APPLIED MEASUREMENT IN EDUCATION, *20*(1), 7–25

RESEARCH ARTICLES

Issues Related to Judging the Alignment of Curriculum Standards and Assessments

Norman L. Webb
Wisconsin Center for Education Research
University of Wisconsin–Madison

A process for judging the alignment between curriculum standards and assessments developed by the author is presented. This process produces information on the relationship of standards and assessments on four alignment criteria: Categorical Concurrence, Depth of Knowledge Consistency, Range of Knowledge Correspondence, and Balance of Representation. Five issues are identified—but not resolved—that have arisen from conducting alignment studies. All of these issues relate to making a decision about what alignment is good enough. Pragmatic decisions have been made to specify acceptable levels for each of the alignment criteria. The assumptions are described. The issues discussed arise from a change in the underlying assumptions and from considering variations in the purpose for an assessment. The existence of such issues reinforces that alignment judgments have an element of subjectivity.

With the increased importance imposed by the No Child Left Behind Act of 2001, procedures for determining the alignment of curriculum standards and assessments have gained the increased attention of state departments of education, the federal government, and the measurement community. States are required to show that their standards and assessments meet fairly explicit criteria of alignment (e.g., comprehensiveness, content and performance match, emphasis, depth, consistency with performance standards, and clarity for users). However, how alignment is judged, what is considered to be acceptable alignment, and what procedures are used to de-

Correspondence should be addressed to Norman L. Webb, Wisconsin Center for Education Research, 1025 W. Johnson Street, Madison, WI 53706. E-mail: nlwebb@wisc.edu

termine the alignment between standards and assessments raises a number of issues. The purpose of this article is to use data collected through alignment studies as a basis for identifying and discussing some of the most critical alignment issues.

The Webb (1997, 2002) alignment process is one of a handful of processes that have been used to determine the match between curriculum standards and assessments (Blank, 2002). In general, this process identifies four criteria that are used to compare the relation between standards and assessments. The process is conducted in two stages. In the first stage, reviewers code the depth-of-knowledge (DOK) levels of standards. In the second stage, reviewers code the DOK levels of assessment items and the corresponding curriculum standards or objectives. Reviewers code assessment items directly to the curriculum standards. Findings are reported for each of the four criteria, along with the attainment of specified acceptable levels. The reviewers' entry of coding and the analysis of data have been automated using a Web-based tool (http://www.wcer.wis.edu/WcAT).

Porter and Smithson (Porter, 2002) developed another process for judging alignment, referred to as the Survey of the Enacted Curriculum (see Porter, Smithson, Blank, & Zeidner, 2007/this issue). Central to this process is a content-by-cognitive level matrix. Reviewers systematically categorize standards, assessments, curriculum, or instructional practices onto the matrix, indicating the degree of emphasis in each cell. Comparisons, or the degree of alignment, are made by considering the amount of overlap of cells on the matrix between any two elements of the analysis (assessment and standards, curriculum and standards, standards and instruction, etc.). One difference between this process and the Webb process is that the Survey of the Enacted Curriculum has reviewers map the standards and assessments to a common framework rather than directly to each other.

Achieve, Inc. has developed another process that is based on a group of experts reaching consensus on the degree to which the assessment-by-standard mapping conducted by a state or district is valid. This process reports on five criteria: (a) Content Centrality, (b) Performance Centrality, (c) Source of Challenge, (d) Balance, and (e) Range. For Content Centrality and Performance Centrality, reviewers reach a consensus as to whether the item and the intended objective(s) correspond fully, partially, or not at all. Achieve prepares an extensive narrative to describe the results from the review and will include a "policy audit" of standards and the assessment system if desired. Determining consensus among reviewers distinguishes this approach from the other two, along with centrally incorporating the state's mapping into the process.

WEBB ALIGNMENT PROCESS

Generally, the alignment process is performed during a 3-day Alignment Analysis Institute. The length of the institute is dependent on the number of grades to be an-

alyzed, the length of the standards, the length of the assessments, and the number of assessment forms under consideration. Five to eight reviewers generally do each analysis. The larger number of reviewers will increase the reliability of the results. Reviewers should be content-area experts, district content-area supervisors, and content-area teachers.

To standardize the language, the process employs the convention of standards, goals, and objectives to describe three levels of expectations for what students are to know and do. *Standard* is used here as the most general (e.g., Data Analysis and Statistics). A standard, most of the time, will be composed of a specific number of goals, which are comprised in turn of a specific number of objectives. Generally, but not always, there is an assumption that the objectives are intended to span the content of the goals and standards under which they fall.

Reviewers are trained to identify the DOK of objectives and assessment items. This training includes reviewing the definitions of the four DOK levels and then reviewing examples of each. Then the reviewers participate in (a) a consensus process to determine the DOK levels of the state's objectives and (b) individual analyses of the assessment items of each of the assessments. Following individual analyses of the items, reviewers participate in a debriefing discussion in which they give their overall impressions of the alignment between the assessment and the state's curriculum standards.

Reviewers are instructed to focus primarily on the alignment between the state standards and the various assessments. However, they are encouraged to offer their opinions on the quality of the standards, or of the assessment activities/items, by writing notes about the items. Reviewers can also indicate whether there is a source-of-challenge issue with the item—that is, a problem with the item that might cause the student who knows the material to give a wrong answer or enable someone who does not have the knowledge being tested to answer the item correctly. For example, a mathematics item that involves an excessive amount of reading may represent a source-of-challenge issue because the skill required to answer is more a reading skill than a mathematics skill. Source of challenge can be considered a fifth alignment criteria in the analysis and was originally so defined by Achieve, Inc.

The results produced from the institute pertain only to the issue of agreement between the state standards and the assessment instruments. Thus, the alignment analysis does not serve as external verification of the general quality of a state's standards or assessments. Rather, only the degree of alignment is discussed in the results. The averages of the reviewers' coding are used to determine whether the alignment criteria are met. When reviewers do vary in their judgments, the averages lessened the error that might result from any one reviewer's findings. Standard deviations, which give one indication of the variance among reviewers, are reported.

To report on the results of an alignment study of a state's curriculum standards and assessments for different grade levels, the study addresses specific criteria related to

the content agreement between the state standards and grade-level assessments. The four alignment criteria receive major attention in the reports: (a) Categorical Concurrence, (b) DOK Consistency, (c) Range of Knowledge Correspondence, and (d) Balance of Representation.

ALIGNMENT CRITERIA USED FOR THIS ANALYSIS

The analysis, which judges the alignment between standards and assessments on the basis of four criteria, also reports on the quality of assessment items by identifying those items with sources of challenge and other issues. For each alignment criterion, an acceptable level is defined for what would be required to ensure that a student had met the standards.

Categorical Concurrence

An important aspect of alignment between standards and assessments is whether both address the same content categories. The Categorical Concurrence criterion provides a very general indication of alignment if both documents incorporate the same content. The criterion of Categorical Concurrence between standards and assessment is met if the same or consistent categories of content appear in both documents. This criterion is judged by determining whether the assessment included items measuring content from each standard. The analysis assumes that the assessment had to have at least six hits measuring content from a standard in order for an acceptable level of Categorical Concurrence to exist between the standard and the assessment. A hit is used here always to designate that a reviewer has mapped an assessment item to one objective. The number of hits for a standard is determined by taking the average number of hits each reviewer codes as corresponding to one objective under a standard. It is possible for one assessment item to have up to three hits, each to a different objective, under the same standard. Different reviewers may have different items as hits for a standard. The Categorical Concurrence criterion represents a summary of the average number of hits assigned to each standard. When reviewers assign the same number of hits to a standard, they agree that there is a similar weighting of information from the assessment to make judgments about students' performance on the standard. Multiple hits for one item are only encouraged if the assessment item measures content related to more than one objective. Multiple hits are only to be used sparingly. In general, then, the number of hits is the same as the number of items.

The number of hits, six, used as an acceptable level of Categorical Concurrence is based on estimating the number of items that could produce a reasonably reliable subscale for estimating students' mastery of content on that subscale. Of course, many factors have to be considered in determining what a reasonable number is,

including the reliability of the subscale, the mean score, and cutoff score for determining mastery. Using a procedure developed by Subkoviak (1988) and assuming that the cutoff score is the mean and that the reliability of one item is .1, it was estimated that six items would produce an agreement coefficient of at least .63. Usually, states do not report student results by standards or require students to achieve a specified cutoff score on subscales related to a standard. If a state did do this, then the state would seek a higher agreement coefficient than .63.

DOK Consistency

Standards and assessments can be aligned not only on the category of content covered by each but also on the basis of the complexity of knowledge required by each. DOK Consistency between standards and assessment indicates alignment if what is elicited from students on the assessment is as demanding cognitively as what students are expected to know and do as stated in the standards. For consistency to exist between the assessment and the standard, as judged in this analysis, at least 50% of the hits corresponding to an objective have to be at or above the level of knowledge of the objective. Fifty percent, a conservative cutoff point, is based on the assumption that a minimal passing score for any one standard of 50% or higher would require the student to successfully answer at least some items at or above the DOK level of the corresponding objectives. For each category—below, at, and above—the percentage of hits is averaged across the reviewers to produce the value for this criterion. To illustrate how the acceptable level for the DOK Consistency criterion is computed, assume an assessment included six items related to one standard and students were required to answer correctly four of those items to be judged proficient—that is, 67% of the items. If three, 50% of the six items, were at or above the DOK level of the corresponding objectives, then for a student to achieve a proficient score would require the student to answer correctly at least one item at or above the DOK level of one objective. Some leeway is used in the analysis on this criterion. If a standard has between 40% and 50% of items at or above the DOK levels of the objectives, then it is reported that the criterion is "weakly" met.

Interpreting and assigning DOK levels to both objectives within standards and to assessment items is an essential requirement of alignment analysis. These descriptions help to clarify what the different levels represent in, for example, mathematics.

Level 1 (recall). Level 1 includes recalling information such as a fact, definition, term, or a simple procedure, as well as performing a simple algorithm or applying a formula. That is, in mathematics, a one-step, well-defined, and straight algorithmic procedure should be included at this lowest level. Other key words that signify a Level 1 include "identify," "recall," "recognize," "use," and "measure."

Verbs such as "describe" and "explain" could be classified at different levels, depending on what is to be described and explained.

Level 2 (skill/concept). Level 2 includes the engagement of some mental processing beyond a habitual response. A Level 2 assessment item requires students to make some decisions as to how to approach the problem or activity, whereas Level 1 requires students to demonstrate a rote response, perform a well-known algorithm, follow a set procedure (like a recipe), or perform a clearly defined series of steps. Key words that generally distinguish a Level 2 item include *classify, organize, estimate, make observations, collect and display data,* and *compare data.* These actions imply more than one step. For example, to compare data requires first identifying characteristics of the objects or phenomenon and then grouping or ordering the objects. Some action verbs, such as *explain, describe,* or *interpret,* could be classified at different levels, depending on the object of the action. For example, interpreting information from a simple graph or requiring the reading of information from the graph also are at Level 2. Interpreting information from a complex graph that requires some decisions on what features of the graph need to be considered and how information from the graph can be aggregated is at Level 3. Level 2 activities are not limited only to number skills but can involve visualization skills and probability skills. Other Level 2 activities include (a) noticing and describing nontrivial patterns; (b) explaining the purpose and use of experimental procedures; (c) carrying out experimental procedures; (d) making observations and collecting data; (e) classifying, organizing, and comparing data; and (f) organizing and displaying data in tables, graphs, and charts.

Level 3 (strategic thinking). Level 3 requires reasoning, planning, using evidence, and a higher level of thinking than the previous two levels. In most instances, requiring students to explain their thinking is at Level 3. Activities that require students to make conjectures are also at this level. The cognitive demands at Level 3 are complex and abstract. The complexity does not result from the fact that there are multiple answers, a possibility for both Levels 1 and 2, but because the task requires more demanding reasoning. An activity, however, that has more than one possible answer and requires students to justify the response they give would most likely be at Level 3.

Other Level 3 activities include (a) drawing conclusions from observations, (b) citing evidence and developing a logical argument for concepts, (c) explaining phenomena in terms of concepts, and (d) using concepts to solve problems.

Level 4 (extended thinking). Level 4 requires complex reasoning, planning, developing, and thinking most likely over an extended period of time. The extended time period is not a distinguishing factor if the required work is only repetitive and does not require applying significant conceptual understanding and

higher order thinking. For example, if a student has to take the water temperature from a river each day for a month and then construct a graph, this would be classified at Level 2. However, if the student is to conduct a river study that requires taking into consideration a number of variables, this would be at Level 4. At Level 4, the cognitive demands of the task should be high and the work should be very complex. Students should be required to make several connections—relate ideas within the content area, or among content areas—and would have to select one approach among many alternatives on how the situation should be solved to be at this highest level. Level 4 activities include (a) developing and proving conjectures, (b) designing and conducting experiments, (c) making connections between a finding and related concepts and phenomena, (d) combining and synthesizing ideas into new concepts, and (e) critiquing experimental designs.

Range of Knowledge Correspondence

For standards and assessments to be aligned, the breadth of knowledge required on both should be comparable. The Range of Knowledge Correspondence criterion is used to judge whether a comparable span of knowledge expected of students by a standard is the same as, or corresponds to, the span of knowledge that students need to correctly answer the assessment items/activities. The criterion for correspondence between span of knowledge for a standard and an assessment considers the number of objectives within the standard for which there is at least one related assessment item/activity. The number of objectives under a standard that reviewers assign at least one hit is averaged across reviewers. As with the other criteria, the specific objectives that reviewers assign hits may not be the same. The value used for this criterion represents the average number of objectives for which the reviewers assigned hits, but the actual objectives hit may differ from one reviewer to the next. The average number of hits across reviewers provides one indicator of coverage and not whether there is exact agreement on what is assessed. Moreover, one item per objective is a very lenient criterion for considering measurement of an objective. However, the set of objectives under a standard does represent some delineation of the content domain for a standard. It is assumed that these objectives partition the content addressed by a standard. Having one item per objective for at least half of the objectives provides a decision rule that ensures the assessment is measuring some breadth in content knowledge and is at least sampling half of the most important partitions of content identified by the objectives. The data produced by the alignment analysis produce more precise information that shows the mapping of each item by each reviewer, so it is possible to determine precisely the distribution of items by objectives under a standard.

Fifty percent of the objectives for a standard have to have at least one related assessment item in order for the alignment on this criterion to be judged acceptable. This level is based on the assumption that students' knowledge should be tested on

content from over half of the domain of knowledge for a standard. This assumes that each objective for a standard should be given equal weight. Depending on the balance in the distribution of items and the necessity for having a low number of items related to any one objective, the requirement that assessment items need to be related to more than 50% of the objectives for a standard increases the likelihood that students will have to demonstrate knowledge on more than one objective per standard to achieve a minimal passing score. As with the other criteria, a state may choose to make the acceptable level on this criterion more rigorous by requiring that an assessment include items related to a greater number of the objectives. However, any restriction on the number of items included on the test will place an upper limit on the number of objectives that can be assessed. Range of Knowledge Correspondence is more difficult to attain if the content expectations are partitioned among a greater number of standards and a large number of objectives. If 50% or more of the objectives for a standard have a corresponding assessment item, then the Range of Knowledge Correspondence criterion is considered to be met. If between 40% to 50% of the objectives for a standard have a corresponding assessment item, the criterion is weakly met.

Balance of Representation

In addition to comparable depth and breadth of knowledge, aligned standards and assessments require that knowledge be distributed equally in both. The Range of Knowledge Correspondence criterion only considers the number of objectives within a standard hit (a standard with a corresponding item); it does not take into consideration how the hits (or assessment items/activities) are distributed among these objectives. The Balance of Representation criterion is used to indicate the degree to which one objective is given more emphasis on the assessment than another. An index is used to judge the distribution of assessment items. This index only considers the objectives for a standard that have at least one hit—that is, at least one related assessment item per objective. The index is computed by considering the difference in the proportion of the standard represented by each objective and the proportion of hits assigned to the objective. The index is computed using this formula:

$$\text{BALANCE INDEX } 1 - \left(\sum_{k=1}^{O} |1/(O) - I(k)/(H)|\right)/2$$

Where 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

If all of the items assigned to a standard are evenly distributed among the objectives, then the index value will be 1. The larger the number of items that correspond to one objective, while other objectives have only one or two corresponding item, the smaller the index value.

Two examples of the use of the Balance index are shown in Figures 1 and 2. In both figures 12 items are distributed among seven objectives under one standard. In Figure 1, the total number of objectives hit is seven, the total number of items is 12, and the total number of items that hit one objective ranges from one to four. The Balance index value is .73 and would be judged to be acceptable as described later. The computation is $1 - (|1/7 - 4/12| + |1/7 - 1/12| + |1/7 - 1/12| + |1/7 - 1/12| + |1/7 - 1/12| + |1/7 - 1/12| + |1/7 - 1/12|)/2$ or $1 - .54/2 = 1 - .27 = .73$. In Figure 2, the total number of objectives and items are the same except the distribution of items by objective differs, now ranging from one to six. The Balance index value is .67 and would be judged to be only weakly acceptable, because Objective 1.1.2 has a relatively high number of corresponding items.

An index value of 1 signifies perfect balance and is obtained if the hits (corresponding items) related to a standard are equally distributed among the objectives for the given standard. Index values that approach 0 signify that a large proportion of the hits address only one or two of all of the objectives hit. Depending on the number of objectives and the number of hits, a unimodal distribution (most items related to one objective and only one item related to each of the remaining objectives) has an index value of less than .5. A bimodal distribution has an index value of around .55 or .6. Index values of .7 or higher indicate that items/activities are distributed among all of the objectives at least to some degree (e.g., every objective has at least two items) and is used as the acceptable level on this criterion. Index

			X						X				
			X						X				
			X						X				
			X	X		X	X		X		X		X
Goal/Obj #	1.1	1.1.1	1.1.2	1.1.3	1.2	1.2.1	1.2.2	1.2.3	1.2.4	1.3.1	1.3.2	1.3.3	1.3.4

FIGURE 1 Example of 12 items judged as corresponding to seven objectives under one standard that produces a Balance index value of .73 (acceptable).

			X										
			X										
			X										
			X										
			X										
			X	X		X	X		X		X		X
Goal/Obj #	1.1	1.1.1	1.1.2	1.1.3	1.2	1.2.1	1.2.2	1.2.3	1.2.4	1.3.1	1.3.2	1.3.3	1.3.4

FIGURE 2 Example of 12 items judged as corresponding to seven objectives under one standard that produces a Balance index value of .67 (weak).

values between .6 and .7 indicate the Balance of Representation criterion is only weakly met.

Four Alignment Criteria

Each alignment criterion is defined to represent a different attribute of how an assessment and a set of standards can relate to each other. To determine whether an assessment and standards are fully aligned (as described here) requires that all four of the criteria to be met. An acceptable value on any one criterion is necessary but not sufficient to have alignment. Thus, alignment is determined if there are sufficient number of items allocated to each standard with an appropriate level of complexity and coverage and without overemphasizing any one content area.

CHALLENGES AND ISSUES

This section identifies some of the issues that have arisen in performing alignment studies and addresses some of the basic principles of aligning content standards and assessments.

Acceptable Level for Number of Items Per Standard

The Webb alignment process uses six items measuring content related to a standard as the acceptable level, but there are questions about six as the minimal level. This number, as discussed earlier, was derived using a procedure developed by Subkoviak (1988) to determine reliability pertaining to judging a person's mastery based on assessment items. The Webb alignment tool has a feature that allows people to vary the number used to define an acceptable level, so the process does have some flexibility. However, some situations have come up that raise questions about six as the number. Table 1 reports the findings from State A's science alignment analysis for grade 3 for the Categorical Concurrence criterion. Of the six science standards, three standards met the acceptable level by having six hits, and three standards did not. The mean number of hits for standards across raters is shown in Table 1, along with the proportions of items specified in the state test blueprint for each standard. Clearly, the state gives more emphasis to two of the standards, 3.2 (Inquiry) and 3.4 (Subject Matter and Concepts), and equal but minor emphasis to the other four standards. This relative emphasis is reflected by the distribution of items by standard on the assessment. However, Standard 3.5 (Design and Applications) and Standard 3.6 (Personal and Social) are more difficult to assess on an on-demand assessment and were given less emphasis, even less than specified by the test blueprint. The report indicated that the alignment was not acceptable because there was an insufficient number of items for three of the Grade 3 standards.

TABLE 1
(State A) Mean Number of Hits and Acceptable Level on Categorical Concurrence for Grade 3 Science

Title/Standard	Test Blueprint (%)	No. of Hits[a]		Categorical Concurrence[b]
		M	*SD*	
3.1. History/Nature	8	1	0	No
3.2. Inquiry	30	17.38	2.12	Yes
3.3. Unifying Themes	8	7.5	4	Yes
3.4. Subject Matter/Concepts	38	33.5	1.94	Yes
3.5. Design/Applications	8	2.12	1.27	No
3.6. Personal/Social	8	4.75	1.09	No
Total		66.25	5.78	

Note. $N = 55$ items.

[a]A hit indicates a reviewer has mapped an assessment item to one objective. One item could be mapped to two objectives. This would represent two hits. Because multiple hits for an item were rare, the number of hits is very close to the number of items.

[b]*Yes* indicates an acceptable level has been met (six or more hits). *No* indicates an acceptable level has not been met.

At issue: Is six items a reasonable minimum, or should adjustments be made in this acceptable level? If adjustments are to be made, then what should be the decision rule?

Distribution of Items Related to a Standard by DOK Level

A second issue regards the distribution of items on an assessment by the DOK level. Is 50% of the items coded to a standard with a DOK level at or above the DOK level of the corresponding objective appropriate as the minimal acceptable level? Table 2 displays the data for one state and one grade where this acceptable level was met for four of the six standards. For Standard 3, only 42% of the more than 13 items coded as corresponding to that standard, on the average, had a DOK level that was the same as or above the DOK level of the corresponding objective. Because this is within 10% of the acceptable level of 50% items with DOK levels at or above that of the corresponding objective, it was judged that this standard and assessment only weakly met the alignment criterion of DOK Consistency. Thus, a student could answer 8 of the 13 items corresponding to Standard 3, generally a level sufficient to be declared proficient on a standard, without ever answering a question with a DOK level that is at least as high as the corresponding objective. Standard 1 and the assessment failed to be acceptable on the DOK Consistency criterion because only 17% of the nearly 10 items that reviewers on average indicated as corresponding to that standard had a DOK level that was at least comparable to the DOK level of the corresponding objective.

TABLE 2
(State B) Percentage of Hits With a Depth of Knowledge Under, At, and Above the Depth of Knowledge of the Corresponding Objective for Each Standard for High School Mathematics

Title/Standard	*No. of Hits (M)*	*% Under (M)*	*% At (M)*	*% Above (M)*	*DOK Consistency*[a]
I. Patterns, Relationships and Functions	10.44	83	17	0	No
II. Geometry and Measurement	13	20	51	29	Yes
III. Data Analysis and Statistics	13.44	58	40	2	Weak
IV. Number Sense and Numeration	2.78	25	61	14	Yes
V. Numerical and Algebraic Operations and Analytical ...	10.67	30	57	12	Yes
VI. Probability and Discrete Mathematics	6.89	42	56	2	Yes
Total	57.22	43	47	11	

Note. $N = 51$ items. DOK = depth of knowledge.

[a]*Yes* represents 50% or higher at or above. *Weak* represents 40% to 50% at or above. *No* represents less than 40% at or above.

The acceptable level for DOK is based on the assumption that proficiency is set at 50% of the items correct and that students judged as proficient should have answered at least one item with a DOK level that is at least at the same level of complexity as the corresponding content objective. However, what is considered as acceptable should depend to some degree on the purpose of the assessment. If the purpose of the assessment is to differentiate between students who are proficient from students who are not, then an argument could be made that all or nearly all of the item DOK levels should be the same as the DOK levels of the corresponding objectives. However, if the purpose of the assessment is to place students on a range of proficiency levels (e.g., *below basic, basic, proficient,* and *advanced*), then it is reasonable to have items with a range of DOK levels in comparison to the corresponding objectives.

Content standards and many objectives under content standards cover a broad range of content that students are expected to attain. Thus, the domain of items for measuring students' knowledge related to an objective or standard can be very large and vary by complexity or DOK level. The alignment process devised by Webb has reviewers assign one DOK level to each objective. Reviewers, who are experts in the content area, are to assign a DOK level to an objective by judging the complexity of the most representative assessment items or content expressed by the objective. Realizing that many objectives cover a broad range of content, it may be reasonable to have items with different DOK levels corresponding to the same objective, some below the DOK level of the objective, some at, and some above.

The decision rule imposed in the alignment analysis discussed here is based on judging whether students are proficient. Another decision rule could be based on having items that are more representative of the range of complexity in objectives and standards, such as 20% with a DOK level of 1, 60% with a DOK level of 2, and 20% with a DOK level of 3, or the range of complexity could be decided by a certain percentage of items that are below, at, or above the corresponding objectives. The issue remains that there are different ways of considering what is an acceptable distribution of items by complexity that depends largely on the purpose of the assessment.

Range of Knowledge in Content Coverage of a Standard

A third issue is related to what constitutes the appropriate breadth of coverage for a standard. The decision rule currently being used is for 50% or more of the objectives under a standard to have at least one corresponding assessment item, for a minimal acceptable breadth of coverage. The number of objectives under a standard is highly related to the difficulty in meeting the Range of Knowledge Correspondence criteria (breadth in content). If a state lists a large number of objectives under a standard, then it is more difficult for a state to meet an acceptable level on Range of Knowledge because of the limited number of objectives that can be used on an assessment. For example, State B (Table 3) had six standards in mathematics. Each standard had from 9 to 18 objectives for a total of 77 objectives. The high school test had a total of 51 items. Except for Standard 5, all of the other standards had from 17% to 38% of the objectives under the standard with at least one hit or corresponding item. This proportion of the objectives with at least one hit or corresponding item is well below the acceptable level of 50% of the objectives.

Having an adequate breadth of content on an assessment can be a trade-off with the length of the assessment. An assessment with fewer items will have more difficulty assessing, at least partially, all of the objectives. Other factors come into play when considering breadth. Some standards may have a larger number of objectives because the standard covers more content. For example, for State B as depicted in Table 3, Standard 2 (Geometry and Measurement) has more objectives than Standard 5 (Numerical and Algebraic Operations), 18 objectives compared to 9 objectives. This suggests that the content under Standard 2 has been partitioned in more ways than content under Standard 5. It could be that the objectives under geometry and measurement are more specific or that the state considered that geometry and measurement had more content to cover. Another factor is that some of the objectives under Standard 2 may be more difficult to assess on an on-demand assessment, particularly if one item only measures content related to one objective. An on-demand assessment could cover more content by including more robust items that measure content associated with more than one objective or standard.

TABLE 3
(State B) Mean Number of Objectives and Percentage of Total With At Least One Hit for High School Mathematics

Title/Standard	*Goals (No.)*	*Objectives (No.)*	*No. of Hits (M)*	*No. of Objectives Hit (M)*	*% Objectives Hit (M)*	*Range of Knowledge*[a]
I. Patterns, Relationships and Functions	2	11	10.44	4.22	38	No
II. Geometry and Measurement	3	18	13	5.78	32	No
III. Data Analysis and Statistics	3	14	13.44	5	35	No
IV. Number Sense and Numeration	3	14	2.78	2.44	17	No
V. Numerical and Algebraic Operations and Analytical …	2	9	10.67	5.22	55	Yes
VI. Probability and Discrete Mathematics	2	11	6.89	3.67	33	No
Total	15	77	57.22	4.39	35	

Note. $N = 51$ items.

[a]*Yes* represents mean of objectives hit as greater than half of the objectives. *No* represents the mean of objectives hit as less than 40% of the objectives.

The current decision rule of 50% of the objectives with at least one hit clearly is a very minimal requirement for alignment. A number of factors could be considered in judging the adequate range of content, including (a) the breadth of content covered by a standard, (b) the length of the assessment, (c) the suitability of the content to be assessed on an on-demand assessment, and (d) differences in importance of the different objectives under a standard. Considering these and other factors, then, other decision rules could be developed, such as randomly sampling objectives under a standard, setting a minimum number of objectives under a standard to have a hit, or differentiating the importance of some standards from others by requiring more objectives under the most important standards being assessed than under the less important standards. As with the other issues, there are multiple considerations that need to be addressed in judging the adequacy of the alignment between an assessment and set of standards.

Balance of Representation Given Some Objectives

It is reasonable that some standards will be more important than other standards and that some objectives under a standard will be more important than other objec-

TABLE 4
State B Balance of Representation High School Language Arts
(3 of 12 Standards)

Title/ Standard	*Goals (No.)*	*Objectives (No.)*	*Balance Index*		*Balance of Representation*[a]
			M	*SD*	
I. Meaning and Communication—Reading	1	5	0.57	0.12	No
II. Meaning and Communication—Writing	1	4	0.68	0.14	Weak

Note. $N = 116$ items.
[a]*No* indicates an index value below .60. *Weak* indicates an index value of .60 to .69.

tives. The Balance-of-Representation alignment criterion, however, assumes that items should be fairly evenly distributed among the objectives under a standard. Table 4 shows the index value for two language arts standards for an analysis for one state. Figure 3 provides a pictorial representation of the distribution of hits (test items) that were coded as corresponding to the different objectives for Standards 1 and 2. For both of these standards, reviewers coded a large number of hits (over 25) as corresponding to one objective under each standard. This resulted in index values of .57 and .68, which are below the acceptable level used of .70 (Table 4).

At issue with Balance is the degree to which the amount of emphasis given to different objectives under a standard should vary. It is possible for a state to accept

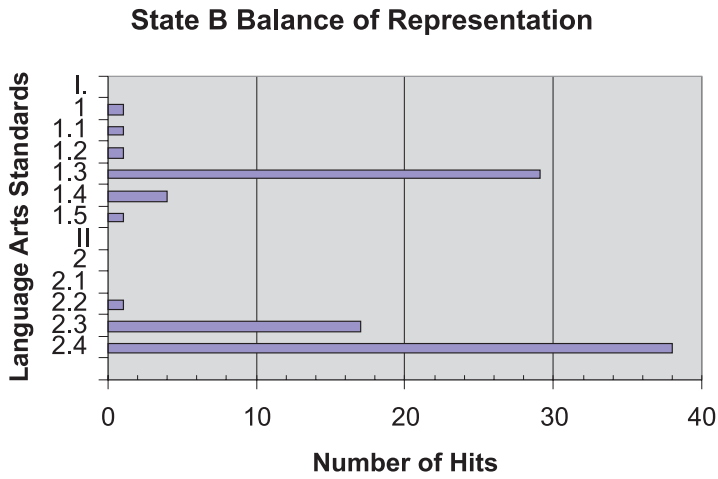

FIGURE 3 Number of hits across all reviewers for each objective under Standards 1 and 2 for State B high school language arts.

a lower Balance index value than .70. So State B could be satisfied with a Balance index value for Standard 2 of .68 (Table 4), with a large emphasis on Objective 2.4, as seen in Figure 3, which includes all of the hits across all of the reviewers for Standards 1 and 2. For Standard 1, a high proportion of the hits for that standard correspond to Objective 1.3. This configuration produced an index value of .57 (Table 4) and the standard was judged not to have met an acceptable level for Balance of Representation.

For Standard 2, a relatively higher proportion of hits were allocated to two of the four objectives that drew any hits, Objectives 2.3 and 2.4. This configuration produced an index value of .68 and the standard was judged only to have weakly met an acceptable level for Balance of Representation. There are many reasons why one objective may be emphasized more on an assessment than others and not always for content reasons. One objective could be emphasized more than other objectives because it is easier to write assessment items for that objective when compared to other objectives. The main issue to be resolved is how should alignment analyses consider the difference in emphasis by objectives. This issue relates to how the assessment blueprint differentiates among objectives and whether or not large variations among objectives are appropriate.

Change in DOK Level Across Grades

The final issue to be discussed is the change in complexity of content across grade levels. It is reasonable to expect that, as students proceed through the grades, more reasoning and analysis will be expected of them and less simple recall and recognition. This was the case for State A in mathematics (Figure 2) and in language arts (Figure 3). For both mathematics and language arts, the percent of objectives with a DOK level of 1 (Recall and Recognition) decreased, while the percent of objectives with a DOK level of 3 (Strategic Reasoning) increased from grade 3 to grade 10 (Figures 4 and 5). However, DOK levels are dependent somewhat on grade level and on what a typical student at a grade level can be expected to know and do. Reviewers in an alignment analysis developed by Webb are instructed to think about what a typical student should be expected to know and do in assigning DOK levels to the content objectives. In reading, the increase in complexity across grades may be due to the greater sophistication of the passages, while the actual behavior or cognitive requirements stay relatively constant, such as determining the main idea. However, if along with more sophisticated passages students are expected to do more with drawing inferences or paraphrasing, then the DOK levels may increase across grades. Currently, there are really no fixed guidelines as to what constitutes an acceptable progression in content complexity from grade to grade. In the absence of such guidelines, the progression of content complexity depicted in Figures 4 and 5 for State A seems to be reasonable.

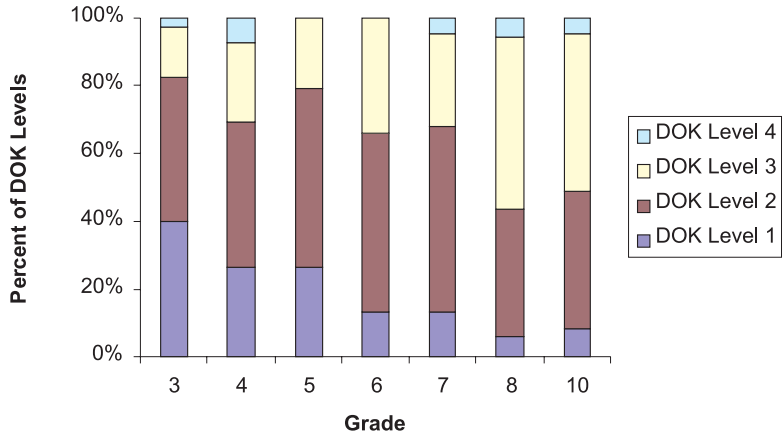

FIGURE 4 State A mathematics DOK levels for objectives by grade.

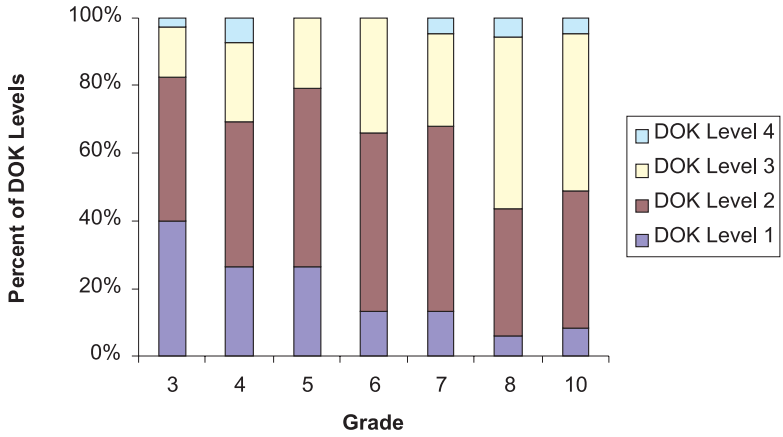

FIGURE 5 State A reading language arts DOK levels for objectives by grade.

CONCLUSIONS

This article describes a process that has been used to analyze the agreement between state academic content standards and state assessments. The Webb alignment process was developed for the National Institute for Science Education and

the Council of Chief State School Officers in 1997 and has evolved over time. A Web-based tool is now available for aiding in conducting the process and analyzing the results. The process produces information about the relation between a set of standards and an assessment by reporting on four main alignment criteria—Categorical Concurrence, DOK Consistency, Range of Knowledge Correspondence, and Balance of Representation.

Five alignment issues were discussed. Each of these issues is related to one or more of the alignment criteria. These issues center around the basic question of when an alignment is good enough. Specific rationales described in this article have been used to set acceptable levels for each of the four alignment criteria. These acceptable levels have been specified for primarily pragmatic reasons, such as assumptions regarding what would be considered a passing score, the number of items needed to make some decisions on student learning, and the relatively low number of items that can be included on an on-demand assessment. The issues discussed arise from a change in these underlying assumptions and from considering variations in the purpose of an assessment. The issues themselves are not resolved in this article, and this was not the intention of the article even if it were possible. The existence of these issues and other related issues just point to the fact that judging the alignment among standards and assessments requires a certain subjectivity and cannot be based solely on a clear set of objective rules. This makes it critical that in any alignment analysis the underlying assumptions and how conclusions are reached must be made clear at the outset.

The issues discussed in this article do not constitute all of the issues related to judging the alignment between standards and assessments. The decisions that are made with this process are clearly stated along with the assumptions. They are judgments and are not derived empirically. The process depends heavily on content experts. There is an underlying question as to who should be making the decision on alignment and if the assumptions for these decisions are accurate. The true test for alignment is the improvement of student achievement as described by the expectations. However, the argument soon becomes circular. Aligned assessments are needed to show that students are meeting the expectations as stated in the standards, but the measurement of the attainment of standards can only be done using the assessments. Alignment as portrayed here is a content analysis that needs to be done by content experts. The definitions of the DOK levels raise another issue. These definitions are intended to describe the complexity of content as determined by a content analysis. The DOK levels relate to cognitive levels but have not withstood the empirical analyses to say the levels correspond to cognitive levels. Cognitive laboratories would be an excellent way to delve deeper in the DOK levels and their relation to cognition. However, simply as representations of content complexity, the DOK levels provide a very valuable function in producing a language for comparing what is expected in assessment items and what is expected in curriculum standards. Alignment itself is a fruitful area of study. The issues outlined in

this article only scratch the surface in trying to increase the systematic link between assessments and standards.

ACKNOWLEDGMENT

This work was supported by a subgrant from the U.S. Department of Education (S368A030011) to the State of Oklahoma and a grant from the National Science Foundation (EHR 0233445) to the University of Wisconsin–Madison. Any opinions, findings, or conclusions are those of the author and do not necessarily reflect the view of the supporting agencies.

REFERENCES

Blank, R. (2002). *Models for alignment analysis and assistance to states.* Council of Chief State School Officers Summary Document. Washington, DC: Council of Chief State School Officers.

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

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

Porter, A., Smithson, J., Blank, R., & Zeidner, T. (2007/this issue). Alignment as a teacher variable. *Applied Measurement in Education, 20,* 27–51.

Subkoviak, M. J. (1988). A practitioner's guide to computation and interpretation of reliability indices for mastery tests. *Journal of Educational Measurement, 25,* 47–55.

Webb, N. L. (1997). *Criteria for alignment of expectations and assessments in mathematics and science education* (Council of Chief State School Officers and National Institute for Science Education Research Monograph No. 6). Madison: University of Wisconsin, Wisconsin Center for Education Research.

Webb, N. L. (2002). *Alignment study in language arts, mathematics, science, and social studies of state standards and assessments for four states.* Washington, DC: Council of Chief State School Officers.