mathematics, and writing had that effect as well. For example, a high school teacher said, "I didn't even know he could read" after the teacher was required for the first time to actually assess the student's reading in the extended assessment. In that sense both types of alternate assessment carry a strong message according to state staff.

Oregon: Extended CLRAS

The routines are anchored in environments, but the related skills are assessed at the same time as the routines assessment. For the related skills there is a manual that defines the skills in terms of expressive communication and receptive communication; those are to be taken from the student's IEP, looking for those skills to be exhibited in the context of the routines.

Vermont

The scoring criteria in Vermont are based on students' IEPs and progress measures that are designed to measure students' progress against IEP goals that are referenced to the learning outcomes or the corresponding Vermont Standards. The learning outcome can be quoted directly, paraphrased, or can be referenced to the corresponding Vermont Standards listed in the rubric. The outcome might be embedded in a larger goal or outcome. In the field studies, we discovered that the IEPs for students with severe disabilities were not very good. One goal of the assessment system was to get people to write good goals for students with severe disabilities and assess students' progress toward those goals.

Self-determination when defined as evidence of system providing choice

Kentucky

Self-determination is defined as evidence of degree of student choice with an impact on student learning across and within entries, and evidence of planning, monitoring, and evaluating of the self-performance to improve or set new goals. State staff see this as a system measure, not a student measure because self-determination is an instructional area for students with severe disabilities.

Vermont

Vermont state staff point out that although they don't have specific scoring criterion on self-determination, the ability to make choices is embedded throughout their learning outcomes. For example, one of the communication learning outcomes is, "Uses communication to indicate preferences and make choices."

Settings (multiple) when defined as system offering multiple settings for student

Arkansas

Settings is defined as the observed settings or environments in which tasks are administered/performed. The entire portfolio (not individual entries) is scored for Settings. It is weighted one-eighth of performance, one-fourth of appropriateness, and one-half of level of assistance. Settings is also a student criteria, included in the definition of "performance."

Kentucky

Settings are defined as the number and variety of settings in which there is evidence of student performance, and the number of entries in which they are observed. Those settings are counted only in which the student is integrated with age appropriate peers (also, see social relationships).

Social relationships when defined as evidence of system provided opportunities

Kentucky

The "Social Relationship" dimension is scored on appropriateness of interaction, frequency, diversity of range of peers, sustainability, and reciprocity of relationships.

Support (access to assistive technology, peer supports, adaptations, accommodations)

Kentucky

Support is defined as degree of peer support, tutoring, or natural supports; use of adaptations, modifications, or assistive technology that demonstrates progress toward independence. Kentucky views this as a system measure, but this could be identified as a student measure depending on how the scoring process looks at actual performance or evidence of provided supports.

The College of Education
& Human Development

UNIVERSITY OF MINNESOTA

NCEO is an affiliated center of the Institute on Community Integration