Alignment Between Standards and Assessments in Mathematics for Grade 8 in one State

Norman L. Webb
John Smithson

Wisconsin Center for Education Research
School of Education
University of Wisconsin-Madison

in collaboration with

The Council of Chief State School Officers

Paper presented at the American Educational Research Association Annual meeting held in Montreal, Quebec, Canada, April 19-23, 1999.

This Alignment Study was supported by a grant to the Council of Chief State School Officers provided by the National Science Foundation (award number REC-9803080) and by the National Institute for Science Education, in cooperative agreement between the National Science Foundation and the University of Wisconsin-Madison (Cooperative Agreement No. RED-9452971). The opinions, findings, and conclusions that are expressed in this report do not necessarily reflect those of the supporting agencies.

Contributors to the Report

Norman L. Webb	*Project Director*
Margaret H. Powell	*Editor*
Lynn Lunde	*Secretary*

Reviewers

Rolf Blank
Jennifer Falls
Michael Kestner
Gerald Kuhm
Michael Lower
Curtis McKnight
Linda Wilson

Table of Contents

Executive Summary

Alignment Between Standards and Assessments in Mathematics for Grade 8 in one State

At a four-day Alignment Analysis Institute, June 29-July 2, 1998, reviewers conducted an analysis mathematics standards and assessments for four states. This is a report of the analysis for grade 8 mathematics for one state, State B. Three reviewers conducted the analysis for grade 8. The analysis was structured on the basis of specific alignment criteria that addressed the content focus between the standards and assessments. The codings by reviewers were analyzed and mean results among the reviewers reported. In general, the standards and assessments for grade 8 were found to be aligned on the criteria of categorical concurrence, depth-of-knowledge consistency, range-of-knowledge correspondence, and balance of representation. The only ways the grade 8 standards and assessments could be more aligned would be to replace a few assessment items by those at a higher depth-of-knowledge level for two standards, and to modify or replace a few items relating to another standard so that the set of items would be more evenly distributed among the objectives.

State B
Alignment Between Standards and Assessments in Mathematics for Grade 8 in one State

Norman L. Webb
John Smithson
Wisconsin Center for Education Research

Introduction

Alignment is not a new phenomenon, but has been studied for a number of years. What has changed is the nature of the assessments, expectations, and other system components to be aligned and the stakes for achieving alignment. In the 1960s, analyses were performed on assessment tasks and behavioral objectives as part of the mastery-learning movement (Cohen, 1987; Carroll, 1963). Exact alignment was achieved if the assessment tasks were equivalent to the instructional tasks. Learning goals were partitioned into narrowly defined behavioral objectives. Domains of all possible test items were specified for each behavioral objective. Content analysis by expert panels remains the primary technique for judging alignment between learning expectations and assessments. But with the advent of standards-based education, systemic reform (Smith & O'Day, 1991), and criterion-referenced tests, judging alignment has become more complex and requires more systematic procedures. The underlying assumptions regarding the form of assessments, such as norm-referenced tests and normally distributed achievement, can result in some misalignment with expectations that are targeted for nearly all students (Baker, Freeman, & Clayton, 1991). Educators increasingly recognize that if system components are not aligned, the system will be fragmented, will send mixed messages, and will be less effective (CPRE, 1991; Newmann, 1993; Spillane, 1998). But in addition to conceptual reasons for assuring alignment, states are also faced with legal reasons. The Improving America's Schools Act explicated how assessments are 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 a key performance indicator for states, districts, and schools striving to meet challenging standards. Within the changing climate of what we know about what works in education and the increasing mandates and pressures on education systems, alignment has become critical in fully understanding how systems function. This study was directed toward refining procedures so that they can become more standardized and useful in order for states and districts to better understand the agreement between expectations and assessments

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

A four-day Alignment Analysis Institute was conducted June 29 through July 2, 1998. Sixteen people, including state content specialists, state assessment consultants, content experts, and researchers, attended the institute, which was coordinated by the Council of Chief State School Officers (CCSSO) with the cooperation of the National Institute for Science Education (NISE). Prior to this institute, most participants attended a one-day meeting in Washington, DC, on April 29, to be introduced to the process and to the alignment criteria to be used at the institute. At the summer institute, six of the participants rated mathematics standards and assessments; seven rated science standards and assessments; and three coordinated the process. Four states volunteered to have their mathematics standards and assessments analyzed for alignment for two or three grade levels. Three of these states agreed to have their science standards and assessments analyzed for two or three grade levels.

A major goal of the institute was to develop a systematic process and analytic tools for judging the alignment between standards and assessments based on the criteria developed in conjunction with CCSSO and NISE (Webb, 1997) that are listed in Appendix A. Because of this, reviewers were not given lengthy training in applying the criteria, but were expected to help refine the process over the duration of the institute. One outcome of the institute is a refined process that can be used under more controlled conditions to make a judgment on the alignment of standards and assessments. Reviewers were instructed to attend to the alignment between the state standards and assessments. There was no opportunity for reviewers to offer their opinion on either the quality of the standards or of the assessment activities/items. 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 Analysis 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 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 one state, identified here as State B, and for grade 8. The study addressed specific criteria related to considering 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. Other criteria such as articulation across grades and ages, equity and fairness, and pedagogical implications were given less emphasis. Wixson and her colleagues (Wixson, Fisk, Dutro, & McDaniel, 1999) have successfully applied the four criteria used in this analysis in the content area of reading.

Initial Methodology Developed at the Institute for the Analysis of Alignment Criteria

Three reviewers analyzed the alignment between the standards and the assessment for State B. Prior to analyzing the documents, the reviewers were only given general instructions and broad definitions for the depth-of-knowledge levels required to satisfy a standard and to successfully answer an assessment activity. One purpose for conducting these alignment studies is to better specify what training reviewers need if they are to validly code assessment activities and standards. Reviewers were given the following levels to judge depth of knowledge:

Level

1. *Recall*
   Recall of a fact, information, or procedure

2. *Skill/Concept*
   Use of information, conceptual knowledge, procedures, two or more steps, etc.

3. *Strategic Thinking*
   Requires reasoning, developing a plan or sequence of steps; has some complexity; more than one possible answer; generally takes less than 10 minutes to do.

4. *Extended Thinking*
   Requires an investigation, time to think and process multiple conditions of the problem or task; and requires more than 10 minutes to do other than routine manipulations.

Reviewers within a content area were encouraged to refine these levels or to add greater clarification, providing they all came to some agreement. One of the intended outcomes for this alignment study is greater clarity for the levels. The revised levels are given in the summary report.

Different states use different terminology to label expectations for what students are to know and do. Some states label the large categories of student expectations as "strands." Other states call these expectations "competency goals." Still other states call these expectations "benchmarks." To improve the interpretation of results, the same convention was used in this analysis to label the different levels of expectations. Standards refers to the most general expectations for a grade and content area. The number of standards in the four states that participated in this analysis ranged from four to ten. Goal refers to the next level of specificity of expectations. Generally the set of goals for a standard covers the full range of knowledge specified by the standard. The number of goals for a standard in this analysis went as high as 20. Objective refers to the third level of specificity. Objectives further delineate expectations stated as a goal. The number of expectation levels can vary. In this analysis, a maximum of three levels of expectations were included. If a state only used two levels of expectations, then the most general level is called standards and the second level is called objectives.

Prior to the Alignment Analysis Institute, reviewers were sent copies of the standards and were asked to become familiar with them. At the institute, reviewers began their analysis by assigning a depth-of-knowledge level for each objective for a standard. Achieving one objective could require students to know the content at more than one depth-of-knowledge level.[1] The assigned level was to represent the highest level of knowledge expected for a student to satisfactorily demonstrate the attainment of the objective. All of the reviewers were to reach consensus on the assigned level for each objective (second level of expectations for State B). This activity served two purposes. First, reviewers became more familiar with what students were expected to know and do for each objective. Second, the assigned levels were used as benchmarks to compare the depth-of-knowledge level of individual assessment items/activities.

Reviewers recorded the depth-of-knowledge level for each objective on a coding matrix prepared prior to the institute. The coding matrix listed all of the objectives for student learning for each standard. These expectations were listed in rows in the same order using the same organization as that used in the state's standards document. For each standard, in sequence, the first row listed the standard, the second row a goal, the third row objectives, and so on. Each standard (competency goal for State B) and objective was assigned a unique numerical-alpha code.

Each column on the coding matrix represented an assessment item/activity. Reviewers, after reaching consensus on the depth-of-knowledge levels for all of the objectives, individually judged and wrote in the corresponding objective-item cell the depth-of-knowledge code for the assessment item/activity. Reviewers assigned the assessment item/activity one level of depth-of-knowledge and wrote this code in the item/activity's column in each row of an objective if a student's response to the item/activity provided information about what the student knew or could do with respect to the objective. Each objective coded for an item was called a hit. Multiple hits were allowed for any one assessment item/activity. Initially, reviewers were not given specifications about limits on the number of hits for any one assessment activity/item. After discussion with other reviewers following the coding of each test, reviewers developed more refined guidelines for multiple hits. This had the effect, as the reviewers gained more experience, of reducing noticeably the number of instances that reviewers marked multiple hits for any one item/activity. The number of multiple hits was one source of variation among reviewers. Reviewers did converge in the number of multiple hits as they became more familiar with the process and developed agreed-upon rules.

Reviewers were asked to code the assessment items/activities independently for each test, with little or no interaction. After all of the reviewers completed coding the instruments, they were asked to select a sample of items and compare their results. The primary purpose of this discussion was to improve the reliability among the reviewers in coding assessment items/activities on the next and subsequent instruments. Reviewers could make changes as they calibrated their work with the other reviewers if they felt it was appropriate. Reviewers discussed both what items/activities were assigned to what objectives and the depth-of-knowledge code assigned to each item.

---

[1] Objective as used in this analysis should not be confused with a behavioral objective designed to express one specific behavior and one depth-of-knowledge level.

Statistics for each standard were computed on four alignment criteria for content focus: categorical concurrence, depth-of-knowledge consistency, range-of-knowledge correspondence, and balance of representation. The mean number of hits was used to judge the categorical concurrence between the assessment instrument and the standards. The frequencies of hits aggregated across the objectives for each standard and by the depth-of-knowledge levels were used to judge the consistency by considering the percentage of hits that were below, at the same level as, or above the level for the objective. The percentage of the objectives hit within a standard was used to judge the range-of-knowledge correspondence within a standard. The distribution of the hits among the objectives for a standard with at least one hit was used to compute the balance of representation for a standard. This analysis is based on the assumption that the set of objectives for a standard spans the entire domain of knowledge and skills a student should demonstrate to fully meet the standard, an assumption not always met.

Reviewers were asked for their comments on other alignment criteria that included articulation across grades, pedagogical implications, and equity. Some offered their comments on these criteria, but because of strong time pressures, systematic procedures were not used to gather information on these criteria. Reviewers reacted to the overall process and made suggestions in a debriefing session held at the end of the institute. These comments are included in the summary report for the Alignment Analysis Institute and not in this document.

All of the statistics were computed for each reviewer. The mean for each statistic was computed using the results for only the reviewers who completed coding all of the items—i.e., at least two reviewers, and up to seven for some tests. The mean among reviewers on each statistic is a reasonable approximation for the summary information that lessens the error of any one reviewer in coding. Of course, statistics based on the coding by a greater number of reviewers will be more accurate. Standard deviations, reported along with the mean, provide one indication of the variation among reviewers. Low standard deviations indicate that there was minimal variation among reviewers on the marginal statistics for a standard. Relatively high standard deviations indicate that the reviewers disagreed more on the marginal statistics for a standard. The total number of objectives and the total number of hits for a standard also have to be considered in judging the agreement among reviewers.

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 if both address the same content categories. The categorical concurrence criterion provides a very general indication if both documents incorporate the same content. The criterion of categorical concurrence between standards and assessment is met if the same or consistent categories of content appear in both documents. This criterion 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 there to be an acceptable categorical concurrence between the standard and the assessment. The number of items, six, is based on estimating the number of items that could produce a reasonably reliable sub-scale for estimating students' mastery of content on that sub-scale. Of course, many factors have to be considered in determining what a reasonable number is, including the reliability of the sub-scale, the mean score, and cutoff score for determining mastery. Using a procedure developed by Subkoviak (1988) and assuming 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 sub-scales 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 cut-off score set at 4 would produce an agreement coefficient of .77. Any fewer items with a mean of one-half of the items and a standard deviation of one item would require a cut-off 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 sub-scale.

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 a standard had between 40% to 50% of its corresponding items at or above the depth-of-knowledge levels of the objectives, then it was reported that the criterion was "weakly" met.

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.

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.

Data Analysis

The codings for all of the reviewers were entered on a spreadsheet to compute summary statistics. For each assessment instrument and standards document, the codes for each reviewer were tabulated by the frequency of hits and the depth-of-knowledge levels for the hits. Data for all of the objectives for one standard were aggregated or listed as a profile for each standard. The results are reported for each standard.

The ratings upon which the initial data analyses were conducted consist of three fundamental pieces of information [variable names are given in brackets]: 1) the level of knowledge assigned to each objective as agreed upon by the panel of reviewers [LVL]; 2) an indication of the objective(s) relevant to each assessment item included in the state assessment (i.e., the base measure upon which categorical concurrence was measured) [OB]; and, 3) the level of knowledge deemed necessary for a student to successfully answer a given assessment item [ITMLVL].

The data were analyzed using the following procedures.

Measuring Level of Knowledge

Ratings were organized within each state for analysis by state, subject, grade level, and reviewer. For each assessment item, the level-of-knowledge rating [ITMLVL] was compared to the level-of-knowledge rating assigned by the panel [LVL] to the relevant objective [OB]. This comparison resulted in the indication of a hit in one of three categories for which a frequency count was established for each objective: [UN] (i.e., ITMLVL < LVL); [AT] (i.e., ITMLVL = LVL); and, [AB] (i.e., ITMLVL > LVL). In addition, the total number of hits for each objective were tabulated and recorded as [HITS]. By dividing [UN], [AT], and [AB] each by [HITS], we computed a percent under [PUN], a percent at [PAT], and a percent above [PAB] the level-of-knowledge [LVL] rating established by the review panel for each objective. By aggregating the categories of [UN], [AT], [AB], and [HITS] across objectives within a given standard, we also computed a summary [PUN], [PAT], and [PAB] measure for each standard.

Measuring Range

In order to establish a measure of range, the variable [#RNG] was used to indicate the total number of objectives included in a given standard. The number of objectives per standard for State B, for example, varied from as few as five to as many as twenty-two objectives per standard. The variable [RNG] was used to tabulate the actual number of objectives which received at least one hit for a given standard. Dividing [RNG] by [#RNG] yields a measure of the percent of objectives for a given standard that are represented in the state assessment [PRNG].

Measuring Balance

The most complex measure calculated concerns the notion of balance. Our calculation of balance involved three steps: First, a measure of the configuration of hits across objectives within a given standard was calculated by dividing the number of hits for a given objective [HITS] by the total number of hits for the standard [#HITS], yielding a percent of hits [PHITS] for each objective. A summary measure of [PHITS] was also calculated for each standard, using the total number of hits across all standards and objectives as the denominator, and the [#HITS] for each standard as the numerator.

The next step in calculating balance involved comparing the percent of [PHITS] (at the level of the objective) with a proxy measure representing perfect balance. We define a *balanced* standard as one in which each of the objectives for a standard are equally represented in state assessments; we used as the basis for this proxy measure the formula {1/[RNG]}. Note that we used [RNG] (number of objectives for which at least one hit was recorded) instead of [#RNG] (the total number of objectives listed for the standard regardless of whether the objective received a hit or not). Our reasoning for using [RNG] rather than [#RNG] as the denominator for this formula was to keep the resulting measure of balance conceptually distinct from the percent range measure [PRNG]. Had we used [#RNG] instead, the measure of balance would have been closely linked to [PRNG]. The interpretive effect of this is that the resulting measure of balance is based solely on those objectives actually measured by the state assessment.

The absolute value of the difference between [PHITS] and {1/[RNG]} was calculated and placed in the variable [ABSVAL]. The absolute value was used so that divergence from the target measure {1/[RNG} was not muted by the canceling effect of positive and negative values that would otherwise have resulted. The resulting measure [ABSVAL] was then aggregated to the standard level, producing a summary measure of the difference between the configuration of assessment items and the configuration of the state standards.

One effect of using absolute value in this way is that the theoretical range of the resulting summary measure is 0-2. In addition, the larger the measure, the *lower* the balance of the assessment relative to the state standard. In order to make the balance measure more understandable, the summary measure of difference (Σ [ABSVAL]) was divided by two and the result subtracted from one, in order to convert the balance measure to a scale with a range of 0-1, with 1 representing perfect balance [BAL].

Note that because we used [RNG] instead of [#RNG] as the denominator in calculating the balance proxy measure, it is possible to get a balance of 1 (i.e., perfect balance) without having assessed every objective included in the standard. The percent of objectives actually assessed was calculated as the variable [PRNG]. In order to provide a summary measure that was sensitive to both range and balance, we calculated one more summary measure, which is defined as the product of {[PRNG] * [BAL]} and which is reported in the tables of Appendix B as the variable [BALxPRNG].

Alignment of Standards and Assessment in Mathematics for Grade 8

Organization of Standards for Mathematics

In a document titled *Standard Course of Study and Grade Level Competencies for Mathematics K-12*, State B describes the expectations for student learning. This document was most recently revised in 1993. For each grade, K-8, expectations are stated in seven competency goals. Under each goal, more specific objectives are provided. The seven competency goals address the same mathematics content areas across the nine grades: numeration, geometry, patterns/relationships, measurement, problem solving/reasoning, probability/statistics, and computation. For the purpose of this report, each competency goal is considered a standard. The expectations listed under the competency goals are considered both in this study and by State B as objectives. An example of a competency goal and some objectives from grade 8 follow:

Competency Goal 2: The learner will demonstrate an understanding and use properties and relationships of geometry.

- 2.1 Use the Pythagorean Theorem to find the missing side of a right triangle; use calculator, when appropriate.
- 2.2 Solve problems related to similar figures, using indirect measures to determine missing sides.
- 2.3 Draw 3-dimensional figures from different perspectives (top, side, front).
- 2.4 Graph on a coordinate plane similar figures, reflections, and translations.
- 2.5 Explore the triangle congruency relationships: ASA, SSS, SAS.
- 2.6 Explore the relationships of the angles formed by cutting parallel lines by a transversal.
- 2.7 Solve problems that relate geometric concepts to real world situations.

Grades 9-12 competency goals (standards) are organized by course—Pre-Algebra, Technical Mathematics, Algebra I – A, Algebra I – B, Algebra I, Geometry, Algebra II, Advanced Mathematics, and Advanced Placement Calculus. The competency goals for courses were consolidated into four goals to represent what students in State B should know by the end of grade 10—1. Number Sense, Numeration, and Numerical Operations; 2. Spatial Sense, Measurement, and Geometry; 3. Patterns, Relationships, and Functions; and 4. Statistics, Probability, and Discrete Mathematics. These goals and the objectives incorporated end-of-course goals and objectives from Pre-Algebra, Algebra, Geometry, and Technical Mathematics. For example, Goal 4 states:

Goal 4 This strand will focus on using statistical methods, analyses, and relationships to collect, organize, and describe data and communicate the results; determining the probability of simple and compound events; and solving related problems.

- 4.1 The learner will use statistics to analyze and solve real-world problems.
- 4.2 The learner will use probability to solve real-world problems.
- 4.3 The learner will fit a line or curve to a set of data and use this line or curve to make predictions about data.

Assessments for Mathematics

End-of-grade tests are administered in the elementary and middle grades. In this study, end-of-grade test for grade 8 was analyzed. The mathematics end-of-grade tests consisted of three parts, incorporated in booklets along with an assessment on reading comprehension. At the beginning of each booklet, students are asked to work 12 multiple-choice items in computation. After a section on reading comprehension, students are asked to complete a section labeled Mathematics Applications. This section includes 72 multiple-choice items. In a separate booklet, students are asked to respond to six open-ended questions in mathematics, as well as reading items, related to one context described in a page of text.

Alignment of Standards and Assessment for Grade 8 Mathematics

Categorical Concurrence

The criterion of categorical concurrence between standards and assessment is met if the same or consistent categories of content appear in both documents. Three reviewers coded the depth-of-knowledge levels of 43 objectives for the seven standards (Competency Goals) and the 86 assessment activities/items (Table BM8-1). They used few multiple hits for the assessment items, indicating that most of the assessment activities/items were judged to measure specific skills, concepts, and thinking confined to one objective. The average number of hits by reviewer was 95, with a standard deviation of 7.21. Reviewers varied in their assignment of assessment activities/items to more than one objective, but even the reviewer who coded the most multiple hits only did so on 17 assessment activities/items, or less than 20% of the items.

All of the seven standards were found to have a sufficient number of corresponding items to make some judgment about a student's knowledge and skills related to each standard. If the 86 activities/items were evenly distributed among the seven standards, the expected number of items per standard would be 12. Only Standard 3 (Patterns, Relationship, and Pre-Algebra) exceeded this expected value by a large number of items, giving this standard greater weighting on the assessment than the other standards.

Reviewers were very consistent in the number of assessment activities/items they coded as related to objectives of three of the seven standards (1, 2, and 4). The standard deviation of .58 for the number of hits by the three reviewers on these standards was the minimum short of perfect agreement. For the other four standards, reviewers varied more on the number of items assigned to specific standards. However, even with this variation, none of the three reviewers assigned fewer than six activities/items corresponding to standards, the minimum for indicating categorical concurrence.

Depth-of-Knowledge Consistency

Depth-of-knowledge consistency between standards and assessment are aligned 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. Depth-of-knowledge consistency was

attained by five of the seven standards (1, 2, 3, 4, and 7) (Table BM8-2). For these standards, the percentage of corresponding assessment activities/items with a level of knowledge that was the same as or above that for the objective was 50% or higher. Over 60% of the activities/items corresponding to Standard 5 (Problem Solving and Reasoning) were coded as below the level of knowledge of the corresponding objective. On this standard, a student could achieve a score of over 60%, a minimal passing level assumed for this analysis, without answering any items at the level of the objectives for Standard 5. The depth-of-knowledge consistency was weak for Standard 6 (Probability and Statistics); 56% of the items had a level of knowledge under the corresponding objectives.

The standard deviations on the level of knowledge codes ranged from 0 to 21. Two of the three reviewers were generally consistent. The third reviewer tended to code the activities/items at the same levels as the objective more often than did the other two reviewers. Reviewers also disagreed noticeably on the few times any one coded items above the objectives' level of knowledge, with standard deviations on four of the standards exceeding 50% of the mean. Although reviewers showed some disagreement in coding depth-of-knowledge consistency, they agreed enough to make a valid judgment on this criterion.

Range-of-Knowledge Correspondence

The range-of-knowledge criterion is used to judge whether a span of knowledge expected of students for meeting a standard is the same as, or corresponds to, the span of knowledge students need in order to correctly answer the assessment items/activities. The correspondence between span of knowledge for a standard and the assessment is the number of objectives within the standard that has at least one related assessment item/activity. Only one standard, Standard 4 (Measurement) (Table BM8-3), failed to meet an acceptable level--50% or more of the objectives with a related assessment activity/item--on the range-of-knowledge correspondence. On average, the three reviewers only coded activities/items related to two of the five objectives, or 40%. This percentage is too low for this standard to be adequately represented on the assessment. The range for the other six standards was high, exceeding 80% of the objectives. Reviewers also were in strong agreement in coding this variable. For five of the seven standards, the standard deviation was nearly 10%, or less, of the mean. The variation in coding ranges on the other two standards still were reasonably low, with standard deviations about 20% of the mean. On both of these standards, one of the reviewers coded at least one assessment activity/item to the standard, but not to any one of the existing objectives. This indicates some lack of coverage of the full span of content on these two standards, as measured by the activities/items on the assessment. Overall, the range-of-knowledge correspondence between the standards and assessment was very good.

Balance of Representation

The range-of-knowledge criterion only considered the proportion of objectives within a standard hit, but did not take into consideration how the hits (or assessment items/activities) were distributed among these objectives. Most of the items corresponded to only one of the objectives, with the other items distributed evenly among the other objectives. The balance-of-representation criterion is used to indicate the degree to which one objective within a standard is given more emphasis on the assessment than another. An index is used to judge the distribution of

assessment items. This index only considered the objectives for a standard that had one hit, e.g., one related assessment item/objective.

For five of the standards, the items among the objectives hit were evenly distributed without one objective having a significantly larger number than the others (Table BM3-3). For these standards, the balance of representation was acceptable. For the other two standards, Standards 4 and 7, one or two objectives were overloaded with related assessment activities/items, compared to the other objectives with at least one corresponding item. For example, most of the assessment items (11 to 16) for Standard 7 (Computation) corresponded to objective 7.1 (Select appropriate operations, strategies, and methods of solving a variety of application problems using real numbers, justifying the selection.). Very few items (0 to 2) were coded as corresponding to objective 7.2 (In meaningful contexts, develop the laws of exponents; solve problems involving exponentiation.). Both of the standards with a low balance of representation had a low number of objectives, either written or hit. This was a related factor, but not the only factor that contributed to low balance. To achieve full balance on the assessment, more activities/items need to be included that relate to a greater number of the objectives for these two standards.

Summary for Grade 8 Mathematics Alignment

Overall, the grade 8 Competency Goals and the grade 8 assessments had a high degree of alignment (Table BM8-4). A sufficient number of assessment items corresponded to each standard (categorical concurrence). Items on the assessment were at, or above, the depth-of-knowledge levels in sufficient proportion for five of the seven standards (depth-of-knowledge consistency). The items on the assessment corresponded to a high proportion of the objectives in six of the seven standards (range of knowledge). The items were adequately distributed among the objectives addressed for five of the seven standards (balance of representation). The alignment could be improved by including assessment items that correspond to more of the objectives under the measurement standard, including assessment items at a higher cognitive level for Standard 5 (Problem Solving and Reasoning) and Standard 6 (Probability and Statistics), and including more items on exponents related to Standard 7 (Computation).

Study Limitations

This analysis did not take into consideration the quality of the standards or of the assessments. It only considered the relationship between the two documents for each grade level and the degree to which the assessment and standards were aligned with each other. Specified cut-off points were set to judge acceptable agreement on each criterion. The assumptions for these cut-off points may not be valid for State B. It could be that the cut-off points for what is acceptable should be higher or lower. Varying the cut-off points would change the results of this analysis. Also, not all of the possible relevant information was incorporated, such as the scoring rubrics, or objectives designed to be assessed in ways other than by the assessment instruments included in this analysis. All of these factors need to be considered in interpreting the results reported here. Based on this analysis, the underlying assumptions, and the conditions under which the analysis was performed, the State B mathematics standards and assessments for grade 8 attained reasonably high alignment.

References

Baker, E.L., Freeman, M., & Clayton, S. (1991). Cognitive assessment of history for large-scale testing. In M.C. Wittrock & E. L. Baker (Eds.), *Testing and Cognition* (pp. 131-153) Englewood Cliffs, NH: Prentice Hall.

Carroll, J.B. (1963). A model of 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.

Consortium 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.

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

Smith, M.S., & O'Day, J. (1991). Systemic school reform. In S.H. Fuhrman & B. Malen (Eds.), *The politics of curriculum and testing.* Politics of Education Association Yearbook. (1990, pp. 233-267). London: Taylor & Francis.

Spillane, J.P. (1998). State policy and the non-monolithic nature of the local school district: Organizational and professional considerations. *American Educational Research Journal, 35* (1), 35-63.

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). *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.* Council of Chief State School Officers and National Institute for Science Education Research Monograph No. 6. Madison, WI: University of Wisconsin-Madison.

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

List of Tables

State B

Grade 8 Mathematics Alignment Analysis Tables

Table BM8-1
Categorical Concurrence Between Standards and Assessment As Rated by Three Reviewers
State B--Grade 8 Mathematics
(Number of Assessment Items—8 Computation, 72 Applications, 6 Open-Ended, Total 86)

Standards			Level by Objective			Hits		Categorical Concurr. Acceptable
Title	Goals #	Objs #	Level	# of objs by Level	% w/in std by Level	Mean	S.D.	
1. Numeration (Real #s)	1	8	2 3	6 2	75 25	16.33	.58	Yes
2. Geometry	1	7	2 3 4	3 3 1	43 43 14	9.67	.58	Yes
3. Patterns, Relationships, & Pre-Algebra	1	7	2 3 4	5 1 1	72 14 14	20.00	3.46	Yes
4. Measurement	1	5	2 3 4	2 1 2	40 20 40	9.67	.58	Yes
5. Problem Solving and Reasoning	1	6	3 4	4 2	67 33	13.67	5.03	Yes
6. Graphing, Probability, & Statistics	1	7	2 3 4	1 5 1	14 72 14	9.33	3.06	Yes
7. Computation (Real #s)	1	3	2 3	1 2	33 67	16.33	3.51	Yes
Total	7	43	2 3 4	18 18 7	42 42 16	95.00	6.03	

Table BM8-2
Depth-of-Knowledge Consistency Between Standards and Assessment
As Rated by Three Reviewers
State B--Grade 8 Mathematics
(Number of Assessment Items—8 Computation, 72 Applications, 6 Open-Ended, Total 86)

Standards			Level by Objective			Hits		Level of Item w.r.t. Standard						Depth-of-Knowledge
								% 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.	
1. Numeration (Real #s)	1	8	2 3	6 2	75 75	16.33	.58	46	12	54	12	0	0	Yes
2. Geometry	1	7	2 3 4	3 3 1	43 43 14	9.67	.58	50	0	44	5	6	5	Yes
3. Patterns, Relationships, & Pre-Algebra	1	7	2 3 4	5 1 1	72 14 14	20.00	3.46	37	3	51	7	11	4	Yes
4. Measurement	1	5	2 3 4	2 1 2	40 20 40	9.67	.58	31	5	69	5	0	0	Yes
5. Problem Solving & Reasoning	1	6	3 4	4 2	67 33	13.67	5.03	64	21	30	18	5	3	No
6. Probability & Statistics	1	7	2 3 4	1 5 1	14 72 14	9.33	3.06	56	18	36	18	7	3	Weak
7. Computation (Real #s)	1	3	2 3	1 2	33 67	16.33	3.51	47	13	52	10	2	3	Yes
Total	7	43	2 3 4	18 18 7	42 42 16	95.00	6.03	48	15	48	16	5	5	

Table BM8-3
Range-of-Knowledge Correspondence and Balance of Representation Between Standards and Assessment As Rated by Two Reviewers
State B--Grade 8 Mathematics
(Number of Assessment Items—8 Computation, 72 Applications, 6 Open-Ended, Total 86)

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/ std by Level	Mean	S.D.	Mean	S.D.	Mean	S.D.		Mean	S.D.	Mean	S.D.	
1.Numeration (Real #s)	1	8	2 3	6 2	75 75	16.33	.58	5.67	.58	82	16	Yes	17	2	.75	.06	Yes
2. Geometry	1	7	2 3 4	3 3 1	43 43 14	9.67	.58	6.00	0.00	86	0	Yes	10	1	.83	.04	Yes
3. Patterns, Relationships, & Pre-Algebra	1	7	2 3 4	5 1 1	72 14 14	20.00	3.46	5.67	0.58	81	8	Yes	21	4	.74	.02	Yes
4.Measurement	1	5	2 3 4	2 1 2	40 20 40	9.67	.58	2.00	0.00	40	0	No	10	1	.60	.01	No
5. Prob Solving & Reasoning	1	6	3 4	4 2	67 33	13.67	5.03	4.67	1.53	94	10	Yes	14	5	.72	.08	Yes
6. Probability & Statistics	1	7	2 3 4	1 5 1	14 72 14	9.33	3.06	5.33	2.08	95	8	Yes	10	3	.80	.05	Yes
7. Computation (Real #s)	1	3	2 3	1 2	33 67	16.33	3.51	2.67	0.58	89	19	Yes	17	3	.55	.10	No
Total	7	43	2 3 4	18 18 7	42 42 16	95.00	6.03	4.57	1.75	81	20		14	5	.71	.10	

Table BM8-4
Summary of Attainment of Acceptable Alignment Level on Four Content Focus Criteria
State B--Grade 8 Mathematics
(Number of Assessment Items—8 Computation, 72 Applications, 6 Open-Ended, Total 86)

Standards	Alignment Criteria			
	Categorical Concurrence	Depth-of-Knowledge Consistency	Range of Knowledge	Balance of Representation
1. Numeration (Real #s)	Yes	Yes	Yes	Yes
2. Geometry	Yes	Yes	Yes	Yes
3. Patterns, Relationships, & Pre-Algebra	Yes	Yes	Yes	Yes
4.Measurement	Yes	Yes	No	No
5. Problem Solving & Reasoning	Yes	No	Yes	Yes
6. Probability & Statistics	Yes	Weak	Yes	Yes
7. Computation (Real #s)	Yes	Yes	Yes	No

Appendix A

Criteria for Alignment of Expectations and Assessments in Mathematics and Science Education

Appendix A

Criteria for Alignment of Expectations and Assessments in Mathematics and Science Education

1 – Content Focus. System components should focus consistently on developing students' knowledge of subject matter. Consistency will be present to the extent components' logic of action and the ends achieved share the following attributes:

A. *Categorical Concurrence.* Agreement in content topics addressed.
B. *Depth-of-Knowledge Consistency.* Agreement in level of cognitive complexity of information required.
C. *Range-of-Knowledge Correspondence*. Agreement in the span of topics.
D. *Structure-of-Knowledge Comparability*. Agreement in what it means to know concepts.
E. *Balance of Representation*. Agreement in emphasis given to different content topics.
F. *Dispositional Consonance*. Agreement in attention to students' attitudes and beliefs.

2 – Articulation Across Grades and Ages. Students' knowledge of subject matter grows over time. All system components must be rooted in a common view of how students develop, and how best to help them learn at different developmental stages. This common view is based on:

A. *Cognitive Soundness Determined by Superior Research and Understanding*. All components build on principles for sound learning programs.
B. *Cumulative Growth in Knowledge During Students' Schooling*. All components are based on a common rationale regarding progress in student learning.

3 – Equity and Fairness. When expectations are that all students can meet high standards, aligned instruction, assessments, and resources must give every student a reasonable opportunity to demonstrate attainment of what is expected. System components that are aligned will serve the full diversity in the education system through demanding equally high learning standards for all students while fairly providing means for students to achieve and demonstrate the expected level of learning. To be equitable and fair, time is required for patterns to form in order to decipher how system components are working in concert with each other. Judging a system on the criterion of equity and fairness will require analysis over a period of time.

4 – Pedagogical Implications. Classroom practice greatly influences what students learn. Other system components, including expectations and assessments, can and should have a strong impact on these practices, and should send clear and consistent messages to teachers about appropriate pedagogy. Critical elements to be considered in judging alignment related to pedagogy include:

A. *Engagement of Students and Effective Classroom Practices*. Agreement among components in a range of learning activities and in what they are to attain.
B. *Use of Technology, Materials, and Tools.* Agreement among components in how and to what ends applications of technology, materials, and tools are to be included.

5 – System Applicability. Although system components should seek to encourage high expectations for student performance, they also need to form the basis for a program that is realistic and manageable in the real world. The policy elements must be in a form that can be used by teachers and administrators in a day-to-day setting. Also, the public must feel that these elements are credible, and that they are aimed at getting students to learn the mathematics and science that are important and useful in society.

Appendix B

Coding Variables Used in Alignment Data Summaries

Appendix B

Coding Variables Used in Alignment Data Summaries

ST Standard #

GO Goal

OB Objective

LVL Cognitive Skill Level

PUN Percent Under Cognitive Level

PAT Percent At Cognitive Level

PAB Percent Above Cognitive Level

 # RNG Number of Range (Number of Goals/Objectives in Standard)

RNG Range (Number of Goals/Objectives Represented on Assessment)

PRNG Percent Range = RNG/#RNG

 # HITS Number of Hits (Total # of Codings* for Assessment Items)

PHITS Percent of Hits = Hits / # Hits

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

Where O = Total number of objectives hit for the standard
$I_{(k)}$ = Number of items hit corresponding to objective (k)
H = Total number of items hit for the standard

BAL X PRNG Product of Balance Index and Percent Range

* A single assessment item may be coded in more than one standard, goal, or objective.