Content-Related Evidence: Alignment of Alternate Assessments with Standards

The purpose of this chapter is to provide states a process for aligning their alternate assessments with grade level content standards. The reason for the focus on alignment in establishing content-related evidence is the primacy of state standards in driving all test blueprints. This alignment is critical for alternate assessments because of the changes made in reductions of depth, breadth, and complexity and the need to ascertain whether the constructs contained in the original (grade level) content standards are maintained with integrity and distributed appropriately in coverage.

In this chapter, three topics are addressed: (a) alignment of alternate assessments with standards, (b) inferences about achievement standards, including alignment of extended standards with grade level content standards as well as alignment of alternate assessments with (extended) grade level content standards, and finally (c) steps to follow in the alignment process.

Alignment of Assessments with Standards

Research on evidence based on test content was primarily between tests, curriculum, and instruction in the 1970s and 1980s; now, this evidence needs to take into account state grade level content standards. Rather than ascertaining whether or not tests and curriculum (or instruction) have common items, both (or all three) need to be analyzed with respect to their alignment with standards.

Alignment at its simplest level is close in definition to the term ‘overlap’ that had been present in the early work from the 1970s through the 1980s. Alignment, however, begs the question of “to what?” or “with what?” In our review of alignment, new dimensions also have appeared that are distinct from the earlier evidence based on test content: It is not only standards-based but also systemic in nature and prospective in development. These three features have the potential for (a) bridging evidence based on test content to a more unified view of validity and (b) focusing our attention on *inferences* as the bedrock for validity. As a consequence, curriculum and instruction may be viewed in the context of standards and assessment. Nevertheless, we still need to be cautious in making the assumption that standards, once enacted interactively in instruction, reach into the classroom with equal parity and consistency in defining what is taught or how it is taught. If this is generally true, then our accountability systems can be trusted; if this is not true, then confusion exists between the outcome and the inference.

Alignment of assessments to *standards* is usually the starting point for most researchers in this area. For example, consider the following three definitions. “Alignment refers to the degree of match between test content and the subject area content identified through state academic standards. Given the breadth and depth of typical state standards, it is highly unlikely that a single test can achieve a desirable degree of match. This fact provides part of the rationale for using multiple accountability measures and also points to the need to study the degree of match or alignment both at the test level and at the system level. Although some degree of match should be provided by each individual test, complementary multiple measures can provide the necessary degree of coverage for systems alignment. This is the greater accountability issue” (La Marca, 2001, p. 1).

More recently, Bhola, Impara, & Buckendahl (2003) stated that "alignment can be defined as the degree of agreement between a state's content standards for a specific subject area and the assessment(s) used to measure student achievement of these standards" (p. 21). Finally, Webb (1997) defines alignment as "the degree to which expectations and assessments are in agreement and serve in conjunction with one another to guide the system toward students learning what they are expected to know and do" (p. 4).

As noted by Webb (1997) in the quote above, another critical feature of alignment (which was not present with test content validity from the 1980s) is a **systemic focus**. This holistic view also is endorsed by La Marca, Redfield, Winter, Bailey, and Hansche (2000):

1. "Alignment is a match between two or more things. *Webster's New World College Dictionary* defines *align* as 'to bring into a straight line; to bring parts or components into a proper coordination; to bring into agreement, close cooperation.' In an aligned system of standards and assessments, all components are coordinated so that the system works toward a single goal: educating students to reach high academic standards" (Hansche, 1998, p. 21).
2. Alignment refers to how well all elements in a system work together to guide instruction and student learning (Webb, 1997).
3. Alignment directly affects the degree to which valid and meaningful inferences about student learning can be made from assessment data (Long & Benson, 1998).

Alignment is more **prospective** rather than retrospective. In the previous research on evidence test content, most tests provided broad surveys of several curricula that were post-hoc analyzed for overlap; but these tests were not designed from the outset with specific planning around any particular curriculum. La Marca (2001) suggests that sound standards and assessment development activities create alignment when the following three conditions are present: (a) as content standards are developed, assessment design should be considered (determining what and how to measure achievement), (b) items and tasks should be designed to measure specific objectives as outlined by academic content standards, and (c) a post hoc review of alignment should be conducted following assessment development.

Eventually, alignment as a construct needs more operational definition. La Marca, Redfield, Winter, Bailey, and Hansche (2000) articulate the following categories to consider in reviewing the alignment of assessments with standards.

1. *Content match.* When evaluating content match between standards and assessment, one should consider whether assessments are designed to match the content standards, whether all items and tasks are related to the content standards, and whether the assessment fully covers the content standards.
2. *Depth match.* When evaluating depth match between standards and assessment, one should take into consideration whether both the assessment as a whole and items/tasks are at a level of difficulty matching that is prescribed by content standards, and whether item/task specifications both indicate the depth at which knowledge should be measured and elicit responses reflecting the depth of knowledge they measure.
3. *Emphasis.* Assessment items and tasks should measure knowledge and skills representative of those in the content standards in order for the assessment and content standards to have emphasis match.

4. *Performance match.* When evaluating performance match, one should consider whether the assessment blueprint specifies: (a) how the entire range of performance descriptors will be measured by the assessment, (b) whether item specifications are referenced to the levels of knowledge and skills in the performance descriptors, (c) whether the assessment as a whole covers knowledge and skills at each defined performance level, and (d) how each aspect of the performance descriptors is covered by one or more items/tasks.

5. *Accessibility.* Accommodations (and modifications) should be available for students with disabilities (SWD) and English Language Learners (ELL). Groups of selected-response items should cover a variety of ways of expressing knowledge and skills related to the content standard(s) and the assessment should be free of irrelevant factors that are likely to interfere with students' opportunity to demonstrate knowledge/skills in order to have accessibility.

6. *Reporting.* To evaluate reporting, one should take into consideration whether score reports clearly illustrate levels of student proficiency on content standards, whether reports contain information that can be used to make valid inferences and decisions, whether they provide information about the standard error of measure regarding reported scores, and whether the reported information can be applied for the intended purpose(s) of the assessment.

Making Inferences about Alternate Achievement Standards

In April 2006, the U.S. Department of Education released a 'Toolkit on Teaching and Assessing Students with Disabilities" (http://www.osepideasthatwork.org/toolkit/index.asp). This web site provides the public with a number of useful papers and procedures, among the most critical of which is a series of eight documents within *Models for Large-Scale Assessment for Students with Disabilities*. These documents provide a rationale to guide the alignment process.

In the *Executive Summary*, the following definition is made for making explicit the inference from alternate achievement standards (Technical Work Group, 2006):
"**Alternate achievement standards** are designed to enable inferences to grade-level expectations that have been extensively prioritized but maintain high expectations for progress in the general curriculum and assume student performance is contingent on having the supports specified for the assessment. Inferences are **stipulated because of the assessment methodology**.

Although a few different systems are available (using Achieve, Webb, or the enacted curriculum), the alignment process essentially addresses the question of which standards are addressed in the assessment, how broadly the objectives within the standards are covered, how consistent is the depth of knowledge between the standards and the assessment items, and what are the sources of challenge. The process is entirely judgmental and is best completed with appropriate group of teachers (e.g. content experts in general education as well as special education teachers).

Developing Alternate Assessments

Generally, states have taken either of two approaches to guide development of alternate assessments: (a) grade-level content standards are *essentialized* (extended or expanded) and alternate assessment items are designed to reflect them, or (b) grade level content standards remain intact and the alternate assessments are reduced directly in their breadth, depth, or complexity. Following is an explanation of these two approaches.

Aligning extended standards to grade level content standards. When 'essentializing' grade level content standards, the focus is on adapting the grade level standards to reduce the breadth, depth, or complexity of the standard. In this process, the essential verb of the standard is 'translated' to be less encompassing.

In most state content standards, the verbs can be considered as concepts. Although all concepts have three components (a label, attributes, and examples and non-examples), it is the attributes that help constrain the range of examples and non-examples.

For example, in the following standard, three key verbs are used: Students explain their choice of estimation and problem-solving strategies and justify results when performing number operations with fractions and decimals in problem-solving situations. In this example, the key verbs are underlined above: justify, explain, and estimate (within context of fractions and decimals). This standard could be 'essentialized' as follows. See Figure 1.

Figure 1. Example for Creating Extended Standards from Grade Level Content Standards

Attributes for Grade Level Standard	Essentialized Attributes for Extended Standards
Explain and Justify • Establish a procedure or process • Use multiple approaches to solving the problem • Use steps that are logical or empirical • Describe the steps (procedures or process)	Explain and Justify • Use any kind of *steps* in arriving at a solution • Describe the steps (procedures or process)
Estimate • Include multiple numbers (though one may be a constant) • Provide a probable answer with a ballpark solution • Use a strategy (that might be implicit or explicit) • Employ one of four basic math operations • Express the answer in units of measurement	Estimate • Employ one of four basic math operations • Provide a probable answer with a ballpark solution

One way to systematically evaluate this 'essentialization' of a grade level content standard is to use a rating scale that denotes the level of constraints made in reducing the breadth, depth, or complexity of the attributes. See Figure 2.

Figure 2. Rating Scale for Evaluating Reduction of Breadth, Depth, Complexity of Standards

Rating	Alignment Descriptors
4	The verbs (or context) in the alternate benchmarks reflect the construct in a manner that fully includes all of the attributes.
3	The verbs (or context) adjust the construct in a manner that marginally limits the attributes.
2	The verbs (or context) adjust the construct in a manner that constrains the construct but still reflects some attributes.
1	The verbs (or context) adjust the construct in a substantial (stipulated) manner with no attributes reflected.

Aligning alternate assessments to grade level content standards. Rather than ‘essentializing’ the grade level content standards (changing them from grade level to extended or expanded standards), the focus could be on the alternate assessments and the manner in which they are similar to other items that reflect the standard. With this process, the items on an alternate assessment may be directly constrained in their breadth, depth, or complexity. With this process, the focus is on the domain (or universe) for sampling items that reflect the grade level content.

Figure 3. Rating Scale for Evaluating the Reduction of Breadth, Depth, Complexity of Items

Rating	Alignment Descriptors
4	The content breadth and depth allows generalization to any (all) items in the universe of items for that standard.
3	The content breadth and depth is constructed so generalization can be made to most items in the universe of items for that standard.
2	The content breadth and depth is constricted so generalization can be made only to a limits number of items in the universe of items for that standard.
1	The content breadth and depth is so severely constricted that generalization is not possible to (m)any items in the universe of items for that standard.

Notice that this language is very consistent with the terminology used in the toolkit. The focus in either strategy is to reduce the breadth, depth, and complexity (of the standard or items) and therefore, to constrain the inference that can be made about achievement of proficiency on the (various) standards.

Alignment of Alternate Assessments Using the Webb System (excerpted from CCSSO, 2006)
Given either the alignment of extended standards to grade level content as a strategy for developing an alternate assessment or a systematic process for developing items that are directly designed to reflect less depth, breadth, or complexity, the alignment process may proceed with a number of different critical components. In this example, the Webb system is described, though other systems are certainly available.

1. *Categorical Concurrence* is the degree to which standards and assessments address the same content categories. This criterion is met if both documents display the same or consistent categories of content.

2. *Depth-of-Knowledge (DOK) Consistency* is the degree to which the DOK required by the standards and assessments are in agreement. If the assessment is as demanding as the expectations standards set for the students, this criterion is met. According to Webb's model, depth-of-knowledge is judged at four levels: (a) recall of fact, information, or procedure; (b) skill in using information, conceptual knowledge, or procedures of two or more steps; (c) strategic thinking, reasoning, developing a plan or sequence of steps, complexity, more than one possible answer, requiring less than 10 minutes to do; and, (d) extended thinking, requiring an investigation, time to think and process multiple conditions of the problem or task, and requiring more than 10 minutes to do non-routine manipulations. In the table below, these four levels have been translated using language specifically designed for alternate assessments that employ portfolios, performances, or observations. See Figure 4.

Figure 4. Levels for Depth of Knowledge (using Webb model) and Adapted Model

Level	Webb Description	Alternate Assessment Description
1	Recall and reproduction: Recall and recognition of a fact, information, or procedure	A 'behavior event' with 1:1 correspondence completed in single context.
2	Skill and concept: Use information or conceptual knowledge with 2 more steps	A 'behavioral event' with more than 1:1 correspondence in more than one context with correct or incorrect responses.
3	Strategic thinking: Requires reasoning, developing a plan or a sequence of steps, some complexity, more than one possible answer (non-routine problem-solving)	A multiple step 'behavioral event' executed in more than one context with more than 1:1 correspondence and with partial correct scoring of responses.
4	Extended thinking: Requires an investigation, time to think and process multiple conditions of the problem (e.g. completing a project, including how to design and execute it.	A multiple step 'behavioral event' executed as an approach (of many) to completing a task that occurs in multiple settings.

3. *Range-of-Knowledge Correspondence* is the degree to which the span of knowledge a standard expects of students matches that required to correctly answer the assessment items or activity.

4. *Balance of Representation* is the extent to which assessment items are evenly dispersed across learning objectives within a standard.

*Definition of an item***.** Before applying this model to an alternate assessment that uses a portfolio or observation, however, the notion of an item needs to be opertionalized. In the adapted model in Figure 4, the term **'**behavioral event' is used. A behavioral event has four attributes: (a) it reflects routines (with a beginning/middle/end), (b) it is captured in one session (e.g., sitting/setting), (c) it comprises a skill or multiple skills, and (d) it contains one or more items that may incorporate observations of students in different settings (and therefore use rating scales and checklists) or collections of work samples in portfolios or performance tasks.

Given a measurement approach and a suitable definition of an item, it is possible to proceed directly to an alignment study. In the next section, seven steps are followed to ensure that the information from an alignment study is useful in generating formative measures that can guide any improvements to the alternate assessments.

Steps in the Alignment Process

Step 1. List the appropriate standards and objectives on a spreadsheet in the first column.

Step 2. Note the format of assessment and develop a (student) sampling plan (if using a portfolio measurement approach).

a. For performance assessments, fill subsequent columns with task labels. The cell that defines the intersection of each standard and task can be filled in with the number of items for that task. The total number of cells with numbers equals the number of tasks (to be considered hits in Step 5 below).

b. For portfolio or observation alternate assessments, make frequency counts of behavioral events. These counts should appear in subsequent columns for each standard and objective. A behavioral event is any entry that was completed within a sitting/setting. For example, a picture of the student performing a task qualifies as a behavioral event; completing a worksheet also serves as a behavioral event.

Step 3. Count the number of *standards* in which a behavioral event (for an alternate assessment task) appears as a performance task or is part of a portfolio or observation. Calculate the percentage by dividing the total number of standards with behavioral events into the total number of standards. The resulting number reflects the degree of *categorical concurrence*.

Step 4. Evaluate each standard having an associated alternate assessment task or behavioral event for the *depth of knowledge*. These evaluations are listed in the table and can be summarized as percentages: (a) matching, (b) alternate assessments > standards, or (c) standards > alternate assessments.

Step 5. Calculate the *range of knowledge* by counting within each standard the number of objectives having associated behavioral events or alternate assessment tasks divided by the total number of objectives in that standard. This percentage is computed within each standard and can be averaged across the standards.

Step 6. Calculate the *balance of representation* for each standard in which objectives have associated alternate assessment tasks or behavioral events (referred to as a hit). The easiest system for calculating balance is to use a spreadsheet and in successive columns (one for each standard) place sufficient rows to equal the number of objectives in that standard. With each objective, calculate the number of tasks in the assessment (those standards with assessment tasks then represent a hit); then the formula for that objective is (1/# hits) – (# tasks for that objective/# total tasks). The absolute value is taken of these values summed, which is then divided by 2 and subtracted from 1. The formula described above is used to ascertain that for this objective, the balance (.78) is sufficient, using definitions articulated by Webb (2002). Balance Index = 1 – (Σ |1/(O) – $I_{(k)}$/(H)|) / 2.

Step 7. Report the results in a table with both the raw results in an appendix and a summary for critical stakeholders.

Figure 5. Example for Balance of Representation (Calculated only on Objectives with Hits)

Standard or Objective	# Items "hit" per objective	1/(O) – $I_{(k)}$/(H)	ABS
1.	3 hits	1/5 - 3/17 =	0.02
2.	4 hits	1/5 - 4/17 =	| - .04 | = 0.04
3.	3 hits	1/5 - 3/17 =	0.02
4.	4 hits	1/5 – 4/17 =	| - .04 | = 0.04
5.	3 hits	1/5 - 3/17 =	0.02

	O = 5
	H = 17
			Σ = 0.14
			0.14/2 = 0.07
			1 – 0.07 = 0.93

			Balance Index = 0.93

O = Total number of objectives hit for the standard
$I_{(k)}$ = Number of items hit corresponding to objective (k)
H = Total number of items hit for the standard
Balance Index = 1 – (Σ |1/(O) – $I_{(k)}$/(H)|) / 2

Once all of these data are collected, they can be displayed in a single table to help guide any changes to improve the alternate assessment. Of course, it is important to develop sensible criteria for interpreting these values. In Table 6 below, it appears that the alternate assessment is not particularly well aligned though any firm interpretations need to be qualified by the design and assumptions of the assessment. See Figure 6.

Figure 6. Example Report of All Webb Dimensions

				Depth of Knowledge		
	Cat. Concur.	Ave. Range	Balance	Std=AA	Std>AA	AA>Std
3	60	.18	.49	60	27	13
4	60	.25	.71	27	47	27

References

Council of Chief State School Officers (2006). *Aligning Assessment to Guide the Learning of All Students: Six Reports on the Development, Refinement, and Dissemination of the Web Alignment Tool*. Washington, D.C.: Author.

The Technical Work Group (April 2006). *Including Students with Disabilities in Large-Scale Assessment: □Executive Summary.* □Eugene, OR: University of Oregon – Behavioral Research and Teaching □and □ Washington, D.C: American Institutes for Research. Retrieved from the World Wide Web on August 28, 2006.

La Marca, P. M. (2001). Alignment of standards and assessments as an accountability criterion. *Practical Assessment, Research & Evaluation, 7*(21). Retrieved June 4, 2003, from http://ericae.net/pare/getvn.asp?v=7&n=21

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

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

Hansche, L. N. (1998). *Meeting the requirements of Title 1: Handbook for the development of performance standards*. Washington, DC: U.S. Department of Education.

Long, V. M., & Benson, C. (1998). Re. Alignment. *The Mathematics Teacher*, *91*(6), 503-508.

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