An Analysis of the Alignment Between Language Arts Standards and Assessments for Four States

by

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

Marshá Horton
Consultant
Dover, Delaware

Sharon O'Neal
Southwest Texas State University, San Marcos

April 2, 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. Earlier work in developing the alignment criteria was supported in part by a cooperative agreement between the National Science Foundation and the University of Wisconsin-Madison (Cooperative Agreement No. RED-9452971). Any opinions, findings, or conclusions are those of the authors and do not necessarily reflect the views of the support agency or of WCER.

An Analysis of the Alignment Between Language Arts Standards and Assessments for Four States

by

Norman L. Webb
Marshá Horton
Sharon O'Neal

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 in mathematics and science (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 language arts standards and assessments from four states, two teams of specialists analyzed mathematics standards and assessments from three of the four states. In language arts, the alignment of standards and assessments was analyzed for three grade levels for three states and four grade levels for one state. This report only attends to the language arts analysis. The grade levels analyzed in this study by state were:

State E	Grades 4, 7, and 10
State F	Grades 5, 8, and 11
State G	Grades 4, 8, and 11
State H	Grades 4, 5, 6, and 9

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' first task was to become familiar with the state standards to be included in the alignment analysis. This required a preliminary look at the standards to understand their structure, what goals were included under each standard and what objectives were included under each goal. Reviewers were then trained to identify the depth-of-knowledge of objectives and assessment items. Depth of knowledge consistency is the cornerstone for the determination of alignment and is evident 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. Training on identifying depth of knowledge 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 the 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. They then compared their individual assignments of content objective and item depth of knowledge with those of others in the group. 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.

States use a variety of labels for identifying levels of expectations. For the purposes of this analysis, we have employed the convention of standards, goals, and objectives to describe three levels of expectations for what students are to know and do. Standard is the most general; it is divided into goals, which are further subdivided into objectives. It is assumed that all of the goals under a standard span the content knowledge expressed in the standard and all of the objectives under a goal span the content knowledge expressed in the goal.

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.

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

This report describes the results of an alignment study of standards and grade-level tests in language arts for four 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.

Language Arts 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. As to what should be the appropriate acceptable level for any of the criteria depends on many factors. For this analysis, an acceptable level was based on a judgment of what number of items or what proportion of objectives under a standard would need to have corresponding items to decide if a student had demonstrated an adequate amount of knowledge to determine if the student had met the standard.

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.

In language arts, four depth-of-knowledge levels were used to judge both reading and writing objectives and assessment tasks. The reading levels are based on Valencia and Wixson (2000, pp. 909-935) and Wixson, Fisk, Dutro, & McDaniel (1999). The writing levels were developed by Marshá Horton, Sharon O’Neal, and Phoebe Winter, consultants to the project.

Reading

Reading Level 1

Level 1 requires students to receive or recite facts or to use simple skills or abilities. Oral reading that does not include analysis of the text as well as basic comprehension of a text is included. Items require only a shallow understanding of the text presented and often consist of verbatim recall from text, or simple understanding of a single word or phrase. Some examples that represent, but do not constitute all of, Level 1 performance are:

- Support ideas by reference to details in the text.
- Use a dictionary to find the meanings of words.
- Identify figurative language in a reading passage.

Reading Level 2

Level 2 includes the engagement of some mental processing beyond recalling or reproducing a response; it requires both comprehension and subsequent processing of text or portions of text. Inter-sentence analysis of inference is required. Some important concepts are covered but not in a complex way. Standards and items at this level may include words such as summarize, interpret, infer, classify, organize, collect, display, compare, and determine whether fact or opinion. Literal main ideas are stressed. A Level 2 assessment item may require students to apply skills and concepts that are covered in Level 1. Some examples that represent, but do not constitute all of, Level 2 performance are:

- Use context cues to identify the meaning of unfamiliar words.
- Predict a logical outcome based on information in a reading selection.
- Identify and summarize the major events in a narrative.

Reading Level 3

Deep knowledge becomes a greater focus at Level 3. Students are encouraged to go beyond the text; however, they are still required to show understanding of the ideas in the text. Students may be encouraged to explain, generalize, or connect ideas. Standards and items at Level 3 involve reasoning and planning. Students must be able to support their thinking. Items may involve abstract theme identification, inference across an entire passage, or students' application of prior knowledge. Items may also involve more superficial connections between texts. Some examples that represent, but do not constitute all of, Level 3 performance are:

- Determine the author's purpose and describe how it affects the interpretation of a reading selection.
- Summarize information from multiple sources to address a specific topic.
- Analyze and describe the characteristics of various types of literature.

Reading Level 4

Higher-order thinking is central and knowledge is deep at Level 4. The standard or assessment item at this level will probably be an extended activity, with extended time provided for completing it. The extended time period is not a distinguishing factor if the required work is only repetitive and does not require the application of significant conceptual understanding and higher-order thinking. Students take information from at least one passage of a text and are asked to apply this information to a new task. They may also be asked to develop hypotheses and perform complex analyses of the connections among texts. Some examples that represent, but do not constitute all of, Level 4 performance are:

- Analyze and synthesize information from multiple sources.
- Examine and explain alternative perspectives across a variety of sources.
- Describe and illustrate how common themes are found across texts from different cultures.

Writing

Writing Level 1

Level 1 requires the student to write or recite simple facts. The focus of this writing or recitation is not on complex synthesis or analysis but on basic ideas. The students are asked to list ideas or words, as in a brainstorming activity prior to written composition; are engaged in a simple spelling or vocabulary assessment; or are asked to write simple sentences. Students are expected to write and speak using the conventions of Standard English. This includes using appropriate grammar, punctuation, capitalization, and spelling. Some examples that represent, but do not constitute all of, Level 1 performance are:

- Use punctuation marks correctly.
- Identify Standard English grammatical structures and refer to resources for correction.

Writing Level 2

Level 2 requires some mental processing. At this level, students are engaged in first-draft writing or brief extemporaneous speaking for a limited number of purposes and audiences. Students are expected to begin connecting ideas, using a simple organizational structure. For example, students may be engaged in note-taking, outlining, or simple summaries. Text may be limited to one paragraph. Students demonstrate a basic understanding and appropriate use of such reference materials as a dictionary, thesaurus, or web site. Some examples that represent, but do not constitute all of, Level 2 performance are:

- Construct compound sentences.
- Use simple organizational strategies to structure written work.
- Write summaries that contain the main idea of the reading selection and pertinent details.

Writing Level 3

Level 3 requires some higher-level mental processing. Students are engaged in developing compositions that include multiple paragraphs. These compositions may include complex sentence structure and may demonstrate some synthesis and analysis. Students show awareness of their audience and purpose through focus, organization, and the use of appropriate compositional elements. The use of appropriate compositional elements includes such things as addressing chronological order in a narrative or including supporting facts and details in an informational report. At this stage, students are engaged in editing and revising to improve the quality of the composition. Some examples that represent, but do not constitute all of, Level 3 performance are:

- Support ideas with details and examples.

- Use voice appropriate to the purpose and audience.
- Edit writing to produce a logical progression of ideas.

Writing Level 4

Higher-level thinking is central to Level 4. The standard at this level is a multi-paragraph composition that demonstrates the ability to synthesize and analyze complex ideas or themes. There is evidence of a deep awareness of purpose and audience. For example, informational papers include hypotheses and supporting evidence. Students are expected to create compositions that demonstrate a distinct voice and that stimulate the reader or listener to consider new perspectives on the addressed ideas and themes. An example that represents, but does not constitute all of, Level 4 performance is:

- Write an analysis of two selections, identifying the common theme and generating a purpose that is appropriate for both.

A few samples will help illustrate the depth-of-knowledge levels:

Objective	DOK Level
Identify idioms, analogies, metaphors, and similes in prose and poetry. (Grade 7)	1
Follow oral instructions that provide information about a task or assignment. (Grade 5)	2
Students use comprehension strategies to make predictions, identify the main idea and supporting details, compare and contrast, and summarize. (Grade 4)	2
Identify social/culture values and beliefs reflected in literature and media. (Grade 5)	3
Students explain how perspectives and purposes define or influence forms of media (i.e., biases, points of view, sensationalism, entertainment, information, and persuasion). (Grade 8)	4

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 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 an assessment considers the number of objectives within the standard with one related assessment item/activity. Fifty percent 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 41% to 49% 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 that knowledge 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); 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—i.e., one related assessment item per 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 (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 are 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 result in some students not answering an assessment item, or answering an assessment item incorrectly, or at a lower level, even though they possess 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 three to six reviewers within each group, generally were .70 and higher (Table 1).

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

Grade	Number of Reviewers	Number of Items	Alpha*	95% CI Lower-Upper
State E Language Arts				
4	5	84	.52	.34-.67
7	3	74	.70	.56-.80
10	3	82	.62	.45-.74
State F Language Arts				
5	6	38	.79	.66-.88
8	6	35	.60	.36-.78
11	6	88	.85	.79-.89
State G Language Arts				
4	6	45	.92	.88-.95
8	6	40	.87	.80-.92
11	6	192	.86	.83-.89
State H Language Arts				
4	5	112	.83	.78-.88
5	4	114	.73	.64-.80
6	3	116	.79	.71-.84
9	3	113	.36	.13-54

* Average Measure Intraclass Correlation

Of the 13 group ratings, ten produced an alpha of .70 or higher. There was some interaction among the states and the reviewers' reliability. This is related to the variation in the depth-of-knowledge among items. Some reviewers had difficulty in coding items from State E. There also was a learning factor that generally resulted in reviewers having higher agreement as they became more experienced. One exception was reviewers coding State H grade 9. Twenty items were of the same type on the State H grade 9. Two reviewers coded these items all as a depth-of-knowledge level 1 whereas one coder coded them as a depth-of-knowledge level 2. As a result, 20% of the items were coded as having the same depth-of-knowledge level as the corresponding objective. This discrepancy did not alter the result of the analysis on the depth-of-knowledge correspondence. Twenty percent of the items measuring a standard at or below the depth-of-knowledge level of the corresponding objectives falls way below the acceptable level of 50%. The appendix includes a sample set of tables used to report findings from the analysis.

Categorical Concurrence

Two states achieved an acceptable level for the categorical concurrence criterion on all of the standards for all of the grade levels analyzed, States E and G (Table 2). This indicates that the assessment had six or more items that measured students' knowledge of

Table 2

Percent of Standards with an Acceptable Level on the Categorical Concurrence Criterion States E, F, G, and H Language Arts Standards and Assessments

State	Grade	Number of Standards	Number of Items	Number of Hits	% Acceptable (Weak) Categorical Concurrence
			Elementary		
State E	4	2	91+2OE	104.40	100%
State F	5	3	40	62.33	67%
State G	4	1	45	46.33	100%
State H	4	3	117	125.67	67%
State H	5	3	117+2OE	128.75	67%
			Middle School		
State E	7	2	90+2OE	84.67[1]	100%
State F	8	3	40	53.83	67%
State G	8	1	45	51.17	100%
State H	6	3	117+2OE	128.33	67%
			High School		
State E	10	2	85	85.00	100%
State F	11	5	100	134.33	60%
State G	11	1	194	222.5	100%
State H	9	3	113+2OE	110.67	67%

[1] Reviewers indicated that on the average 16.67 items were not codeable.

each of the standards. Attaining an acceptable level on this criterion was not too difficult considering that the states generally had from 40 to nearly 200 items to distribute among two or three standards. The language are standards for each of the states included:

State G one standard—reading—for all grades,
State E two standards—reading and writing—for all grades,
State F three standards—reading process, responding to text, and information and research—for grades 5 and 8,
State F five standards—reading process; responding to text; information and research; grammar, usage, and mechanics; and literature---for grade 11, and
State H three standards—comprehension, communication, and use of literature—for all four grades.

States F and H had one or two standards that reviewers coded less than six corresponding items. The State F assessments were judged as having less than six items corresponding to the information and research standard and literature standard (grade 11). The State H assessments had little or no items corresponding to the literature standard for all grades.

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 the opportunity to meet an acceptable level. 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 interpret the main idea and determine word usage. Performance assessment and open-ended items are more likely to require students to demonstrate their knowledge of more than one topic. Two states had two open-ended items on their assessments. All of the other items were multiple-choice. The structure of the standards may be another reason why items have multiple hits. Some sets of language arts standards include a standard on reading and one on literature. Generally students will be required to read when responding to questions about literature. In these cases, it may be appropriate for an item to be coded as corresponding to both the reading and literature 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. The analysis does not distinguish between any of these reasons.

Depth-of-Knowledge Consistency

States were less successful in achieving an acceptable level on the depth-of-knowledge consistency criterion (Table 3). Only State E (grade 4) and State G (grades 4 and 8) had 50% or more of the assessment items with a depth-of-knowledge level at or above the depth-of-knowledge level of the corresponding objectives. One third or fewer of the objectives for the language arts standards in any of the states were judged to have a depth-of-knowledge level of 1 (receive, recite, or write simple facts). Most of the objectives were judged to have a depth-of-knowledge level of 2 or 3. However, many of the items using the multiple-choice format were judged to have a depth-of-knowledge levels 1 or 2. States with one or more standards not even weakly meeting an acceptable level on this criterion had over 60% of the items judged as being less complex than what was expected by the standards. States were advised to replace existing items with more demanding items in order to improve the degree of alignment. In some cases only a few of the items needed replacement.

Table 3

Percent of Standards with an Acceptable Level on the Depth-of-Knowledge Consistency Criterion States E, F, G, and H Language Arts Standards and Assessments

State	Grade	Number of Standards	Number of Items	Number of Hits	% Acceptable (Weak) Depth-of-Knowledge Consistency
			Elementary		
State E	4	2	91+2OE	104.40	100%
State F	5	3	40	62.33	33%
State G	4	1	45	46.33	100%
State H	4	3	117	125.67	33%
State H	5	3	117+2OE	128.75	33%
			Middle School		
State E	7	2	90+2OE	84.67[1]	50% (50%)
State F	8	3	40	53.83	(33%)
State G	8	1	45	51.17	100%
State H	6	3	117+2OE	128.33	50%[2]
			High School		
State E	10	2	85	85.00	50%
State F	11	5	100	134.33	25% (25%)[2]
State G	11	1	194	222.5	(100%)
State H	9	3	113+2OE	110.67	0%[2]

[1] Reviewers indicated that on the average 16.67 items were not codeable.

[2] The number of hits on one standard was less than one and insufficient to rate the standard on this criterion.

Range-of-Knowledge Correspondence

Of the thirteen analysis performed, over half of them successfully met an acceptable level on the range-of-knowledge correspondence criterion by having assessment items correspond to 50% or more of the objectives under a standard (Table 4). State E and G met this criterion for all grades. State F met it for one grade. State H did not meet the criterion for any of the four grades being analyzed, but did weakly meet the criterion for one or two standards for three of the grades. An acceptable level on this criterion indicates some distribution of items among the objectives for a standard, but it falls short in determining what proportion of the content that is covered by items for any one objective or standard. Other alignment techniques such as those developed by Achieve or the American Association for the Advancement of Science (AAAS) have incorporated in their alignment analysis some judgment based on the proportion of content in an objective measured by an item. Achieve asks reviewers to come to agreement if an item measures part of the content of the objective or all of the content. AAAS asks reviewers to judge if the item is necessary to measure the content and if the item is sufficient to measure the content. The strength of the process used in the Webb/CCSSO analysis is that reviewers independently code the items as corresponding to objectives so that some judgment can be made about the consistency among reviewers.

Table 4
Percent of Standards with an Acceptable Level on the Range-of-Knowledge Correspondence Criterion
States E, F, G, and H Language Arts Standards and Assessments

State	Grade	Number of Standards	Number of Items	Number of Hits	% Acceptable (Weak) Depth-of-Knowledge Consistency
			Elementary		
State E	4	2	91+2OE	104.40	100%
State F	5	3	40	62.33	33% (67%)
State G	4	1	45	46.33	100%
State H	4	3	117	125.67	(67%)
State H	5	3	117+2OE	128.75	0%
			Middle School		
State E	7	2	90+2OE	84.67[1]	100%
State F	8	3	40	53.83	33% (33%)
State G	8	1	45	51.17	100%
State H	6	3	117+2OE	128.33	(50%[2])
			High School		
State E	10	2	85	85.00	100%
State F	11	5	100	134.33	100%[2]
State G	11	1	194	222.5	100%
State H	9	3	113+2OE	110.67	(50%[2])

1 Reviewers indicated that on the average 16.67 items were not codeable.

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

Balance of Representation

Only two states, each at one grade level, met an acceptable level on the balance of representation criterion on all of the standards, State E grade 10 and State F grade 8 (Table 5). Five of the analyses—State E grades 4 and 7, State F grades 5 and 11, State H grade 5—had all of the standards either fully or weakly meet an acceptable level. On the other analysis, one or more standards had a large proportion of items corresponding to only one or two objectives. For example, State H grade 6 had 21 objectives under its first standard related to comprehension organized under four goals. However, every nearly

80% of the 67 hits, on the average, were coded by the reviewers as corresponding to three (14%) of the objectives. The balance index compares the proportion of items for each objective to the proportion if the items were evenly distributed among all possible objectives, in this case 67 hits among 21 objectives or about 3 items per objective. The State H grade 6 assessment had 13 to 21 items corresponding to only three objectives. This produced a balance index value of .57 well below the .7 value needed to be acceptable. There may be reasons for why a state may want to test more heavily students' knowledge of specific objectives over others. This may be the case if the standard covers a range of grade levels and only certain objectives are taught at the grade level being tested. However, lack of balance can come from a test being created using an item bank with a very limited choice of item types that does not reflect the expectations of the state.

Table 5
Percent of Standards with an Acceptable Level on the
Balance of Representation Criterion
States E, F, G, and H Language Arts Standards and Assessments

State	Grade	Number of Standards	Number of Items	Number of Hits	% Acceptable (Weak) Depth-of-Knowledge Consistency
			Elementary		
State E	4	2	91+2OE	104.40	50% (50%)
State F	5	3	40	62.33	33% (67%)
State G	4	1	45	46.33	0%
State H	4	3	117	125.67	33% (33%)
State H	5	3	117+2OE	128.75	33% (67%)
			Middle School		
State E	7	2	90+2OE	84.67[1]	(100%)
State F	8	3	40	53.83	100%
State G	8	1	45	51.17	0%
State H	6	3	117+2OE	128.33	(50%)[2]
			High School		
State E	10	2	85	85.00	100%
State F	11	5	100	134.33	75% (25%)[2]
State G	11	1	194	222.5	0%
State H	9	3	113+2OE	110.67	(50%)[2]

[1] Reviewers indicated that on the average 16.67 items were not codeable.

[2] The number of hits on one standard was less than one and insufficient to rate the standard on this criterion.

Discussion

This alignment analysis is based on the assumption that a multiple of criteria are necessary to judge the degree that an assessment and standards represent the same content and are likely to send a common message to teachers and others that will lead towards students achieving what is expected of them. The analysis is a content analysis that is based on the judgment of experts. Ideally any analysis would be done with five or more reviewers. Some of the analyses reported here were done with three reviewers. The reliability among the reviewers gives some indications of the consistency among the reviewers. If with three reviewers, a couple of the groups achieved a reliability of .70 or higher. The results of the groups of reviewers with lower reliabilities were checked to see if the variability among the reviewers related to the alignment judgments. In most cases it did not. Variability among the reviewers is incorporated into making the final judgment on if a standard and the assessment met an acceptable level because the average among the reviewers is used.

Information is reported here by criterion. This was intended to study the process and understand more about over all procedures used to study alignment. In the reports to the states the results across the four criteria, along with any source of challenge issues, are summarized and recommendations are given for what needs to be done to achieve a higher degree of alignment. Considering all four criteria, all four states could improve alignment between their standards and assessments in at least some way. State G had alignment on three of the four criteria for all three grade levels. Of course, the degree of alignment was help by only having one standard and a fairly large number of items. State G did not achieve an acceptable level for each of the grade levels on the balance of representation. Even thought the assessment had an adequate number of items at appropriate depth-of-knowledge levels and measuring at least a range of the objectives, a few of the objectives were over emphasized on the assessment. State E with only two standards at each of the grade levels had a reasonably high degree of alignment on all four criteria. The alignment at the three grade levels could be easily improved by replace less than five items with those at a higher depth-of-knowledge level and measuring other objectives currently less emphasized on the assessment. The alignment for States F and H is more problematic. Both these states had three or more standards including one that was more difficult to assess—information and research (State F) and use of literature (State H)—using an on-demand assessment. The assessment only included a very small number of items, if any, that corresponded to these standards. These states have included more demanding standards than the other two, but the assessment does not reflect these higher expectations. In addition to needing more items for standards not currently assessed, the alignment for both of the states can be improved greatly by replacing existing items with those that have a higher depth-of-knowledge level, assess content corresponding to more of the objectives, and more evenly distributing the items among the objectives.

There are a number of issues that have to be considered in judging the alignment among standards and assessments. Each process used to determine alignment has to explicitly or implicitly make assumptions about these issues:

- 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.

It is also apparent from this analysis of language arts standards and assessments that the number of standards is an important factor in achieving alignment. What is striking about studying the alignment in language arts compared to mathematics and science is the high proportion of standards and assessments that did not meet the balance of representation criterion. This indicated that the tests had a large number of items only measuring two or three objectives. Whereas, from the analysis of mathematics assessments and standards, it has been less rare for the balance criterion not to have an acceptable level.

In conclusion, the process does produce information that distinguishes different qualities in relationship between the assessments and standards in the different states. The process does not produce all of the desired information particularly on the degree to which content of an objective or standard is measured by an assessment item. Reviewers are able to code the depth-of-knowledge levels of objectives and items with fairly high consistency. The acceptable levels for each criterion were set with some rationale, but in the end they are only based on judgment. What is an acceptable level still is an open question. The process remains under development. The results from the studies for the CCSSO TILSA group will be used to make revisions to the process. There are principles of testing that are implied in the process that would be useful considerations for assessment development such as attending to the depth-of-knowledge levels. More formal studies of the alignment processes are needed where the different process are used to analyze the same sets of standards and assessment to help determine more what are the differences and similarities among the different methods and how valid each one is.

None of the state standards and assessments met all four of the criteria except for

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.

Valencia, S. W., & Wixson, K. K. (2000). Policy-oriented research on literary standards and assessment. In M. L. Kamil, P. B. Mosenthal, P. D. Pearson, & R. Barr (Eds.), *Handbook of reading research: Vol. III.* Mahwah, NJ: Lawrence Erlbaum.

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.

Wixson, K. K., Fisk, M. C., Dutro, E., & McDaniel, J. (1999). *The alignment of state standards and assessments in elementary reading*. A report commissioned by the National Research Council’s Committee on Title I Testing and Assessment. Ann Arbor: The University of Michigan.

APPENDIX

Set of Tables Reporting Data from the Alignment Analysis
State E Grade 4

Brief Explanation of Data in the Alignment Tables by Column

Tables LA4-1, LA7-1, LA10-1

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 LA4-2, LA7-2, LA10-2

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 40% 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 40% items were rated as "at" or "above" the Depth-of-Knowledge level of the corresponding objectives.

Tables LA4-3, LA7-3, LA10-3

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

Note: BALANCE INDEX $1 - \left(\sum_{k=1} \left| 1/(O) - I_{(k)}/(H) \right|\right)/2$

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

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 .6 to .7 (a high percentage of items coded as corresponding to two or three objectives).

"No" indicates that the Balance Index was less than .6 (a high percentage of items coded as corresponding to one objective.)

Tables LA4-4, LA7-4, LA10-4

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

Table LA4-1
Categorical Concurrence Between Standards and Assessment as Rated by Five Reviewers
State E Grade 4 Language Arts
(Number of Assessment Items—91 Multiple Choice Items & Two Writing Prompts)

Standards			Level by Objective			Hits		Categorical Concurr. Acceptable
Title	Goals #	Objs #	Level	# of objs by Level[1]	% w/in std by Level	Mean	S.D.	
I. Reading	3	18.80[1]	1 2 3 4	6 9 4 1	30 45 20 5	59.00	8.15	YES
II. Writing	2	17.00	1 2 3	8 5 3	50 31 19	45.40	4.93	YES
Total	5	35.80	1 2 3 4	14 14 7 1	39 39 19 3	104.40	8.55	

[1] Includes two generic objectives (1A and 1B) because raters did not find existing matching objective.

Table LA4-2
Depth-of-Knowledge Consistency Between Standards and Assessment
as Rated by Five Reviewers
State E Grade 4 Language Arts
(Number of Assessment Items—91 Multiple Choice Items & Two Writing Prompts)

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[1]	% w/in std by Level	M	S.D.	M	S.D.	M	S.D.	M	S.D.	
I. Reading	3	18.80[1]	1 2 3 4	6 9 4 1	30 45 20 5	59.00	8.15	44	44	44	42	12	30	YES
II. Writing	2	17.00	1 2 3	8 5 3	50 31 19	45.40	4.93	30	41	43	40	27	38	YES
Total	5	35.80	1 2 3 4	14 14 7 1	39 39 19 3	104.4	8.55	38	42	43	38	19	32	

[1] Includes two generic objectives (1A and 1B) because raters did not find existing matching objective.

Table LA4-3
Range-of-Knowledge Correspondence and Balance of Representation Between Standards and Assessment as Rated by Five Reviewers
State E Grade 4 Language Arts
(Number of Assessment Items—91 Multiple Choice Items & Two Writing Prompts)

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[1]	% w/in std by Level	Mean	S.D.	Mean	S.D.	Mean	S.D.		Mean	S.D.	Mean	S.D.	
I. Reading	3	18.8[1]	1 2 3 4	6 9 4 1	30 45 20 5	59.00	8.15	14.80	1.30	79	6	YES	56	4	.67	.05	WEAK
II. Writing	2	17.00	1 2 3	8 5 3	50 31 19	45.40	4.93	12.40	1.14	73	7	YES	44	4	.71	.06	YES
Total/Mean	5	35.80	1 2 3 4	14 14 7 1	39 39 19 3	104.4	8.55	13.60	1.71	76	7		50	.08	.69	.06	

[1] Includes two generic objectives (1A and 1B) because raters did not find existing matching objective.

Table LA4-4
Summary of Attainment of Acceptable Alignment Level on Four Content Focus Criteria
State E Grade 4 Language Arts
(Number of Assessment Items—91 Multiple Choice Items & Two Writing Prompts)

Standards	Alignment Criteria			
	Categorical Concurrence	Depth-of-Knowledge Consistency	Range of Knowledge	Balance of Representation
I. Reading	YES	YES	YES	WEAK
II. Writing	YES	YES	YES	YES