

***PROPOSED AMENDMENTS JUNE 2004 (Middle column)***

**I Say Potato, You Say Potahto:  
The Assessment-speak Gap  
Between General and Alternate Assessment Experts**

**A SIDE-BY-SIDE GLOSSARY**

**Joseph M. Ryan  
Arizona State University West**

**Rachel F. Quenemoen and Martha L. Thurlow  
National Center on Educational Outcomes  
University of Minnesota**

**April 15, 2004**

Paper presented at the annual meeting of the American Educational Research Association (AERA). It was prepared, in part, by the National Center on Educational Outcomes through a Cooperative Agreement (#H326G000001) with the Research to Practice Division, Office of Special Education Programs, U.S. Department of Education. Points of view or opinions expressed in the paper are not necessarily those of the U.S. Department of Education, or Offices within it.

<b>Assessment Term, Concept, or Procedure</b>	<b>Traditional connotations as used in assessing students with significant cognitive disabilities</b>	<i>Meeting in the middle: Building on the expertise of all partners</i>  <i>Implications, insights, and inspiration</i>	<b>Traditional connotations as used in assessing students in the general education population</b>
<b>Population</b>	<p>Very small group of students (dozens in a state)</p> <p>State-to-state variation of students who take alternate assessment/aa, multiple alternate assessments/aa, pressure from 1% Rule</p> <p>This is a highly variable population in terms of learner characteristics, available response repertoires, and often competing complex medical conditions</p> <p>“Outliers” can be a large proportion of this very small population</p>		<p>Tens or hundreds of thousands of students</p> <p>Rules for inclusion and exclusions vary across time and setting</p> <p>Often homogeneous in the aggregate with respect to what is being measured (e.g., the construct has the same meaning for most students although students may vary in amount of knowledge/skill).</p> <p>“Outliers” who are not homogeneous are a relatively small proportion of the large population</p>

<p><b>Construct domain</b></p>	<p>The applicable construct domains for students are often defined individually, through extended content with flexible access points</p> <p>Generally refers to observable behaviors related to performance of content related skills and knowledge</p> <p>May be defined through IEP process in states that are training on standards referenced IEPs; may involve progress on standards-based IEP goals</p> <p>No consensus theory of learning exists in the academic content areas for these children, that is, what patterns of growth they show on the path to competence</p>		<p>State standards generally define grade-level construct domains all students</p> <p>Defines learning targets in terms of content, cognitive processes, and performance</p> <p>Usually emphasizes content, also refers to cognition, e.g., remembering, comprehending, applying, and more complex processes</p> <p>Specifies the boundaries (what's in and what's not), structures, and relationships among elements</p>
<p><b>Assessment Format: Tests and Items</b></p>	<p>The majority of states use portfolio, body-of-evidence, or other performance-based models for their alternate assessments on alternate achievement standards (aa)</p>		<p>Test are generally given under standard conditions in terms of content, format, timing, and response mode</p> <p>A common test blueprint is</p>

	<p>Most alternate assessment assessments/aa include relatively few open-ended tasks that are often tailored to the individual student.</p> <p>In many states, teachers are trained to design assessment tasks to demonstrate student knowledge and skills, often embedded in ongoing instruction</p>		<p>used across test forms</p> <p>Item formats vary widely and include selected response, short answer, extended open-ended task and/or response, complex constructed response, and performance assessments</p> <p>Item formats vary to reflect the different learning objective being assessed</p>
<b>Generalization / Generalizability</b>	<p>Given the limited understanding of the construct domain, and lack of consensus on a theory of learning in the academic content for these students, and the varying coverage of the domain, generalization as traditionally defined in measurement is a challenge</p> <p>The term “generalization” is a foundational term used by special educators, and is a common scoring criterion, meaning: <i>Student performance of skills or knowledge learned in one</i></p>		<p>Assessments should provide representative coverage of the construct domain content and processes so that score interpretation is not limited to the sampled tasks on the specific assessment</p> <p>Generalizability is usually considered an aspect of validity although the “consistency” connotation reflects the concept of reliability</p> <p>Generalizability studies are rarely part of local assessment programs and are not always</p>

	<i>setting or for one purpose is evidenced in additional settings or for different purpose</i>		included in state programs.
<b>Reliability</b>	<p>Often refers to whether a student can demonstrate the same behavior two or three times, or through triangulated data sources</p> <p>Cannot be easily quantified in terms of classical test theory concepts of true and error scores</p> <p>Some states report inter-rater reliability statistics as one indicator of reliability for alternate assessments. Although reporting the consistency of scoring processes is valuable, reporting inter-rater agreement statistics as if they are reliability coefficients is misleading</p>		<p>Usually refers to consistency in response to items, which are viewed as sampled replications from a construct domain</p> <p>Used to evaluate inferences about the likelihood that students would perform similarly on the same or parallel form of the assessment</p> <p>Easily quantified in indices of internal consistency, alternative form, tests-retest reliability</p>
<b>Error of Measurement</b>	Very difficult to index because of small sample sizes and narrowly defined behavioral domains		Provides a quantification of the amount of error that can be expected in students' scores

			<p>Used to establish confidence intervals or bands within which students' true scores are known with a specified level of probability</p> <p>Straight forward in both classical and IRT approaches</p>
<p><b>Validity</b></p>	<p>Some validity studies have looked at the process used in alternate assessment design in states, specifically around defining the scoring criteria. Stakeholder agreement on criteria reflecting achievement for students with significant disabilities then shapes the design of the alternate assessment.</p> <p>A few studies have looked at concurrent validity of alternate assessment scores against other measures of quality programming and outcomes for students with significant disabilities</p> <p>Current work is being done</p>		<p>An integrated evaluative judgment about the degree to which evidence and theory support the adequacy and appropriateness of inferences and actions based on assessment information (Messick, 1989, p 13)</p> <p>In most settings, validity rest largely on demonstrating that the assessment reflects the content standards it is designed to measure.</p> <p>The degree to which items reflect the content standards is usually assessed by a content review panel.</p> <p>Evidence about adequate care</p>

	<p>on content validity or at a minimum, alignment of extended standards to general standards, and ultimately to the alternate assessment.</p> <p>Documentation of adequate care and implementation of recognized procedures in setting of alternate achievement standards has occurred in a few states. There is limited understanding in special education of what setting standards involves, what it means</p> <p>Correlational studies have documented rapid shifts in instruction and curriculum in the desired directions in several states through teacher surveys and observational protocols</p>		<p>and implementation of recognized procedures in the item and test development processes often is used as validity evidence</p> <p>Evidence about adequate care and implementation of recognized procedures in the setting performance standards often is used as validity evidence</p> <p>Correlations with external variables (convergent and divergent) are frequently used as validity evidence</p>
<p><b>Fairness</b></p>	<p>A layman’s version of a fairness discussion is a common aftermath to the</p>		<p>Often seen as an aspect of validity</p>

	<p>first year of alternate assessment/aa. These discussions focus on whether these assessments measure the skill of the teacher or the skill of the student, whether scoring processes are of high quality and are applied consistently, and whether it is appropriate or desirable to expect these students to learn academic content</p> <p>Generally, the discussion is focused on how unfair the new assessments are to teachers. States respond with additional training support in many cases, although some states have reduced requirements considerably in the face of the outcry</p> <p>The accountability requirements of NCLB may change the nature of fairness discussions</p> <p>Proponents of alternate assessments/aa suggest that OTL is the major fairness</p>		<p>Deals specifically with evaluating assessments for bias, meaning that tests scores are influenced by factors irrelevant to the construct being measured</p> <p>Generally examined through studies of differential item functioning (DIF)</p> <p>Bias-sensitivity panels review assessment items and task for any offensive features and for opportunity to learn (OTL) as a standard element in test development</p> <p>Sources of construct irrelevant variance (e.g., language skills in math or social studies) are also examined judgmentally and empirically</p> <p>OTL is an aspect of fairness that is examined in some assessment programs</p> <p>Assessment data are disaggregated and the validity of the assessment for each</p>
--	--	--	--

	issue for this group of children		subgroup is considered
<b>Test Administration</b>	<p>These assessments tend to be individually tailored to the response repertoire of the individual student. The content, items, format, timing, and response mode are all individualized</p> <p>Level of challenge is a criterion in several states on which the evidence is scored</p> <p>A few states have developed common tasks with flexible modes of response, scoring on level of prompting needed before a student can respond</p>		<p>The critical feature of test administration is that tests are generally given under standard conditions in terms of content, items, format, timing, and response mode</p> <p>In most cases, students take exactly the same test or a form that is equivalent in content and difficulty</p> <p>In a few instances, like NAEP, students take a subset or sample of items but in such cases individual scores are not reported</p> <p>Amount of time student have to take the tests may vary from a fixed period to un-timed conditions</p>
<b>Scoring</b>	Performance assessments are scored against carefully developed standards-referenced rubrics applied by trained raters in many states.		<p>Selected-response questions are machine-scored against a key</p> <p>Short answer, extended</p>

	<p>The scoring rubrics reflect the task and content domain structure and, thus, are part of the validity evidence</p> <p>The raters are trained to a mastery criterion and then check papers, read behinds, and rater agreement indices are employed to monitor scoring</p> <p>Some states have regional certified scorers administer the tasks or checklist, or they may document the evidence supporting teacher scoring in a sample of cases</p> <p>Other states permit teacher scoring and reporting of student performance. Some require a sample audit; others rely on teacher judgment</p>		<p>response, and other performance assessments are scored against carefully developed standards-referenced rubrics applied by trained raters</p> <p>The scoring rubrics reflect the task and content domain structure and, thus, are part of the validity evidence.</p> <p>The raters are trained to a mastery criterion and then check papers, read behinds, and rater agreement indices are employed to monitor scoring</p>
<b>Interpretation</b>	<p>In a few states, student performance is interpreted relative to achievement standards resulting in students being classified into various achievement levels</p>		<p>Student performance is interpreted normatively (percentiles, stanines, etc)</p> <p>Student performance is interpreted relative to</p>

	<p>Achievement standards are generally based on panel review of score patterns and student work, and cutscores are selected using various recognized procedures</p> <p>Achievement levels often have substantively rich descriptions that aid in interpretation</p> <p>Concern focuses on “How high is high enough,” challenge, appropriateness</p>		<p>performance standards resulting in students being classified into various achievement levels</p> <p>Performance standards are generally based on an examination of item content when cutscores are selected using various recognized procedures are employed</p> <p>Achievement levels often have substantively rich descriptions that aid in interpretation</p>
<p><b>Consequence</b></p>	<p>Consequential validity is the primary area of study of the effects of alternate assessment on alternate achievement standards.</p> <p>Correlational studies have documented rapid shifts in instruction and curriculum in the desired directions in several states through teacher surveys and observational protocols.</p> <p>These are students who in</p>		<p>Often incorporated as an aspect of validity</p> <p>Involves examining the intended and unintended consequences of the intended assessments use</p> <p>Not always evaluated</p> <p>The impact of an assessment applications that has a specific purpose (e.g., identify students in need of remediation) should be</p>

	<p>many cases have had no access to the general curriculum.</p>		<p>examined to see if the impact is achieved (e.g., did the students receive remediation)</p> <p>Assessments designed to yield information to be used in educational decisions should be examined to determine what, if any, role the results play in decisions making.</p> <p>Unintended outcomes should be examined to determine if they are related to characteristics of the students that are not related to the construct being measured.</p>
--	---	--	---