An Analysis of the Alignment Between Mathematics Standards and Assessments for Three States

by

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

April 3, 2002

A paper presented at the American Educational Research Association Annual Meeting in New Orleans, Louisiana April 1-5, 2002.

This paper reports results from one of the alignment studies conducted and supported by the Technical Issues of Large-Scale Assessment (TILSA) group of the Council of Chief State School Officers. The alignment institute was held May 21-24, 2001. Any opinions, findings, or conclusions are those of the author and do not necessarily reflect the views of the support agency or of WCER.

An Analysis of the Alignment Between Mathematics Standards and Assessments for Three States

by

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

Alignment is an important attribute for educational systems. Although the concept has been known for some time (Cohen, 1987; Carroll, 1963), alignment gained more prominence in the early 1990s with the advent of standards (NCTM, 1989) and systemic reform (Smith & O'Day, 1991). Educators increasingly recognized that if policy elements are not aligned, the system will be fragmented, will send mixed messages, and will be less effective (Consortium for Policy Research in Education, 1991; Newmann, 1993). For example, the Systemic Initiatives program of the National Science Foundation was directed toward states, districts, and regions setting ambitious goals for student learning through a coherent policy system focused , in part, on assessments aligned with those goals. The Improving America's Schools Act explicated how assessments were to relate to standards: " . . . such assessments (high quality, yearly student assessments) shall . . . be aligned with the State's challenging content and student performance standards and provide coherent information about student attainment of such standards . . ." (U.S. Congress, 1994, p. 8). The U.S. Department of Education's explanation of the Goals 2000: Educate America Act and the Elementary and Secondary Education Act (which includes Title I) indicated alignment of curriculum, instruction, professional development, and assessments as key performance indicators for states, districts, and schools striving to meet challenging standards.

This is a report of a study of an alignment analysis conducted on the standards and assessments of three states. An earlier study analyzed the alignment for four other states (Webb, 1999). In this report, alignment is defined and the process used to do the analysis is described. Then findings from the study are reported, along with reliabilities of reviewers. This report ends by identifying a number of issues related to judging alignment, with some discussion of these issues.

Alignment

Alignment of expectations for student learning and assessments for measuring students' attainment of these expectations is an essential attribute for an effective standards-based education system. Alignment is defined 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. As such, alignment is a quality of the relationship between expectations and assessments and not an attribute of any one of these two system components. Alignment describes the match between expectations and assessment that can be legitimately improved by changing either student expectations or assessments. As a relationship between two or

more system components, alignment can be determined by using the multiple criteria described in detail in a National Institute for Science Education (NISE) research monograph, *Criteria for Alignment of Expectations and Assessments in Mathematics and Science Education* (Webb, 1997).

Alignment Institute

A four-day Alignment Analysis Institute was conducted May 20 through May 24, 2001. Four people, including state assessment consultants, content experts, and researchers, analyzed the agreement between the mathematics standards and assessments. The institute was coordinated by the Council of Chief State School Officers (CCSSO) as a function of the TILSA (Technical Issues in Large-Scale Assessment) collaborative among states. At the institute, concurrently with two teams analyzing the mathematics standards and assessments from three states, two teams of specialists analyzed language arts standards and assessments from four states including the same three states participating in the mathematics alignment study. In mathematics, the alignment of standards and mathematics assessments was analyzed for three grade levels for two states and two grade levels for one state. This report only attends to the mathematics analysis. The grade levels analyzed in this study by state were:

State E	Grades 4, 7, and 9
State F	Grades 5 and 8
State G	Grades 4, 8, and 11

A major goal of the institute was to further develop a systematic process and analytical tools for judging the alignment between standards and assessments based on the criteria developed in conjunction with CCSSO and National Institute for Science Education (Webb, 1997). In addition to training reviewers to use the process, they were asked to provide suggestions for improving the process. Further feedback has been elicited from the TILSA group, which was presented results from the analysis at the end of January, 2002.

Alignment Coding Process

Reviewers were trained to identify the depth-of-knowledge of objectives and assessment items. This training included reviewing the definitions of the four depth-of-knowledge levels and then reviewing examples of each. For the first step of the review process, the team of reviewers read each objective for each standard and reached consensus on the appropriate depth-of-knowledge level of that the objective. This step afforded the team of reviewers opportunity to gain greater familiarity with the objectives themselves and the four depth-of-knowledge levels. Before independently coding the items from each assessment, the reviewers independently coded the sample of five to ten items from the assessments and then compared the assigned depth-of-knowledge level assigned to these items and the corresponding content objectives to each item. In this way, the reviewers calibrated their coding of the depth-of-knowledge level and the assigned objective. The process is not designed for the reviewers to reach exact

agreement. The reviewers' responses are averaged and the variances among reviewers on what constitutes a corresponding objective to an item are considered valid differences in opinion that are a result of a lack of clarity in how the objectives were written and/or the robustness of an item that may legitimately correspond to more than one objective. Reviewers were allowed to identify more than one objective to an individual assessment item. They could assign a primary hit, the main objective the assessment item corresponds to, and up to two secondary hits.

Reviewers were instructed to attend to the alignment between the state standards and assessments. They were able to offer their opinion on the quality of the standards or of the assessment activities/item(s) by writing a note about the item(s). Reviewers also could identify whether the item presented a source-of- challenge issue, a problem with the item that may cause the student who knows the material to answer wrong, or someone who does not have the knowledge being tested to answer correctly. For example, a mathematics item that requires an excessive amount of reading may present a source-of-challenge issue because the item is more a reading item than a mathematics item.

The results produced from the institute pertain only to how the state standards and the state assessment are in agreement and do not serve as external verification of the general quality of a state's standards or assessments. The results of the alignment institute do provide judgments of content area experts, independent of any of the participating states, who are very familiar with state and national standards. The means of the reviewers' coding were used to determine whether the alignment criteria were met. When reviewers did vary in their judgments, the means lessened the error that might result from the input of any one reviewer. The standard deviations are reported, which give one indication of the variance among reviewers.

This report describes the results of an alignment study of standards and grade-level tests in mathematics for three states. The study addressed specific criteria related to the content agreement between the state standards and grade-level assessments. Four criteria received major attention: categorical concurrence, depth-of-knowledge consistency, range-of-knowledge correspondence, and balance of representation.

Mathematics Alignment Criteria Used for This Analysis

This analysis judged the alignment between the standards and the assessment using four criteria. For each criterion, an acceptable level was defined based on what would be required to assure that students have met the standards.

Categorical Concurrence

One aspect of alignment between standards and assessments is whether both address the same content categories. The categorical concurrence criterion provides a very general indication whether 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 was judged

by determining whether the assessment included items measuring content from each standard. The analysis assumed that the assessment had to have at least six items measuring content from a standard in order for an acceptable categorical concurrence between the standard and the assessment to exist. The number of items, six, 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 the cutoff score for determining mastery. Using a procedure developed by Subkoviak (1988) and assuming that the cutoff score is the mean and the reliability of one item is .1, it was estimated that six items would produce an agreement coefficient of at least .63. This indicates that about 63% of the group would be consistently classified as masters or nonmasters, if two equivalent test administrations were employed. The agreement coefficient would increase if the cutoff score is increased to one standard deviation from the mean to .77 and, with a cutoff score of 1.5 standard deviations from the mean, to .88. None of the four states included in the analysis reported student results by standards or required students to achieve a specified cutoff score on subscales related to a standard. If a state did do this, then the state would want a higher agreement coefficient than .63. Six items were assumed as a minimum for an assessment measuring content knowledge related to a standard and as a basis for making some decisions about students' knowledge of that standard. If the mean for six items is 3 and one standard deviation is one item, then a cutoff score set at 4 would produce an agreement coefficient of .77. Any fewer items with a mean of one-half of the items would require a cutoff that would only allow a student to miss one item. This would be a very stringent requirement, considering a reasonable standard error of measurement, on the subscale.

Depth-of-Knowledge 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. *Depth-of-knowledge 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 items corresponding to an objective had 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 60% or higher would require the student to successfully answer at least some items at or above the depth-of-knowledge level of the corresponding objectives. For example, 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—i.e., 67% of the items. If three, 50% of the six items, were at or above the depth-of-knowledge 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 depth-of-knowledge of one objective. Some leeway was used in this analysis on this criterion. If between 40% and 50% of the items on a standard were at or

above the depth-of-knowledge levels of the objectives, then it was reported that the criterion was "weakly" met.

Interpreting and assigning depth-of-knowledge levels to both objectives within standards and assessment items is an essential requirement of alignment analysis. These descriptions help to clarify what the different levels represent in mathematics:

Level 1 (Recall) includes the recall of 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. In science, a simple experimental procedure, including one or two steps, should be coded as Level 1. 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) 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. Keywords 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 objective of the action. For example, if an item required students to explain how light affects mass by indicating a relationship between light and heat, this was considered a Level 2. Interpreting information from a simple graph and requiring reading information from the graph are also a 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 a Level 3. Caution is warranted in interpreting Level 2 as only skills because some reviewers will interpret skills very narrowly, as primarily numerical skills, and such interpretation excludes from this level other skills, such as visualization skills and probability skills, which may be more complex simply because they are less common. Other Level 2 activities include explaining the purpose and use of experimental procedures; carrying out experimental procedures; making observations and collecting data; classifying, organizing, and comparing data; and organizing and displaying data in tables, graphs, and charts.

Level 3 (Strategic Thinking) 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 a Level 3 attribute. 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 a Level 3.
Other Level 3 activities include drawing conclusions from observations; citing evidence and developing a logical argument for concepts; explaining phenomena in terms of concepts; and using concepts to solve problems.

Level 4 (Extended Thinking) 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 as a Level 2. However, if the student is to conduct a river study that requires taking into consideration a number of variables, this would be a 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 have to select one approach among many alternatives on how the situation should be solved, in order to be at this highest level. Level 4 activities include designing and conducting experiments; making connections between a finding and related concepts and phenomena; combining and synthesizing ideas into new concepts; and critiquing experimental designs.

Range-of-Knowledge Correspondence

For standards and assessments to be aligned, the breadth of knowledge on both should be comparable. *The range-of-knowledge 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 in order to correctly answer the assessment items/activities*. The criterion for correspondence between span of knowledge for a standard and the assessment considers the number of objectives within the standard with one related assessment item/activity. At least 50% of the objectives for a standard had 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 need to have 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 an assessment to 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 had a

corresponding assessment item, then the range-of-knowledge criterion was met. If between 40% to 50% of the objectives for a standard had a corresponding assessment item, the criterion was "weakly" met.

Balance of Representation

In addition to comparable depth and breadth of knowledge, aligned standards and assessments require the knowledge to be distributed equally in both. The range-of-knowledge criterion only considers the number of objectives within a standard hit (a standard with a corresponding item), but does not take into consideration how the hits (or assessment items/activities) were 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—i.e., one related assessment item/objective. The index is computed by considering the difference in the proportion of objectives and the proportion of hits assigned to the objective. An index value of 1 signifies perfect balance and is obtained if the hits (items/assessment) 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 (items/assessment) were on 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 values between .6 and .7 indicate the balance-of-representation criterion has only been "weakly" met.

Source-of-Challenge Criterion

The source-of-challenge criterion is only used to identify items on which the major cognitive demand is inadvertently placed and is other than the targeted language arts skill, concept, or application. Cultural bias or specialized knowledge could be reasons for an item to have a source-of-challenge problem. Such item characteristics may cause some students to not answer an assessment item, or answer an assessment item incorrectly or at a lower level, even though they have the understanding and skills being assessed.

Findings

Reviewers rated the depth-of-knowledge levels of individual items with moderate to high consistency. The average measure of intraclass correlations (Shrout & Fleiss, 1979), which compared the ratings of the four reviewers within each group, generally were .75 and higher (Table 1). Of the ten group ratings, eight produced an alpha of .75 or higher. For two of the analyses, State E grade 7 and State G grade 4, a second group

coded the assessment items independently of the first group. The results for these two groups are denoted in Table 1 as a replication analysis. The differences in the number of items between the original analysis and the replication analysis indicates that a few of the reviewers did not code an item. A sample set of tables of data reported in an alignment analysis is given in the appendix.

Table 1
Reliability of Depth-of-Knowledge Levels Ratings of Items for States E, F, and G in Mathematics

Grade	Number of Reviewers	Number of Items	Alpha*	95% CI Lower-Upper
State E Mathematics				
4	4	61	.47	.12-.68
7	4	61	.93	.89-.95
7 (Replication)	4	63	.77	.66-.85
9	4	63	.81	.72-.88
State F Mathematics				
5	4	49	.79	.67-.87
8	4	46	.58	.34-.75
State G Mathematics				
4	4	48	.88	.81-.93
4 (Replication)	4	52	.89	.84-.93
8	4	146	.86	.82-.89
11	4	56	.91	.86-.94

* Average Measure Intraclass Correlation

Categorical Concurrence

Only one of the three states achieved an acceptable level for the categorical concurrence criterion on all of the standards at one or more grades (Table 2). Each of the other two states had at least one standard where there was an inadequate number, less than six, of corresponding items on the assessment. State E met the categorical concurrence criterion on all four grade 7 standards and on the one grade 9 standard. The grade 9 standard only addressed algebra, as did all of the 65 items on the assessment. Only four of the seven State G standards for grade 11 were judged by the reviewers to have, on the average, six or more corresponding items. The test was heavily weighted with items measuring algebra (Standard IV), geometry (Standard II), statistics and probability (Standard V), and problem solving and mathematical reasoning (Standard VII). Two standards had less than six corresponding items—number operations and concepts (Standard I, 4.25 items) and measurement (Standard III, 3.75 items). One standard, tools and technology (Standard VI), did not have any corresponding items.

A replication study was done for two states, one grade each (Table 3). Overall, the two groups of reviewers agreed on the acceptable level for categorical concurrence criterion on 82% of the standards, had partial agreement on one standard, and no agreement on one standard.

Table 2
Percent of Standards with an Acceptable Level on the Categorical Concurrence Criterion—States E, F, and G Mathematics Standards and Assessments

State	Grade	Number of Standards	Number of Items	Number of Hits	% Acceptable (Weak) Categorical Concurrence
			Elementary		
State E	4	4	65	65.25	75%
State F	5	6	50	51.75	83.33%
State G	4	7	52	63.75	71% (14%)
			Middle School		
State E	7	4	65	66.00	100%
State F	8	4	50	50.00	75%
State G	8	7	150	179.00	86%
			High School		
State E	9	1	65	64.50	100%
State G	11	7	60	67.50	57%

Table 3
Agreement Between Two Groups of Reviewers on Categorical Concurrence Criterion—Mathematics

State	Grade	Number of Standards	Exact Agreement on Standards		Partial Agreement on Standards[1]		No Agreement on Standards	
			N	%	N	%	N	%
State G	4	7	5	71	1	14	1	14
State E	7	4	4	100				
Total		11	9	82	1	9	1	9

Note 1: Partial agreement is defined by the results of one group indicating a Weak Alignment and the other group indicating Alignment or No Alignment.

A number of issues are related to whether an assessment and a standard meet an acceptable level on the categorical concurrence criterion. Reviewers were allowed to code an item as corresponding to up to two secondary objectives, in addition to the primary objective. Thus, the number of hits would increase, and the opportunity to meet the acceptable level, if reviewers coded an item as corresponding to multiple objectives.

There are different reasons why an item would be coded as corresponding to more than one objective. The item may require students to apply knowledge from more than one topic, such as an item requiring them to compute and apply statistics. Performance assessment and open-ended items are more likely to require students to demonstrate their knowledge of more than one topic. The structure of the standards may be another reason why items have multiple hits. Some mathematics standards include both *process standards*—problem solving, reasoning, communication—and *content standards*—number, geometry, etc. With process and content standards, an item requiring students to solve a complex problem could be coded as corresponding to two or three standards. A reviewer's indecision could be another reason an item is coded to more than one objective. A reviewer may not be able to decide between which two objectives an item measures and code the item as corresponding to both.

In Table 2, by comparing for each grade level and set of standards the number of hits with the number of items, it is easy to estimate the number of items with multiple hits. All of the items for all of the states were multiple-choice items. Only State G grade 8 had a noticeable number of multiple hits, 29 on the average, or 19% of the items. Even with this number of additional hits and a large total number of hits, one standard (Tools and Technology) received less than one hit. Although not recognized before the analysis, it is clear from the results, and later confirmed by the state, that the assessment was not designed to measure this one standard. A major reason the two groups of reviewers disagreed on one standard in the replication study was because one group used secondary hits corresponding to the problem solving and reasoning standard along with content standards. The other group did not use secondary hits and did not code any items as corresponding to the problem solving and reasoning standard.

It is not always clear what standards are to be measured by a test. A state may have high school standards, but only cover some of the standards on the grade 11 test. This appears to be the case for State G grade 11. Content analysis does point to potential issues of how assessment items are distributed among the set of standards. A state may have reasons for not assessing a standard on a test—e.g., a state's standards may cover grade ranges—K-4, 5-8, and 9-12—while the test at a specific grade level is only designed to measure select standards. This implies that additional information may be required to fully explain the distribution of items and to understand the relationship between the test and the standards in the context of a standards-based system.

Depth-of-Knowledge Consistency

Nearly all of the standards and assessments analyzed failed to fully meet an acceptable level on the depth-of-knowledge consistency criterion (Table 4). This means that more than half of the objectives under the standards required a more complex depth-of-knowledge than the corresponding items. Only State E with one grade 9 standard had an acceptable level for all of the standards. Most of the other states and grade levels met the criterion for all but one of the standards. Only state F grade 8 and state G grade 11 failed to meet an acceptable level on more than one standard.

Table 4
Percent of Standards with an Acceptable Level on the Depth-of-Knowledge Consistency Criterion— States E, F, and G Mathematics Standards and Assessments

State	Grade	Number of Standards	Number of Items	Number of Hits	% Acceptable (Weak) Depth-of-Knowledge Consistency
			Elementary		
State E	4	4	65	65.25	75% (25%)
State F	5	6	50	51.75	83% (17%)
State G	4	7	52	63.75	86%[1]
			Middle School		
State E	7	4	65	66.00	75%
State F	8	4	50	50.00	25% (25%)
State G	8	7	150	179.00	86%[1]
			High School		
State E	9	1	65	64.50	100%
State G	11	7	60	67.50	57%[1] (14%)

Note 1: The number of hits on one standard was less than one and insufficient to rate the standard on this criterion.

In the replication study, the two groups that coded the assessment items for two states, one grade each, did not have strong agreement on the proportion of standards that met depth-of-knowledge consistency. The two groups only had exact agreement on 55% of the standards, partial agreement on 22% of the standards, and no agreement on 22% of the standards. Since analysis compares the depth-of-knowledge level of an item in relationship to the depth-of-knowledge level of an objective—under, at, or above—groups could vary in assigning a depth-of-knowledge level to an objective or to an item. However, the groups were consistent when they independently coded the depth-of-knowledge levels of the objectives. For example, in coding the depth-of-knowledge levels of the 30 objectives for State G grade 4, one group coded 39% as Level 1 (Recall), 52% (skills and concepts), and 9% (strategic thinking). The other group coded 42% as Level 1, 49% as Level 2, and 9% as Level 3. Where the groups differed was in coding the depth-of-knowledge levels of the items and mainly in the distinction between Levels 1 and 2. For example, for State G grade 4, both groups coded about the same number of items as corresponding to objectives under the Statistics and Probability standard, 12 and 13. However, one group coded 5 items (46%) as Level 1 while the other group coded 8 items (64%) as Level 1. The difference of three items was sufficient to vary the results with respect to the acceptable level. Within groups reviewers were consistent in assigning the depth-of-knowledge levels (Table 1). However, between groups the consistency was lower than desired, indicating the need for additional training on assigning levels to items.

Table 5
Agreement between Two Groups of Reviewers on Depth-of-Knowledge Consistency Criterion—Mathematics

State	Grade	Number of Standards	Exact Agreement on Standards		Partial Agreement on Standards		No Agreement on Standards	
			N	%	N	%	N	%
State G	4	5[1]	2	40	2	40	1	20
State E	7	4	3	75			1	25
Total		9	5	55	2	22	2	22

Note 1: Both groups coded an insufficient number of items for one standard and one group did the same for a second standard to rate this criterion.

Range-of-Knowledge Correspondence

For a standard to be judged as having an acceptable level on the range-of-knowledge correspondence criterion, 50% or more of the objectives under a standard had to have at least one corresponding item, or hit. One state, for two grade levels, met this specification for all standards, State E grade 7 and grade 9 (Table 6). On the other analyses, generally all standards met this criterion except for one. State F grade 8 only fully met the range-of-knowledge correspondence criterion on one of four standards and weakly met it on one other standard. On the other two standards, the items were distributed among less than half of the objectives.

In the replication study, there was high agreement between the two groups in the number of standards that met the range-of-knowledge correspondence criterion. The two groups had exact agreement on all of the standards that this criterion was met.

Table 6
Percent of Standards with an Acceptable Level on the Range-of-Knowledge Correspondence Criterion— States E, F, and G Mathematics Standards and Assessments

State	Grade	Number of Standards	Number of Items	Number of Hits	% Acceptable (Weak) Range of Knowledge Correspondence
			Elementary		
State E	4	4	65	65.25	75% (25%)
State F	5	6	50	51.75	83% (17%)
State G	4	7	52	63.75	87%[1]
			Middle School		
State E	7	4	65	66.00	100%
State F	8	4	50	50.00	25% (25%)
State G	8	7	150	179.00	71%[1]
			High School		
State E	9	1	65	64.50	100%
State G	11	7	60	67.50	87%[1]

Note 1: The number of hits on one standard was less than 1 and insufficient to rate the standard on this criterion.

Table 7
Agreement Between Two Groups of Reviewers on Range-of-Knowledge Correspondence Criterion— Mathematics

State	Grade	Number of Standards	Exact Agreement on Standards		Partial Agreement on Standards		No Agreement on Standards	
			N	%	N	%	N	%
State G	4	5[1]	5	100				
State E	7	4	4	100				
Total		9	9	100				

Note 1: Both groups coded an insufficient number of items for one standard and one group did the same for a second standard to rate this criterion.

Balance of Representation

Whereas range considers the breadth of the objectives under a standard with at least one corresponding item, balance of representation is related to the degree of emphasis. The underlying assumption is that items should be spread evenly among the objectives under a standard. There may be reasons for one objective to be measured with a greater number of items than another objective, particularly if the former is more

complex. However, if an objective is to be weighed more heavily on an assessment, students and teachers should be informed of this emphasis. State E at grade 7 and 9 met the acceptable level on the balance-of-representation criterion, computed using an index (Table 8). State E grade 4 and State F grade 5 met this criterion for all but one of its six standards. For the other four analyses, the balance-of-representation criterion was not met by two standards. A frequent reason for an assessment and standard not meeting this criterion is that the assessment items corresponding to one standard tend to be all of one type—e.g., an algebra standard for all of the item requiring students to solve simple linear equations of the same format.

Table 8
Percent of Standards with an Acceptable Level on the Balance of Representation Criterion—States E, F, and G Mathematics Standards and Assessments

State	Grade	Number of Standards	Number of Items	Number of Hits	% Acceptable (Weak) Balance of Representation
			Elementary		
State E	4	4	65	65.25	75% (25%)
State F	5	6	50	51.75	83% (17%)
State G	4	7	52	63.75	71%[1]
			Middle School		
State E	7	4	65	66.00	100%
State F	8	4	50	50.00	50%
State G	8	7	150	179.00	71%[1]
			High School		
State E	9	1	65	64.50	100%
State G	11	7	60	67.50	71%[1]

Note 1: The number of hits on one standard was less than one and insufficient to rate the standard on this criterion.

In the replication study, the two groups had exact agreement on 78% of the nine standards and partial agreement on 22% of the standards in judging the balance of representation. This is reasonable agreement, considering the relatively small number of standards included in the analysis. Lack of agreement between the two groups of reviewers generally would be the result of a variation in the coding of two or three items.

Table 9
Agreement Between Two Groups of Reviewers on the Balance of Representation Criterion—Mathematics

State	Grade	Number of Standards	Exact Agreement on Standards		Partial Agreement on Standards		No Agreement on Standards	
			N	%	N	%	N	%
State G	4	5[1]	3	60	2	40		
State E	7	4	4	100				
Total		9	7	78	2	22		

Note 1: Both groups coded an insufficient number of items for one standard and one group did the same for a second standard to rate this criterion.

Discussion

This content analysis presents an approach to judging the alignment between curriculum standards and assessments. The analysis is based on four criteria used to judge the relationship between a set of standards and an assessment designed to measure students' attainment of those standards—categorical concurrence, depth-of-knowledge consistency, range-of-knowledge correspondence, and balance of representation. A number of issues or findings implicit in the relationship between standards and an assessment emerged in the analysis:

- Assessment items corresponding to more than one objective;
- Assessments addressing only a part of a set of standards;
- The depth-of-knowledge level of a set of assessment items corresponding exactly to the depth-of-knowledge level of the corresponding objectives and not to a range;
- Assessment items measuring only a small fraction of the objectives under a standard;
- The majority of the assessment items corresponding to only one or two objectives under a standard and only one or two items corresponding to other objectives.

The appropriateness of these findings or issues and how they should be resolved depends, in part, on what is viewed as a good assessment. Those who favor using a greater number of open-ended or performance assessment items will believe that most items should measure more than one objective. This alignment analysis brings out these and other issues, which must be addressed.

None of the state standards and assessments met all four of the criteria except for State E grade 9, which only had one standard. All of the other state standards and assessments failed to meet at least one of the criteria, indicating that the alignment could be improved. For one set of standards and assessment, there was variation in the degree to which each of the criteria was met, indicating that the criteria address different

attributes of the alignment. The replication study identified some lack of consistency in the coding process, mainly in coding the depth-of-knowledge level. This lack of consistency in assigning a depth-of-knowledge level of two or three items was attributable to insufficient training of the reviewers. Overall, the process was judged to produce valuable information on the alignment between standards and assessments that can provide states with feedback for improving the way in which these documents work in common toward higher student achievement.

References

Carroll, J. B. (1963). A model for school learning. *Teachers College Record, 64*, 723-733.

Cohen, S. A. (1987). Instructional alignment: Searching for a magic bullet. *Educational Researcher, 16*(8), 16-20.

Cosortium for Policy Research in Education. (1991). *Putting the pieces together: Systemic school reform* (CPRE Policy Briefs). New Brunswick, NJ: Rutgers, the State University of New Jersey, Eagleton Institute of Politics.

National Council of Teachers of Mathematics. (1989). *Curriculum and evaluation standards for school mathematics.* Reston, VA: Author.

Newmann, F. M. (1993). Beyond common sense in educational restructuring: The issues of content and linkage. *Educational Researcher, 22*(2), 4-13, 22.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86,* 2, 420-428.

Smith, M. S., & O'Day, J. (1991). Systemic school reform. In S. H. Fuhrman & B. Malen (Eds.), *The politics of curriculum and testing* (pp. 233-267). Bristol, PA: Falmer.

Subkoviak, M. J. (1988). A practitioner's guide to computation and interpretation of reliability indices for mastery tests. *Journal of Educational Measurement, 25*(1), 47-55.

U.S. Congress, House of Representatives. (1994, September 28). *Improving America's Schools Act*. Conference Report to accompany H. R. 6 Report 103-761.Washington, DC: U.S. Government Printing Office.

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

Webb, N. L. (1999). *Alignment of science and mathematics standards and assessment in four states* (NISE Research Monograph No.18). Madison: University of Wisconsin–Madison, National Institute for Science Education. Washington, DC: Council of Chief State School Officers.

APPENDIX

Set of Tables Reporting Data from the Alignment Analysis
State F Grade 8

Brief Explanation of Data in the Alignment Tables by Column

Tables M8-1

Goals #	Number of goals (second level) for each standard.
Objs #	Average number of objectives (third level) for reviewers. If the number is greater than the actual number in the standard, then at least one reviewer coded an item for the goal/objective but did not find any objective in the goal that corresponded to the item.
Level	The Depth-of-Knowledge level coded by the reviewers for the objectives for each standard.
# of objs by Level	The number of objectives coded at each level
% w/in std by Level	The percent of objectives coded at each level
Hits	
Mean & SD	Mean and standard deviation number of items reviewers coded as corresponding to standard. The total is the total number of coded hits.
Cat. Conc. Accept.	"Yes" indicates that the standard met the acceptable level for criterion. "Yes" if mean is six or more. "Weak" if mean is five to six. "No" if mean is less than five.

Tables M8-2

First eight columns are the same as Table 1.

Level of Item w.r.t. Stand	Mean percent and standard deviation of items coded as "under" the Depth-of-Knowledge level of the corresponding objective, as "at" (the same) the Depth-of-Knowledge level of the corresponding objective, and as "above" the Depth-of-Knowledge level of the corresponding objective.
Depth-of-Know. Consistency Accept.	"Yes" indicates that 50% or more of the items were rated as "at" or "above" the Depth-of-Knowledge level of the corresponding objectives. "Weak" indicates that 45% to 50% of the items were rated as "at" or "above" the Depth-of-Knowledge level of the corresponding objectives. "No" indicates that less than 45% items were rated as "at" or "above" the Depth-of-Knowledge level of the corresponding objectives.

Tables M8-3

First eight columns are the same as Table 1 and 2.

Range of Objectives	
# Objs Hit	Average number and standard deviation of the objectives hit coded by reviewers.
% of Total	Average percent and standard deviation of the total objectives that had at least one item coded.
Range of Know.	
Accept.	"Yes" indicates that 50% or more of the objectives had at least one coded objective. "Weak" indicates that 40% to 50% of the objectives had at least one coded objective. "No" indicates that 40% or less of the objectives had at least one coded objective.
Balance Index	
% Hits in Std/Ttl Hits	Average and standard deviation of the percent of the items hit for a standard of total number of hits (see total under the Hits column).
Index	Average and standard deviation of the Balance Index.
Bal. of Rep Accept.	"Yes" indicates that the Balance Index was .7 or above (items evenly distributed among objectives). "Weak" indicates that the Balance Index was .55 to .7 (a high percentage of items coded as corresponding to two or three objectives). "No" indicates that the Balance Index was .55 or less (a high percentage of items coded as corresponding to one objective.)

Tables M8-4

Summary if standard met the acceptable level for the four criteria by each standard.

Table M8-1
Categorical Concurrence Between Standards and Assessment as Rated by Four Reviewers
State F Grade 8 Mathematics
(Number of Assessment Items—50 Multiple Choice Items)

Standards			Level by Objective			Hits		Categorical Concurr. Acceptable
Title	Goals #	Objs #	Level	# of objs by Level	% w/in std by Level	Mean	S.D.	
I. Number Sense	2	5.5[1]	1 2	3 3	50 50	16	1.41	YES
II. Algebraic Operations	3	8	1 2	3 5	38 62	14.5	.50	YES
III. Geometry-Solid	1	3	2	3	100	3.75	.43	NO
IV. Data Analysis & Statistics	1	6.25[1]	1 2 3	1 5 1	14 72 14	15.75	.43	YES
Total	7	22.75	1 2 3	7 16 1	29 67 4	50.00	1.22	

[1]Includes one generic objective because coded items did not correspond to existing objectives.

Table M8-2
Depth-of-Knowledge Consistency Between Standards and Assessment
as Rated by Four Reviewers
State F Grade 8 Mathematics
(Number of Assessment Items—50 Multiple Choice Items)

Standards			Level by Objective			Hits		Level of Item w.r.t. Standard						Depth-of-Knowledge Consistency Acceptable
								% Under		% At		% Above		
Title	Goals #	Objs #	Level	# of objs by Level	% w/in std by Level	M	S.D.	M	S.D.	M	S.D.	M	S.D.	
I. Number Sense	2	5.5[1]	1 2	3 3	50 50	16	1.41	49	42	35	40	15	36	WEAK
II. Algebraic Operations	3	8	1 2	3 5	38 62	14.5	.50	35	44	58	46	8	27	YES
III. Geometry-Solid	1	3	2	3	100	3.75	.43	83	29	17	29	0	0	NO
IV. Data Analysis & Statistics	1	6.25[1]	1 2 3	1 5 1	14 72 14	15.75	.43	71	35	29	35	0	0	NO
Total	7	22.75	1 2 3	7 16 1	29 67 4	50.00	1.22	54	43	39	42	7	26	

[1]Includes one generic objective because coded items did not correspond to existing objectives.

Table M8-3
Range-of-Knowledge Correspondence and Balance of Representation Between Standards and Assessment as Rated by Four Reviewers
State F Grade 8 Mathematics
(Number of Assessment Items—50 Multiple Choice Items)

Standards			Level by Objective Level 1=Recall Level 4=Complex Reasoning			Hits		Range of Objectives				Range of Know. Accept.	Balance Index (1 perfect-0 no Balance)				Balance of Representation Acceptable
								# Objs Hit		% of Total			% Hits in Std/Ttl Hits		Index		
Title	Goals #	Objs #	Level	# of objs by Level	% w/in std by Level	Mean	S.D.	Mean	S.D.	Mean	S.D.		Mean	S.D.	Mean	S.D.	
I. Number Sense	2	5.5[1]	1 2	3 3	50 50	16	1.41	3.25	.43	59	6	YES	32	2	.50	.02	NO
II. Algebraic Operations	3	8	1 2	3 5	38 62	14.5	.50	3.25	.43	41	5	NO	29	1	.47	.01	NO
III. Geom.-Solid	1	3	2	3	100	3.75	.43	1.00	0	33	0	NO	8	1	1.00	0.0	YES
IV. Data Analysis & Statistics	1	6.25[1]	1 2 3	1 5 1	14 72 14	15.75	.43	3.00	.71	48	9	WEAK	31	0	.76	.10	YES
Total	7	22.75	1 2 3	7 16 1	29 67 4	50.00	1.22	2.63	1.05	45	11		25	10	.68	.22	

1 Includes one generic objective because coded items did not correspond to existing objectives.

Table M8-4
Summary of Attainment of Acceptable Alignment Level on Four Content Focus Criteria
State F Grade 8 Mathematics
(Number of Assessment Items—50 Multiple Choice Items)

Standards	Alignment Criteria			
	Categorical Concurrence	Depth-of-Knowledge Consistency	Range of Knowledge	Balance of Representation
I. Number Sense	YES	WEAK	YES	NO
II. Algebraic Operations	YES	YES	NO	NO
III. Geometry-Solid	NO	NO	NO	YES
IV. Data Analysis & Statistics	YES	NO	WEAK	YES