Research Monograph No. 6

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

Norman L. Webb

National Institute for Science Education (NISE) Publications

The NISE issues papers to facilitate the exchange of ideas among the research and development community in science, mathematics, engineering, and technology (SMET) education and leading reformers of SMET education as found in schools, universities, and professional organizations across the country. The NISE Occasional Papers provide comment and analysis on current issues in SMET education including SMET innovations and practices. The papers in the NISE Research Monograph series report findings of original research. The NISE Conference and Workshop Reports result from conferences, forums, and workshops sponsored by the NISE. In addition to these three publication series, the NISE publishes Briefs on a variety of SMET issues.

The research reported in this paper was supported by a cooperative agreement between the National Science Foundation and the University of Wisconsin–Madison (Cooperative Agreement No. RED-9452971). At UW–Madison, the National Institute for Science Education is housed in the Wisconsin Center for Education Research and is a collaborative effort of the College of Agricultural and Life Sciences, the School of Education, the College of Engineering, and the College of Letters and Science. The collaborative effort is also joined by the National Center for

Improving Science Education, Washington, DC. Any opinions, findings, or conclusions are those of the author and do not necessarily reflect the view of the supporting agencies.

Research Monograph No. 6

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

Norman L. Webb

National Institute for Science Education
University of Wisconsin-Madison

Council of Chief State School Officers
Washington, DC

April 1997

About the Author

Norman Lott Webb is a senior research scientist for the Wisconsin Center for Education Research at the University of Wisconsin-Madison. He directs evaluations of curriculum and professional development projects. Currently he is leading the Strategies for Evaluating Systemic Reform Project for the National Institute for Science Education, a cooperative agreement between the National Science Foundation and the Wisconsin Center for Education Research. He is directing the evaluation of the Interactive Mathematics Program and the Park City/IAS Mathematics Institutes and is serving as a consultant on the National Evaluation of Library Power project. Some of the recent projects he has directed include the *Wisconsin Performance Assessment Development Project*, funded by the Wisconsin Department of Public Instruction, and the mathematics case studies for the Case Studies of United States Innovations in Mathematics and Science and Technology Education in an International Context Project. His work with NCTM has included consulting with the Assessment Standards writing group. He chaired the evaluation working group, one of four groups who wrote the Curriculum and Evaluation Standards for School Mathematics. He edited the NCTM 1993 yearbook on classroom assessment. A book he edited with Tom Romberg on the Urban Mathematics Collaborative Project was published in 1994. He has worked on or directed evaluations on a number of projects including the Woodrow Wilson National Fellowship Foundation Leadership One-week Summer Institutes. He has had the privilege to present lectures and seminars to a number of groups in the United States and in other places in the world including Oman, Iceland and Spain. He is a member of different groups working toward change in assessment and chairs the advisory board for the *Assessment in Practice* newsletter produced by Mathematical Sciences Education Board. His current views on assessment are best reflected in "Assessment of Students' Knowledge of Mathematics: Steps Toward a Theory," a chapter in the *Handbook of Research on Mathematics Teaching and Learning*.

Contents

Contents (Continued)

Abstract

Alignment is central to current efforts of systemic and standards-based education reforms in mathematics and science. More than four-fifths of the states have content frameworks in place in mathematics or science. A large number of these have some form of a statewide assessment to measure student attainment of expectations given in the frameworks. These reforms are based, in part, on the premise that student outcomes will be improved through creating coherent systems of expectations and assessments. Expectations are major elements of educational policy on what students should know about mathematics and science and what they should be able to do with that knowledge. Assessments are major elements of educational policy used to measure student achievement and are classroom tools used by teachers. All assessments used within a state or district constitute an assessment system. The purpose of this monograph is to define criteria for judging the alignment between expectations and assessments.

States and districts were found to use three general approaches for judging the alignment among expectations and assessments. A number of states and districts develop documents in sequence, such as first standards, then curriculum frameworks, and then assessments. These documents are aligned because they were developed to be, with the previously developed document forming the blueprint for the next document. Other systems hire experts to review expectations and assessments. A third approach, used by the Third International Mathematics and Science Study, is to systematically analyze both expectations and assessments using a common metric. Specific criteria for judging the alignment, however, were missing from all three approaches.

Twelve criteria for judging alignment grouped into five general categories are described along with examples and levels of agreement. The five general categories are content focus, articulation across grades and ages, equity and fairness, pedagogical implications, and system applicability. The criteria were created to provide guidance to state or district officials on the important aspects of expectations and assessments that need to be considered in some detail to have a coherent system where the power of these policy documents converges to better support students' learning of important mathematics and science.

The criteria were developed with the input of an expert panel formed as a cooperative effort between the Council of Chief State School Officers and the National Institute for Science Education. The monograph was supported by the Council under a research and evaluation grant from the National Science Foundation on State Curriculum Frameworks and Standards in Mathematics and Science Education. A shorter version of this paper was published as an NISE *Brief*, January 1997.

Introduction

Many states and school districts are making concerted efforts to boost student achievement in mathematics and science. These are not efforts aimed at simple face lifts, but attempts to develop deep, lasting changes in how students learn these critical subjects. This monograph is directed towards those in states and districts who are working to improve student learning through creating coherent systems of expectations and assessments. Other audiences are those who study reform, make decisions about reform, and are affected by reform. The intention of this monograph is to help people think more clearly about the concept of alignment and to examine what is required for important system elements of expectations and assessments to converge.

Educators, notably through efforts spearheaded by national professional associations, increasingly recognize the need for major reform in K-12 mathematics and science curricula and are embracing a vision of ambitious content for all students. Making this vision a reality means encouraging “a far deeper and dynamic level of instructional decision making” (Baker, Freeman, & Clayton, 1991). This is not something that can be done simply by mandating new accountability measures. At the heart of these efforts to make deep changes in instruction is the concept of "alignment."

The major elements of an education system must work together to guide the process of helping students achieve higher levels of mathematical and scientific understanding. Educators increasingly recognize that if policy elements are not aligned, the system will be fragmented, will send mixed messages, and will be less effective (CPRE, 1991; Newmann, 1993). For example, the Systemic Initiatives program of the National Science Foundation is directed toward states, districts, and regions setting ambitious goals for student learning, through a coherent policy system established, in part, on assessments aligned with those goals. The Improving America's Schools Act explicated how assessments are to relate to standards: ". . . such assessments (high quality, yearly students 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. As more and more weight is given to accountability throughout education systems, alignment between assessments and expectations becomes not only critical, but also essential.

Importance of Alignment

Assuring the alignment between expectations and assessments can strengthen an education system in important ways. Teachers give more credence to documents they understand are in agreement, are useful, and will serve to benefit their students. Teachers, already overloaded with responsibilities, are better able to attend to expectations and assessments if they provide a consistent message and have credibility. District curriculum guides, large-scale assessments,

standards, state frameworks, and other policy documents overwhelm teachers. Often teachers' only strategy for coping with this barrage of directions is to ignore them (Cohen, 1990). Carefully aligned assessments and expectations, with input from teachers and others, add to the value teachers give to these documents and their willingness to make sense of these documents. Teachers are more apt to attend to student outcomes listed in expectations if they know these outcomes will be assessed by instruments that will give them feedback on how well their students did on these outcomes. Through understanding the link between expectations and assessment, teachers are more likely to find ways to translate what is being advanced by these documents into their daily work with students.

Aligning assessments with expectations can improve the efficiency and effectiveness of the education system. An aligned system can more effectively set priorities and allocate limited resources. Both expectations and assessments are important statements of what the system believes students should know and do. Better aligned goals and measures of attainment of these goals will increase the likelihood that multiple components of any district or state education system are working towards the same ends. Most students in their K-12 school career will have a number of science and mathematics teachers, maybe as many as 18 or more. Aligning goals carefully with the assessment system—all forms of gathering information about students' learning—is an important tool for mapping students' learning progress as they proceed through the system, allocating curriculum responsibly to teachers, and verifying when students' knowledge of important ideas is assessed. Not all learning can be assessed by large-scale assessments. Teachers are in a much better position to assess important learning such as how well a student is able to perform a scientific inquiry or devise a mathematical proof. Aligning the assessment system with expectations serves as an inventory to help assure all outcomes are being assessed in some way. That is, the expectations are covered by the assessments. A careful analysis of alignment between expectations and assessments also will help reduce unnecessary repetition in the assessment system caused by overassessing a few outcomes at the expense of ignoring others. Other important system functions, such as professional development, soliciting public support, and textbook selection, can be more effectively planned if the goals and the measurement of those goals are in agreement. A formal process for assuring agreement of policy elements and the verification of their agreement both are means for dealing with the complexity within the system and marshalling the support of others.

A formal process for assuring assessments and expectations are in alignment provides a response to those who challenge the system. Both state education departments and school districts face public scrutiny and review. In some instances, legal suits have been waged against education systems for not adequately educating students. A system will be in a better position to respond to any challenges if it has a well-defined and thoughtful process of assuring that it is accountable. Part of this assurance comes from confirming that assessments and expectations are aligned. That is, students are being assessed on what they are expected to know. Having a system where assessments are aligned with expectations, however, will not prevent challenges to the system, nor will it eliminate the need to validate the quality of the expectations. A formal alignment process employed by a district or state is one indication that these systems are assuming responsibility for assuring that students are learning what is expressed as important knowledge in standards, frameworks, or other statements of expectations.

Alignment in Principle

Two or more system components are aligned if they are in agreement or match each other. In the past, the most common educational use of the concept of alignment referred to the match between an assessment instrument (or instruments) and a curriculum. Here, alignment is analogous to instructional or curricular validity of a test (Harmon, 1991). A legal ruling in the 1981 Florida case Debra P. vs. Turlington emphasized the importance of assuring agreement between a curriculum and tests. According to this ruling, for a test to be fair, both curriculum and instruction must match the content coverage of the test (Madaus, 1983). Legally, high stakes tests need to be fair by being "aligned" with curriculum and instruction.

Alignment has been used as a comparison between the psychometric qualities of an assessment and instructional uses of results. Baker, Freeman, and Clayton (1991) questioned the viability of using standardized norm-referenced tests to assess either the improvement of an individual's education or the impact of systemic education reform. Because traditional standardized test results provide only a ranking, not a level of performance, they obscure the meaning of test scores and are less aligned with instruction.

The form of an assessment can be as important as the content in judging alignment. "The content and form of an assessment task must be congruent with what is supposed to be measured" (NRC, 1996, p. 83). For example, an assessment using a short-answer format is not aligned with an intended purpose of measuring students' ability to frame questions for conducting scientific inquiry and to design an inquiry to address the questions.

Alignment does not only refer to a comparison between one assessment instrument with a curriculum, but extends to a set of assessment instruments or the assessment system. "The term *alignment* is often used to characterize the congruence that must exist between an assessment and the curriculum. Alignment should be looked at over time and across instruments" (MSEB, 1993, p. 123). A single assessment may not be well aligned with curriculum because it is too narrowly focused, but it may be part of a more comprehensive collection of assessments that is in full alignment with the curriculum.

With the advent of systemic reform and the increased prominence of standards, alignment is increasingly being used to characterize the agreement or match among a set of documents or multiple components of a state or district system. A model used for the evaluation of statewide systemic initiatives assumed that student attainment of high standards of learning requires significant improvements in classroom teaching, the use of more challenging curricula and materials, and the regular assessment of student learning integrated and aligned with instruction (Zucker, Shields, Adelman, & Powell, 1995, p. 1). Building on this more recent view of the match of important components within a system, alignment can be defined.

> *Alignment* is 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.

Alignment is intimately related to test "validity." However, important and useful distinctions can be drawn between the two concepts. Validity refers to the appropriateness of inferences made from information produced by an assessment (Cronbach, 1971). Alignment refers to how well all policy elements in a system work together to guide instruction and, ultimately, student learning. Of the many different types of validity (Messick, 1989, 1994; Moss, 1992), alignment corresponds most closely with content validity and consequential validity. For example, the degree to which a test is aligned with a curriculum framework may affect a test's validity for a single purpose, such as making decisions on the curriculum's effectiveness. But a test, or tests, and a curriculum framework that are in alignment will work together to communicate a common understanding of what students are to learn, to provide consistent implications for instruction, and to be fair for all students. In addition, tests aligned with a curriculum will be based on compatible and sound principles of cognitive development.

Alignment of Expectations and Assessments

This monograph is primarily confined to the discussion of alignment of expectations and assessments, two major elements of educational policy

- **Expectations** of what students should know about mathematics and science and what they should be able to do with that knowledge. Expectations can be communicated in different ways. Educators can, for example, craft sets of standards or frameworks, ranging from broad vision statements to precise indications of expected performance and recommended instructional practices.

- **Assessments** used to gauge student achievement in science and mathematics and to indicate whether the expectations are being achieved. They can be used to formulate policy, monitor policy effects, enforce compliance with policies, demonstrate accountability, make comparisons, monitor progress toward goals, and/or make judgments about the effectiveness of particular programs. Assessments can refer to the collection or system of procedures used by teachers in classrooms and to district or state tests. Assessments include paper-and-pencil tests, but also encompass other forms of gathering information about students such as interviews and observations. The nature of assessments may be best represented by their development specifications along with instruments rather than by instruments alone.

Both expectations and assessments are now of great concern among educators and policymakers as key to standards-based education, systemic reform, and accountability (Chubin, in press; Ferrini-Mundy & Johnson, 1996; National Academy of Sciences, 1997). Because of the centrality of these two policy elements to current thinking on reform, this monograph is

restricted to specifying ways of judging the agreement between only these two elements. There are, of course, many other important elements in any education system, such as professional development, instructional materials, college requirements, teacher certification, resource allocations, and state mandates. These elements cluster into four different strata: purpose, policy, programs, and practice (Bybee, 1995). How all of these serve to provide a coherent system is important.

The number of different components within a system suggest there are several ways to think about agreement among components. The type of alignment discussed in this monograph will be referred to as "horizontal alignment," meaning the degree to which standards, frameworks, and assessments work together within an education system and mainly at the policy level. This is different from "vertical alignment," which is the degree to which the elements among the strata in an education system (e.g. textbook content, classroom instruction, professional development, and student outcomes) are aligned with each other and with outside forces (e.g. national standards, public opinion, and work force needs) (Figure 1). Horizontal alignment and vertical alignment will be used when necessary to distinguish between these two different directions of agreement.

Alignment in Practice

Judging alignment between expectations and assessments is difficult for several reasons. For one, both expectations and assessments frequently are expressed in multiple pieces or documents, making it difficult to assemble a complete picture. Also, it is difficult to establish a common language for describing different policy elements: The same term may have very different meanings when used to define a goal than when used to describe something that can be measured by assessment. For example, a viable goal in mathematics is for students to be able to use multiple of strategies to solve problems. The intent of this goal is for students to have an assortment of strategies that they can draw upon to solve a problem. Sometimes this goal is measured by giving students a problem to solve and directing them to solve this problem in more than one way or explain their answer in more than one way. Solving a problem in more than one way does not fully assess whether students have a range of strategies for solving a variety of problems. Further, the policy environment in an education system can be constantly changing. New goals can be mandated, for example, while old forms of assessment are still in place. Ever-expanding content areas, advancing technology, and a growing body of research on learning also contribute to the complexity of identifying expectations and assessments.

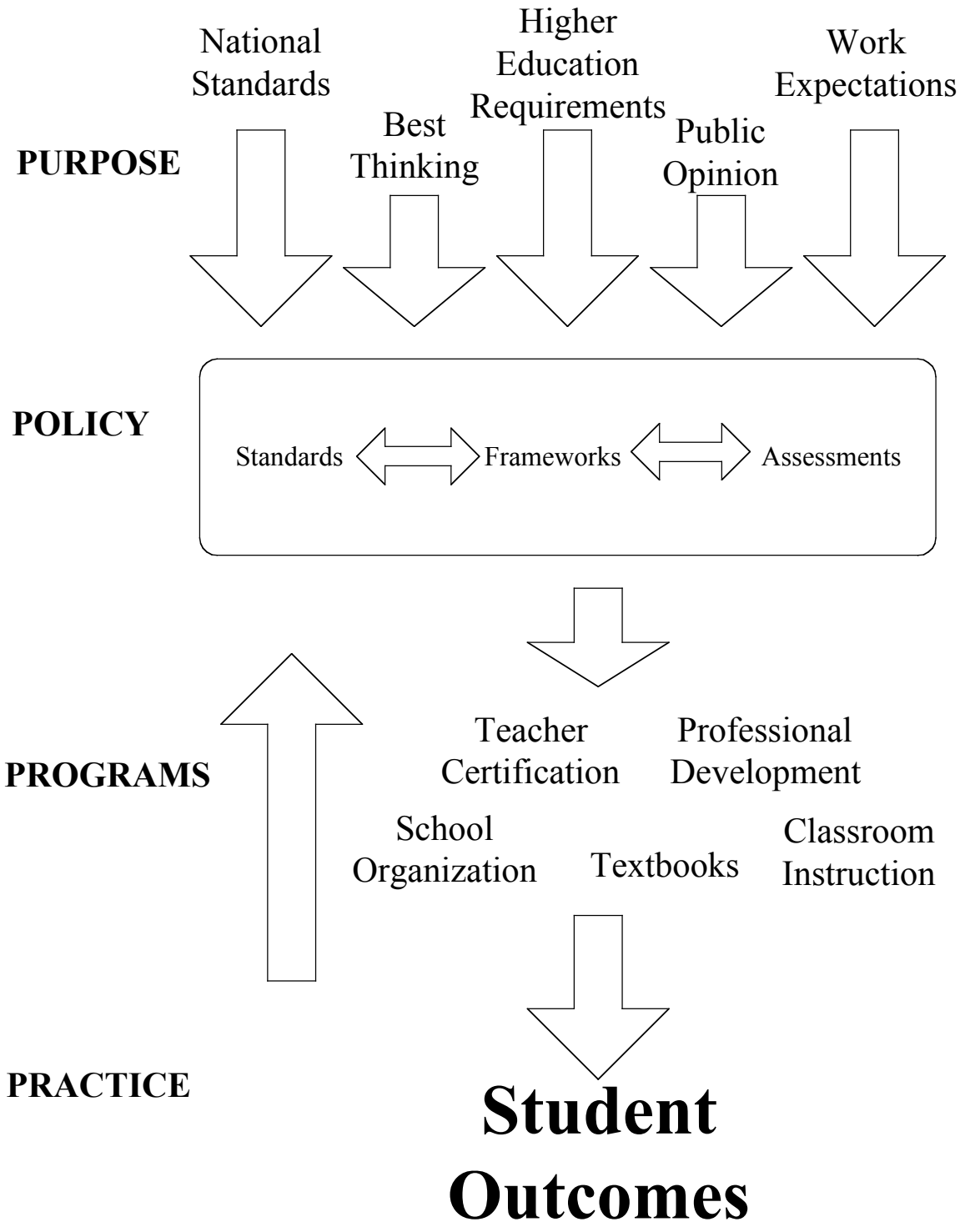

Figure 1. Vertical and horizontal alignment within an educational system

Increasingly more states and districts are expressing expectations for student learning in frameworks or standards. At the end of 1994, 42 of the 50 states, the District of Columbia, and Puerto Rico had mathematics frameworks or curriculum-related documents (Blank & Pechman, 1995). About the same number had science frameworks. Ten of these states had combined mathematics-science frameworks. Nearly all of the other states were in the process of developing frameworks. Even though most of the frameworks were less than 10 years old, some states, such as California, have had frameworks for a number of years. The first version of California's *Mathematics Framework* was published in 1963 (Webb, 1993). Since then a revised framework had been produced by California about every 5-7 years as the first step in adopting each new round of K-8 textbooks. By spring 1996 40 states had content standards ready in mathematics and 38 had standards ready in science (CCSSO, 1996).

What constitutes a framework varies among the states. Some frameworks consist only of content standards while others go into more detail in specifying learning objectives and recommended instructional practices. Arkansas prepared a mathematics framework based on 13 content standards applicable to all grades; South Carolina defined mathematics content standards by grade ranges for each of 6 core strands; New Jersey specified 18 mathematics standards, including content, process, and learning environment standards for K-12. Science frameworks were as varied: Arizona defined 8 goals (K-12) that called for teaching higher-order thinking skills and use of conceptual thinking; New Hampshire identified 6 curriculum strands, with curriculum standards for grades K-12 and proficiency standards for grades 6 and 10; Virginia outlined standards of learning objectives as well as descriptive statements that gave examples of the objectives; and Ohio provided local school districts with sample instructional performance objectives for four strands.

Interpretation of Alignment by States

An analysis conducted in 1993-94 of the 25 states participating in the National Science Foundation's Statewide Systemic Initiative (SSI) program indicated that, even though many of the states had state assessments, a number of these states did not have assessments aligned with the systemic reform goals for student learning (Laguarda, Breckenridge, Hightower, & Adelman, 1994). Many of these states were shifting their assessment policies. In science, the problem was not so much alignment as it was having any form of statewide assessment. Nearly all of the states had statewide assessments in mathematics, but fewer than two-thirds of the states had statewide assessments in science. SSI staff from nearly all states reported the availability of statewide assessment data judged to be appropriate for evaluating the SSI. However, when the SSI staff were asked to judge the alignment between the statewide assessments and the curriculum and instruction promoted by the SSI, fewer indicated there was an agreement. Only 10 of the 22 states with statewide assessments in mathematics and six of 16 states with statewide assessments in science reported content alignment based on the judgement of SSI staff.

In a fall 1995 annual survey conducted by the Council of Chief State School Officers, state assessment directors were asked, What does "alignment" mean in your state? In general, responses given to this question were very brief. The most common response indicated that assessment activities and content standards were aligned by design. For New Jersey, "Aligned

means assessments will be based on the standards and indicators." The response from Michigan was that “the assessments are actually . . . designed to measure . . . outcomes and requirements stated in goals & objectives. Committees approve and reject items based upon their fit with goals and objectives." For North Carolina, alignment meant, "Curriculum frameworks provide the assessment framework for developing tests. All test questions, etc. are developed to meet the curriculum objectives." Some states selected existing tests to measure what was in their curriculum frameworks, but reported only a partial match. Although virtually all that was on the commercial tests used by North Dakota measured something in its frameworks, 40% of the mathematics frameworks was not assessed by those instruments. In a slightly different vein, Nevada noted that alignment meant there was consensus among teachers and curriculum specialists that what was being tested was what was intended to be taught.

For most of the states, frameworks and assessments were judged to be aligned if goals and learning objectives were considered in some way in the design or selection of the assessment instruments. Most states lacked a formal and systematic process for determining the alignment among standards, frameworks, and assessments.

Three Methods for Aligning Documents

Current practice and the literature suggest three major approaches—sequential development, expert review, and document analysis—for determining whether expectations and assessments are in alignment. These are not the only approaches, nor should they be considered as being applied only in their purest form. In most situations, some combination of the approaches is most applicable.

Sequential development. Policy elements, such as expectations and assessment, most frequently are aligned by design. A set of standards, for instance, might be converted directly into specifications for developing an assessment. Once one policy element is established, it becomes the blueprint for subsequent elements. The order in which the different documents are prepared can vary, with periodic revision of any one or all documents common. South Carolina had one of the more formal processes of sequential development extending over a period of eight years to reach all of the content areas. For a content area, an in-state committee developed state goals that were revised through public review. Committees of teachers and other educators used these goals to develop drafts of curriculum frameworks and content standards. These drafts were presented for broad public review and input. This type of review helped to assure vertical alignment between the expectations and public opinion. Frameworks and content standards were revised based on the feedback and then went to the State Board of Education for adoption. In fall 1995, the South Carolina State Board of Education approved academic achievement standards for mathematics based on the state's framework. These were measurable outcomes that could be used to develop assessment instruments. Public review of the draft science achievement standards were completed in the summer of 1996.
California has had some process for judging the alignment between the state assessment and the curriculum framework for a number of years. Up to around 1983 the same key people who wrote the frameworks, the intellectual and official leadership of the committees, developed the specifications for the assessment including what knowledge of content and what skills would be

measured. Booklets entitled *Rationale and Content* were then produced and widely circulated to schools to help them focus on the framework and what would be assessed. After 1983, and with the advent of performance assessment, translating ideas from the framework into test specifications was more difficult. Judgements on the linkage of the assessment with the frameworks became much more dependent on people's judgment (D. Carlson, personal communication, May 22, 1996).

States vary greatly in their policy context. These differences make it difficult to categorize states as using any single model to establish alignment, or in using one model in the same way. In South Carolina, the state legislature had to appropriate funds for the state assessment and, thus, had a strong influence on the form of assessment that would be used by issuing funds only for the form of assessment it approved. The Maryland Board of Education had the responsibility for operating the education system in that state and, with this authority, could institute an assessment system. Maryland's assessment system in 1996 had a strong influence on instructional programs in schools and was considered to be a greater influence on reform than other Maryland policy documents. Georgia had mandated that each district develop its own framework. In July 1994, the Ohio State Board of Education adopted *Ohio's Model Competency-Based Science Program* designed to guide the development of districts' curricula. In states such as Iowa, with less centralized education systems, the state department of education provided technical assistance rather than develop standards or assessments; the authority for establishing alignment in these states resided more at the district and school level than with the state.

Developing standards, frameworks, and assessments in sequence has the advantage of proceeding in a logical process and, after the development of the first document, having known criteria for the development of subsequent documents. Checks of alignment by educators, other experts, and the public can be built into the process. One disadvantage of this approach is the amount of time needed to put a sequentially developed program in place. This approach also ignores a synergism among policy elements: The development of assessments, for example, can provide useful information for thinking about instruction and what students can be expected to learn. For this reason, an iterative process may be more effective than a sequential process. Another disadvantage to sequential development is that it frequently does not reflect reality: In many states, the process for developing expectations and assessments is not linear or sequential, but more dynamic and recursive.

Expert review. In some states and districts, a panel of experts reviews the policy elements and makes some judgement on their alignment. The formality of the review process will vary. In many states and districts, the development of frameworks and standards is an open process, turning to committees and community forums of teachers, administrators, parents and the public to offer reactions and, in some, to reach consensus. Along with a process of sequential development of documents, as discussed above, most states and districts institute some review process. However, in sequential development, the more precise content reviews are conducted internally by those who are engaged in the writing or development of the documents. Reviews by external panels are not as frequent. In Michigan, the review for alignment is part of the selection process. Committees are selected to approve and reject items for inclusion in the assessment process based on their fit with the state goals and objectives. The Oregon Department of

Education convened a national panel to look at various issues related to its standards (Roeber, 1996). A subpanel looked at the alignment of the planned assessments and the standards.

Content area specialists are essential to serve as members of any review panel with the purpose of judging the match between expectations and assessment. The number in a review panel can vary, but generally should be five or more. Complex distinctions need to be made requiring a level of sophistication far exceeding general lay knowledge in understanding how students learn. Just judging the quality of either expectations or assessment and their internal consistency can be complex. In a quantitative review of the curriculum frameworks from 12 states, committees of content area specialists found frameworks lacking internal coherence in both style and content, open to misinterpretations from the use of imprecise language, containing poor translations of fundamental concepts and principles from national standards, and inaccurately interpreting learning approaches (Humphrey & Shields, 1996). All of these issues arising from the analysis of one element are confounded in comparing the alignment among multiple elements.

The tools needed for an expert review of documents will vary. A review panel given an opportunity to meet and discuss the match between documents will need less structure by experts than reviewing in separate locations. Remote reviews will require more structure and rating forms to focus reviewers' comments and allow more reasonable aggregation of comments. In either case, reviewers need clear directions of what their task is and what it is not; what will be useful and what will not. Any content analysis or analysis of agreement among complex elements can be very detailed.

A content comparison between a draft of the *National Science Education Standards* (NRC, 1996) and the *Benchmarks for Science Literacy* (AAAS, 1993) illustrate some of the difficulty (AAAS, 1995). The task of describing correspondence between these two documents was time consuming, demanded an extensive familiarity with both documents, and required deciphering two different organizations and aggregations of content. Whereas the *NSES* may provide one statement of a fundamental concept such as the food change, the same ideas were distributed among several different grade levels in the *Benchmarks*.

In another effort, Project 2061 of the American Association for the Advancement of Science experienced the complexity of designing alignment studies (Roseman, Kesidou, & Stern, 1996). The developers found they needed a minimum of four days to train three-person review teams to produce valid and reliable findings. Teams were trained to use a four-step procedure to analyze curriculum materials and judge their likely contribution to the attainment of specific learning goals as specified in the *Benchmarks* and the *National Science Education Standards*. The four steps included a preliminary inspection, content analysis, instructional analysis, and a summary report. Critical to this analysis were eight distinguishing features—specific learning goals, instruction tied to learning goals, evidence-based arguments, feedback, clarification of learning goals, specific criteria, clarification of criteria, and concrete examples of applying criteria.

Document analyses. Alignment can be judged by coding and analyzing the documents that convey the expectations and assessments. The Third International (IEA) Mathematics and Science Study (TIMSS) effectively used document analysis to judge the alignment between

national and regional curricula and assessment documents (McKnight, Britton, Valverde, & Schmidt, 1992a, 1992b; Schmidt & McKnight, 1995). An analysis of curricula along with the assessment was critical so that educational opportunity could be used as an important link among direct measures of the curriculum, teacher opinions about the content coverage, and achievement data. Initially, repeated ratings two years apart by national committees with the same members produced unstable results in judging the relationship between the curriculum and assessment instruments. The TIMSS staff then adapted a more analytic approach to comparing curricula with assessment.

TIMSS staff devised a three-dimensional grid (content by performance expectations by perspectives) to describe both the curriculum and assessment (Robitaille et al., 1993). They used 10 major content categories for mathematics:

numbers;
measurement;
geometry—position, visualization, and shape;
geometry—symmetry, congruence, and similarity;
proportionality;
functions, relations, and equations;
data representation, probability, and statistics;
elementary analysis;
validation and structure; and
other content.

These major categories were divided further into 44 subcategories. Five big categories described performance expectations:

knowing,
using routine procedures,
investigating and problem solving,
mathematical reasoning, and
communicating.

The five categories delineated the general perspectives:

attitudes,
careers,
participation by underrepresented groups,
interest, and
habits of mind.

TIMSS staff used eight broad categories for the science content dimension:

earth sciences;
life sciences;
physical sciences;
science, technology, and mathematics;
history of science and technology;
environmental and resource issues;
nature of science; and
science and other disciplines.

These categories were divided further into a total of 77 subcategories. The science categories for performance expectations and general perspectives were comparable in number and names used for mathematics, with the addition of safety to the science perspectives.

Trained national committees partitioned the textbooks, curriculum guides, and assessment instruments into blocks defined by changes in content or expected performance by students. The analysis scheme defined a block as a piece of text that was part of a lesson or unit devoted largely to one topic. The different forms of textbook content blocks included a narrative block, graphic block, exercise/question sets, activity blocks, and worked examples. The raters assigned each block a "signature" code determined by the different cells checked on a three-dimensional grid. More than one cell of the grid could be marked for any one block, but usually the number of cells for any one block did not exceed three. The national committees used the same grid to perform a similar analysis on the other document to be compared—an assessment instrument or curriculum guide.

The committees' analyses generated a set of content by performance expectations by perspectives matrices with filled-in cells. Circles (O) marked the block signatures produced by the curriculum guide analysis and plus signs (+) marked the block signatures produced by the assessment analysis. The proportion of cells with both a circle and plus represented the degree of match between the two documents. Cells with only one of the symbols indicated a mismatch. Easily computed measures of reliability determined the quality of the process. In sessions of 22 hours, TIMSS staff successfully trained committees in fifty countries to do this form of document analysis.

One practical issue arose in the document analyses and other data gathering for TIMSS. Teachers, in judging whether assessment activities matched their curriculum, were strongly

influenced by the form of the assessment activity. Teachers were more likely to judge an assessment activity to be aligned with the curriculum content demands if the activity had the same form as activities used in the curriculum. They were more likely to reject assessment activities as being aligned with the curriculum if the activities measured the same content, but in an unfamiliar way. Having examples of student work on assessment activities helped to alleviate this problem.

Porter (1995) devised another system for analyzing content to describe enacted curriculum indicators for school improvement. His system used estimates of instructional time spent with students engaged in cognitive activities and by topics to indicate the emphasis given to activities and topics. By estimating the amount of time devoted on an assessment to the same content topics and cognitive categories, the alignment between the assessment instrument and curriculum could be determined. Document analysis requires using a common metric to compare the curriculum and the assessments. TIMSS used blocks of content and Porter used degree of emphasis based on time.

Quality control. Each of the three approaches for establishing the alignment among the policy documents lend themselves to different ways of determining their quality. Any of the three approaches can be verified by using one of the other techniques as a check.

Alignment by sequential development is frequently controlled within an agency and will be less likely to have some form of external review. Even though policy elements will undergo public or expert review, the match between the documents may be given less attention or done more informally by staff. Alignment among the documents would be more credible with some external review; however, frequently time lines are so stringent due to legislative mandates or administrative pressures that time is not available for any external review. In the absence of such a review, the assurance of alignment can be strengthened by incorporating checking procedures to be used by agency staff.

The quality of expert reviews will depend on the qualifications and expertise of the reviewers. At least some reviewers need to be very knowledgeable of the content areas and learning. The number of members of a review panel and the mix of the panel will depend on a number of variables—grade ranges, depth of analysis, technical level of the documents, and purposes of the documents. Five members are a minimum to have the important perspectives represented and to obtain some measure of reliability. Providing an opportunity for reviewers to interact with each other and to build consensus will help improve the quality of the review.

Sampling techniques can be used to verify the quality of document analyses. The reliability of partitioning documents into blocks and then coding blocks can be determined by accomplished coders checking a sample of the materials.

Specific Criteria

The ultimate goal for alignment is a fully functional system working towards students learning important mathematics and science. This requires a very deep analysis of both horizontal alignment among the policy elements and vertical alignment with purposes, programs, practices, and student outcomes. Partial progress toward this ultimate goal can be determined by analyzing the alignment among policy elements using one of the above approaches, some other viable approach, or a combination of these approaches to a desired level of agreement.

Judging alignment is strengthened by using specific criteria to analyze agreement among expectations and assessments. A review of national and state standards and different alignment studies suggested the following categories of criteria—content focus, articulation across grades and ages, equity and fairness, pedagogical implications, and system applicability. A panel of assessment experts, state specialists in curriculum and assessment, and education researchers, convened by CCSSO and NISE, refined further these criteria (Appendix).

The criteria presented here are intended to provide a means for thinking about alignment. They are not considered to be the definitive list. They have not withstood the test of time or use. It is expected that this set of criteria will evolve over time. The criteria are presented in an order first to consider content, then students, then instruction, and finally application to a system. Each criterion is summarized by a brief description of what can be compared between expectations and assessments (unit of comparison) and levels (full, acceptable, and insufficient) for judging the degree of agreement.

1 - Content Focus

Expectations and assessments should focus consistently on developing students' knowledge of mathematics and science. This consistency will be present to the extent expectations and assessments share the following attributes:

A - Categorical concurrence. The same or consistent categories of content appear in both expectations and assessments. Categorical concurrence is achieved if comparable topic headings and subheadings of content appear in each. This level of detail, however, may vary: standards and frameworks can be statements of general expectations, or they can be more refined description of content. Assessment specifications and activities, designed to provide evidence of students' expected attainment, may be even more specific.

The *National Science Education Standards* (NRC, 1996) used the same eight content topics for each of three grade ranges (K-4, 5-8, and 9-12): Unifying Concepts and Processes, Science as Inquiry, Physical Science, Life Science, Earth and Space Science, Science and Technology, Science in Personal and Social Perspectives, and History and Nature of Science. For an assessment system to have categorical concurrence with these national standards, at least the most general categories of content assessed should coincide with the eight broad categories of the national standards. Alignment would be even greater if assessment results were reported by those eight categories.

Categorical Concurrence Criteria	
Unit of Comparison	
Expectations	Content topics, subtopics, or both, identified by standards or main areas of content specified.
Assessment	Topics by which results are reported (most stringent); subunits topics of instruments; or topics of clusters of assessment activities.
Scale of Agreement	
Full	A one-to-one correspondence between topics given in expectations and topics by which assessment results are reported.
Acceptable	Assessments cover a sufficient number of topics in expectations so that a student judged to have acceptable knowledge on the assessments will have demonstrated some knowledge on nearly all topics in expectations.
Insufficient	Important topics are excluded from assessments to the extent students can perform acceptably on assessments and still lack understanding of important expectation topics.

B - Depth of knowledge consistency. Depth of knowledge can vary on a number of dimensions, including level of cognitive complexity of information students should be expected to know, how well they should be able to transfer this knowledge to different contexts, how well they should be able to form generalizations, and how much prerequisite knowledge they must have in order to grasp ideas. The depth of knowledge or the cognitive demands of what students are expected to be able to do is related to the number and strength of the connections within and between mental networks. Understanding in mathematics, for example, is described as making connections between ideas, facts, or procedures (Hiebert & Carpenter, 1992, p. 67). The depth of knowledge required by an expectation or in an assessment is related to the number of connections of concepts and ideas a student needs to make in order to produce a response, the level of reasoning, and the use of other self-monitoring processes. In addition, other factors influence the cognitive demands of performance including the social or contextual requirements, the variety of representations students are expected to use (written, verbal, pictorial, and variations within each), and requirements for transfer and generalization to new situations.

Expectations and assessments are aligned if what is elicited from students on the assessments is as demanding cognitively as what students are expected to know and do. For example, the *Curriculum and Evaluation Standards for School Mathematics* published by the National Council of Teachers of Mathematics (1989) state that students in grades 9 through 12 should study data analysis and statistics, so that all students can "design a statistical experiment to study a problem, conduct the experiment, and interpret and communicate the outcomes" (p. 167). An assessment system requiring students only to interpret an existing set of data, without designing

and experimenting, would fall short and not be aligned with the depth of knowledge specified in this standard. The standard calls on students to identify a problem, draw upon their knowledge of statistics to create a design for an experiment, perform the experiment, organize the information produced, give meaning to this information, and then explain the findings. Assessing students' interpretation of a given data set would require students to demonstrate only a small part of what the standards intend.

Reality needs to influence the comparison of the cognitive demands of expectations as expressed by standards, goals, and objectives with how students are held accountable to perform on assessments for the purpose of alignment. Ideally cognitive studies would be conducted to delineate in some detail what depth of knowledge is required by an expectation and what mental operations students actually used on the corresponding assessments. Such studies can be costly and time consuming. A more realistic analysis would be to seek some expert help and conduct a content analysis using verbs and their objects to judge the match between expectations and assessments. For example, the expectation for a student to "*use and produce a variety of classification systems* to identify organisms" is more demanding cognitively than the expectation to "*discuss the classification* of organisms." Generating a variety of systems generally will require making more connections than discussing what are classifications. The full cognitive demands implied by both statements lack some clarity without more delineation of the specific criteria that will be employed to judge the successful application, use, and production of classification systems or the extent of the discussion of the classification of organisms.

An assessment activity of comparable depth of knowledge to the expectation that students will use and produce a variety of classification systems could be something like this example. A teacher shows her grade 4 students pictures and lists of a range of animals, birds, and insects. Then she asks them to form groups of these by common characteristics and to assign a name and description for each grouping. This assessment activity, in general, would require students to draw upon more specific features of the organisms and require them to sort the organisms in new ways. A satisfactory discussion by elementary students would be a general recognition of the idea of species as a basis for classifying organisms.

Expectations and assessments will be fully aligned if both are cognitively complex or both are cognitively simple. How closely a comparison between expectations and assessments can be made will depend on the specificity of the expectations. The Adams Twelve Five Star Schools, Northglenn, Colorado (1995) clarified a very general statement of standards with more specific indicators of performance. One grade 9-12 science standard stated: "The student knows and is able to demonstrate an understanding that energy appears in different forms, and can move (be transformed), and change (be transformed)" (p. 3).

This statement is too general to really judge what depth of knowledge is sought. The indicators of performance given along with the general statement gave more clarity to what was expected:

- investigates the quantitative relationships among pressure, volume, and temperatures of gases;
- characterizes chemical or physical changes as endothermic or exothermic;

- measures, calculates, and compares voltage, current, and resistance (Ohm's Law);
- measure, calculates and compares forces and changes in speed (Newton's First and Second Laws);
- contrasts properties of materials that transmit, reflect, absorb, and/or diffract visible light;
- examines the effect of lenses, mirrors, diffraction gratings, and opaque objects on visible light; and
- describes the results of water wave interactions (p. 3).

The verbs used in these indicators of performance—investigate, measure, calculate, compare, contrast—suggest that what is meant by understanding is more than recall of information but requires reasoning, applications of skills, and knowledge of scientific concepts. Assessment aligned with these expectations will have to elicit from students a fairly deep level of knowledge.

Depth of Knowledge Criteria	
Unit of Comparison	
Expectations	Rating of most cognitively demanding expected performance for a topic and for all students as determined by number of ideas integrated, depth of reasoning required, knowledge transferred to new situations, multiple forms of representation employed, and mental effort sustained.
Assessment	Rating of most cognitively demanding assessment activity for a topic and taken by all students as determined by number of ideas integrated, depth of reasoning required, knowledge transferred to new situations, multiple forms of representation employed, and mental effort sustained.
Scale of Agreement	
Full	For each major topic, the most cognitively demanding expected performance for all students is comparable to the most cognitively demanding assessment activity taken by all students.
Acceptable	For nearly all major topics, the most cognitively demanding expected performance for all students is comparable to or can be inferred from the most cognitively demanding assessment activity taken by all students.
Insufficient	Students can be judged as performing at an acceptable level on the assessments without having to demonstrate for any topic the attainment of the most cognitively demanding expected performance for all students.

C - Range of knowledge correspondence. Expectations and assessments cover a comparable span of knowledge within topics and categories. Even with strong categorical concurrence (criterion A), the span of expected knowledge within categories may not always be entirely covered by the assessment system. Tradeoffs frequently have to be made. Time limitations, scoring costs, availability of instruments, and other constraints on assessment can prohibit the measurement of the full range of content or performance expectations for every student. Completely aligned expectations and assessments requires an assessment system designed to measure in some way the full range of expected knowledge within each specified topic. For example, according to the standards published by the Virginia Board of Education (1995), students should be able to read four different types of maps: bathymetric, geologic, topographic and weather. An assessment corresponding to the range of knowledge of this map-reading expectation, would need to measure how well students can interpret information using all four map types.

Thoughtfully designed assessments help to increase the correspondence between the assessed range of knowledge with the expected range of knowledge. Clear assessment specifications can be more helpful in determining that a full range of knowledge is being assessed than the actual assessment instruments. This is particularly true if activities on an instrument represent a selection of possible activities from a conceptual domain defined in detail in the assessment specifications.

One important consideration in judging the correspondence of the range of knowledge is the reference group for reporting results—system, district, school, or individual. Formally assessing the full range of knowledge as described in standards, goals, and objectives for every student in a large system can be extremely difficult, if not impossible. However, with sampling techniques, performance information on the full range of expected knowledge could be gathered for a large group of students. Construct and content domain definitions included in assessment specifications generally identify the range of content, or some decision rule for generating assessment activities, to assure full coverage of content. There will be a high degree of match between expectations and assessments for the total population if the assessment design assures that a reasonable range of knowledge within topics is assessed and appropriate sampling procedures are used so that inferences can be made on this knowledge for the total population. There will be a low degree of match of assessments with expectations for the total population if the designed assessment measures only one or a few special cases of a complex concept or topic.

Alignment between expectations and assessments on the range of knowledge can be achieved in a system even if formal assessments or large-scale assessments do not fully span every topic. Teachers incorporating in their classrooms assessment practices that attend to measuring knowledge within specific topics not fully covered by large-scale assessments contribute towards a fully aligned system.

Range of Knowledge Criteria	
Unit of Comparison	
Expectations	All types or forms of major concepts or ideas within and across performance standards, such as type of graphs, all students are expected to know how to use and all forms of representations they are to be able to use.
Assessment	All types or forms of major concepts or ideas included on assessments or in the specifications of the content domains used to select assessment activities.
Scale of Agreement	
Full	Students are required on assessments to demonstrate knowledge of all forms or the full range of each major concept or idea expressed in the expected performance.
Acceptable	Assessment specifications account for nearly all forms or the full range of each major concept or idea expressed in the expected performance so there is a strong likelihood that students' knowledge and use of all forms will be assessed.
Insufficient	Important forms or specific cases of major concepts and/or ideas given in the expected performance are excluded from or ignored on assessments or their specifications.

D - Structure of knowledge comparability. The underlying conception of science and mathematics knowledge in expectations and assessments are in agreement. For example, if standards indicate that students should understand mathematics “as an integrated whole” (NCTM, 1989) or “science as inquiry” (NRC, 1996), then the assessment activities should be directed toward the same ends. Both expectations and assessments should embody similar requirements for how students are to draw connections among ideas. Assessment of knowledge only as fragmented skills, for example, would not be in full alignment with the national standards.

The preliminary draft of the Illinois Academic Standards (1996) expressed, "Solving problems is at the heart of mathematics. Mathematics is a collection of concepts and skills; it is also a means of investigation, reasoning, and communicating" (p. 25). One of five applications of learning in this document was "Making Academic Connections: Recognize and apply connections of important information and ideas within and among academic learning areas" (p. 26). The depiction of mathematics in these statements imply that students should learn a collection of concepts and skills along with how to solve problems and apply information to other fields. An assessment system aligned with a comparable structure of knowledge as represented by these expectations would have to gather information on students' understanding of specific concepts

and skills, ability to solve problems, and application of mathematics to other academic learning areas.

Judging the structure of knowledge comparability, along with many of the other criteria, requires considering the full assessment system. Expectations and assessment are fully aligned according to this criterion if the structure of mathematical understanding expressed in expectations is represented in the assessment system. In the case of judging whether an assessment system is aligned with the Illinois Academic Standards, some part of the system should gather information on how students are able to identify appropriate applications of mathematical ideas both within mathematics and in other fields. For example, students could be asked to develop a mathematical model of a physics experiment. Alignment with the Illinois Academic Standards will be less if little or no part of the assessment system asks students to demonstrate knowledge of how they are able to draw relationships among mathematical ideas and among mathematical ideas with applications to other fields. The standards and assessments will have very low alignment in the structure of knowledge if students are only asked to demonstrate their knowledge of mathematics by recalling or applying isolated concepts, procedures, and skills.

Structure of Knowledge Criteria	
Unit of Comparison	
Expectations	Expression of how students are expected to form relationships among ideas such as no relationship (single or isolated ideas), equivalent forms of the same idea, connection of many ideas within the content area, and connection of ideas within the content area and with applications to other areas.
Assessment	Expression of the relationships among ideas students are required to successfully perform on assessments or as indicated on the assessment specifications such as no relationship among ideas, equivalent forms of same idea, connection of many ideas within the content area, and connection of ideas within the content area and with applications to other areas.
Scale of Agreement	
Full	The relationships among ideas expressed in the performance expectations are the same as those required to perform successfully on the assessments.
Acceptable	Inferences can be made on how each student has formed relationships among ideas as expressed in the performance expectations.
Insufficient	A student could be judged as fulfilling the expectations without adequate demonstration on assessments that the student is able to draw relationships among ideas as represented by all important structural forms.

E - Balance of representation. Similar emphasis is given to different content topics, instructional activities, and tasks. The expectations and assessments give comparable emphasis to what students are expected to know, what they should be able to do, and in what contexts they are expected to demonstrate their proficiency. Fulfillment of the four prior criteria requires expectations and assessments to cover comparable topics in the same depth and breadth based on a common understanding of how knowledge is organized. For the Balance of Representation criterion to be met, the degree of importance of different ideas given in the assessments and expectations should be the same.

Standards frequently describe what all students are expected to know and do without differentiating the priority or emphasis one standard should be given over another. It is assumed, if not stated directly, all of the standards are important. The equal importance of all content standards given in the *National Science Education Standards* is reinforced by statements such as "None of the eight categories of content standards should be eliminated. . . . No standards should be eliminated from a category." (NRC, 1996, pp. 111-112). The New Jersey Mathematics Coalition and the New Jersey Department of Education (1996), in the *New Jersey Mathematics Curriculum Framework*, indicated some differentiation among the 18 standards (Rosentein, Caldwell, & Crown, 1996). This document described four "mathematical processes" standards as the *major* focus for all students—posing and solving problems, communication, making connections, and reasoning. Each of the 10 mathematics content standards was stated as what all students will develop, but without a superlative.

Assessments, as well as curricula, designed to fulfill expectations and standards are constrained by very pragmatic factors such as time, sequencing, and a high variation in the rate of learning. These constraints force those who develop assessments to make decisions about the amount of emphasis or weight that will be given to different topics on a test. A state assessment for grade 5 may be limited to one class period for each content area. If every student has to take the same test during this class period, then tradeoffs will have to be made among the number of open-ended activities, requiring time and extensive effort, and the number of short answer activities, requiring recall or reproduction of information. The psychometric qualities of the test also will need to be considered.

For example, although the national science standards place an equal weight on each of the eight standards for each of three grade ranges, when these expectations are considered in more depth and across grade ranges there is some variation in emphasis. Under the Earth and Space Science category of standards, emphasis for grades K-4 is on developing observation and description skills and basing explanations on observations (NRC, 1996, p. 134). In the middle grades, on the same category, more emphasis is given to constructing models that explain visual and physical relationships (p. 159). For an assessment system to have a comparable balance of representation with the *National Science Education Standards* would require a shift from grades K-4 to grades 5-8 in how students' knowledge is judged from primarily observing and describing to primarily drawing inferences from relationships. In the lower grades more emphasis on assessments should be given to determine whether students can observe and describe properties of earth materials and objects in space. In the middle grades, more emphasis on assessments would have to be

given to students demonstrating their understanding of the relationship among components of the earth system and using a model of the physical relation among earth, sun, moon, and the solar system to explain such phenomena as the phases of the moon.

Balance of Representation Criteria	
Unit of Comparison	
Expectations	Assigned importance on a scale of 100 by topic over full spectrum of performance expectations (Total for all topics should be 100.)
Assessment	Weight by topic or subtopics for full spectrum of assessments (weight could be determined by the proportion of activities by topic, proportion of average time allocated to do an assessment activity by topic, or according to some other rule).
Scale of Agreement	
Full	The proportion of assigned importance for topics in performance expectations is equivalent to the weight topics are given on assessments.
Acceptable	Distribution of importance by topics in performance expectations nearly matches the weight of topics in assessments without major exclusions.
Insufficient	Weights on assessment by topic are sufficiently different from the assigned importance in the performance expectations such that a student could be judged as meeting performance expectations without knowledge of highly emphasized topics.

F - Dispositional consonance. When expectations include more than learning concepts, procedures, and their applications—such as molding student attitudes and beliefs about science and mathematics—assessments also should support that broader vision. For example, the *National Science Education Standards* underscore the importance of students becoming self-directed learners. The ability for students to self-assess their understanding is an essential tool for this. Assessment practices aligned with this goal will include opportunities for students to critique their own work and to explain how work samples provide evidence of understanding. To achieve this, teachers need to give students opportunities to reflect on their scientific understanding and abilities, so they can begin to internalize the expectation that they can learn science.

Determination of dispositional consonance between expectations and assessments requires considering the assessments administered on different levels. Classroom assessment of students' habits and attitudes towards science and mathematics frequently are done informally by teachers looking for cues of students' developing dispositions. Even though evidence intimately links affect and cognition (McLeod, 1992), formally measuring individual attitudes in conjunction with achievement is difficult. What may be more easily achieved is obtaining some indication of attitudes and dispositions for large groups. The 1994 North Carolina Competency-Based

Curriculum listed as one of five program goals, "Develop responsible attitudes toward the environment, science, technology, and society" (p. 5). The document went on to clarify that "student attitudes **will not** (emphasis in original text) be a part of an achievement [end-of-course or end-of-grade] score" (p. 11). The state assumed the responsibility for measuring how students view science through its state assessments of accumulative knowledge and not at the end of courses. Some degree of alignment with respect to dispositional consonance then, in this case, is achieved by gathering group assessment indicators related to expectations rather than individual assessment indicators.

Dispositional Consonance Criteria	
Unit of Comparison	
Expectations	List of desired dispositions toward the content area students are to develop including habits, attitudes, and other qualities.
Assessment	List of dispositions toward the content area on which information is gathered, either formally and informally, reported, and used to make decisions about students.
Scale of Agreement	
Full	Main expected dispositional qualities are observed, monitored, and reported at designated levels within the system.
Acceptable	Some effort is made and attention is given to observing students' development of expected dispositional qualities.
Insufficient	Little or no attention is given to observing or monitoring dispositional qualities even though these qualities are advanced as important; or only easily measured attitudes are incorporated in assessments while more prominently desired qualities are ignored.

2 - Articulation Across Grades and Ages

Students' knowledge of mathematics and science grows over time. Expectations and assessments should be rooted in a common view of how students develop, and how best to help them learn at different developmental stages. Over time, students' growing understanding of science concepts and processes will allow them to perform higher levels of analysis and work with a greater tolerance of criticism and uncertainty. Students' increased understanding of number, operations, and generalizations will promote more abstract and algebraic thinking. For expectations and assessments to be aligned, they need to be grounded in a similar view of cognitive development and advancement in knowledge. This common view should be based on:

A - Cognitive soundness determined by best research and understanding. There has been considerable research on the learning of mathematics and science, which has produced extensive

knowledge of how students mature in their understanding of these content areas (Romberg & Carpenter, 1986; Stein, Grover, & Henningsen, 1996). Expectations and assessments should build on this knowledge to develop a sound learning program, and they should do so in ways that are aligned. Students' knowledge of mathematics can be thought of as internal networks of representations (Hiebert & Carpenter, 1992). Understanding is built gradually as new information is connected to existing networks of ideas. According to this view of understanding, new networks are constructed by reconfiguring old networks and forming new connections. Thinking and reasoning develops along with understanding of routine skills (Resnick & Resnick, 1992). The alignment between expectations and assessments will be strengthened if they both are grounded in a common view of how understanding is developed, such as through continual expansion of networks of representations.

The preliminary draft of the Illinois Academic Standards (1996) for mathematics listed for each academic standard learning benchmarks to further delineate across five grade ranges what was expected for students. Benchmarks for a measurement academic standard—measure and compare quantities using appropriate units, instruments and methods—represented a developing understanding of measurement ideas over four grade ranges. Early elementary students were to measure quantities with customary and metric systems, relating a metric to a quantity. Late elementary students were to compare and convert units of measures of quantities, expanding their internal "measurement networks" to relationships among different forms of units. Middle or junior high school students were to apply the concepts and attributes of measures to practical situations, forming new links between their understanding of measurement to useful applications. Early high school students were to apply units, domains or ranges and scales to describe and compare functions, numerical data and physical objects, expanding the "measurement networks" even further to more abstract quantities. For an assessment system to be aligned with this measurement standard, assessment instruments used over these grade levels need to be based on a similar view of how students grow in understanding of measurement from forming single relationships between units and quantities to a more complex network of ideas incorporating more abstract units and quantities.

Cognitive Soundness Criteria	
Unit of Comparison	
Expectations	The expressed or implied underlying understandings and theory of how students' learning matures over time (behavioral, constructive, developmental levels, combination, etc.) and accepted individual differences among students.
Assessment	Analysis across grades of the implied theory (theories) of learning represented in assessment system instruments, procedures, and attention to individual differences.
Scale of Agreement	
Full	Assessment instruments for each grade level are developmentally

	appropriate and, across levels, represent a reasonable progression in learning as depicted or implied in the expectations.
Acceptable	Assessment instruments for each grade level represent most, but not all, developmental expectations and generally follow the same progression in learning depicted or implied in the expectations.
Insufficient	Assessment instruments include activities out of sync with the explicit or implied developmental expectations and the sequence of instruments across grades do not depict the same progression of learning as stated or implied in the expectations so that students' knowledge is overstated or understated.

B - Cumulative growth in content knowledge during students' schooling. Expectations and assessments should be linked by underlying rationales of mathematics and science as content areas. Although the learning of mathematical and scientific concepts over time doesn't follow a strict order of steps, students often need to grasp certain concepts and ideas in order to address more advanced ideas. In order for high school students to take part in scientific inquiry, for example, they first need to learn to identify questions and concepts that guide scientific investigations, to design and conduct such investigations, to use technology and mathematics, to formulate and revise scientific explanations, and to recognize and evaluate alternative explanations. For high school students to incorporate more abstract knowledge, such as the structure and function of DNA, they need to have developed in grades K-8 foundational understanding of life sciences. Aligned expectations and assessments describe and represent, in complementary fashion, the underlying structure of content knowledge students need to develop, and how their instructional experiences should be organized. Assessments at a given grade level should reflect whether students are expected by that grade level to have a foundational knowledge, expanding knowledge, or fully mature knowledge of the assessed concept or idea.

Cumulative Growth in Content Knowledge Criteria	
Unit of Comparison	
Expectations	The expressed or implied understanding of how students' knowledge of content will be structured and will mature over time.
Assessment	Analysis across grades of how the growth in understanding of content ideas and the relationships among content ideas are represented in assessment system instruments and procedures.
Scale of Agreement	
Full	Assessment instruments elicit information compatible with how students' knowledge of content ideas develops over time and how students relate these ideas with each other as reflected in expectations.

Acceptable	Assessment instruments elicit information according to general patterns of how students' knowledge of content ideas develops over time and how students relate these ideas with each other as reflected in expectations.
Insufficient	Assessment instruments across the grades do not represent a logical or sequential growth in student content knowledge over time implied in the expectations. Assessments in lower grades require a more advanced understanding of ideas than do those in later grades as depicted in the expectations. Or, important stages or indicators in the development of content areas as depicted in the expectations are excluded from the assessment system.

3 - Equity and Fairness

When expectations are that all students can learn to high standards, aligned assessments must give every student a reasonable opportunity to demonstrate attainment of what is expected. Expectations and assessments 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 demonstrate the expected level of learning. The knowledge a student will demonstrate on an assessment can vary by the form of assessment (Baxter, Shavelson, Herman, Brown, & Valadez, 1993). Even a slight variation in the context of a question can alter performance, such as giving a bare arithmetic problem or the same mathematical content in a word problem (Clements, 1980; Van den Heuvel-Panhuizen, 1996). Rarely will one form of assessment be capable of producing valid evidence for all students. For example, only half of students who answered a multiple-choice question on percents given on the 1986 National Assessment of Education Progress gave a satisfactory response to a following open-ended question that asked them to explain their response (Gay & Thomas, 1993). Students judged to have understood the concept by a correct response to the multiple-choice question had a clear misunderstanding of the concept as noted by their written response.

A student's ability to perform well on a particular assessment can depend on a number of factors in addition to the level of knowledge, including culture, social background, and experiences (Santos, Driscoll, & Briars, 1993). Therefore, expectations and assessments will be better aligned, and more equitable, if alternative forms of assessment to measure student attainment and procedures are in place to assure that the knowledge of each student has been fairly assessed. The challenge becomes developing and maintaining an aligned system with a variety of means of assessment that ensures each student's full attainment of an expectation is measured. Students' unsuccessful demonstration on one measure could be an interaction with the form of assessment rather than the content. An aligned assessment system will provide opportunities for students who are unsuccessful on one form to demonstrate their understanding by taking an alternative form or taking the same form under different conditions as long as validity and reliability are maintained.

Assessment of expectations that are more open-ended and require students to apply knowledge to their life experiences will produce very diverse responses. Grades 3-6 students in South Carolina were expected to "participate in problem-solving activities through group and individual investigations so that they can relate the use and understanding of numeration systems to their world" (South Carolina State Department of Education, 1993, p. 48). Because students' life experiences are very different, their demonstration of the application of numbers to their world can vary greatly. For assessment practices to be equitably and fairly aligned with expectations, such as this one from South Carolina, student responses need to be judged on the adequacy of the application of numbers and not on what experiences students portray, some of which may not be within the experiences of the person scoring the assessment. The South Carolina Mathematics Framework emphasized this point under its principles and goals for mathematics assessment, "To be fair to all, assessments must be sensitive to cultural, racial, and gender differences" (p. 110). Expectations for students to apply their knowledge of mathematics and sciences can be met in many different ways. For assessments to be aligned with these expectations, they also need to be similarly robust so as not to misrepresent what science and mathematics a student truly knows or can do. An underlying principle is to maximize the participation of students achieving the expectations and demonstrating on assessments what is the full extent of their knowledge.

It may be difficult to gauge alignment between expectations and assessments on the criteria of fairness and equity until both have been in place for some time. Consistently low scores over time on an assessment of a particular learning goal may be the result of many factors, including misplaced expectations, rather than poor instruction or lack of effort by students. Students may be developmentally unprepared to attain a particular expectation, for example, or the structure of the curriculum may keep them from attaining sufficient experiences to learn what is expected. Time is required for patterns to form in order to decipher how expectations and assessments are working in concert with each other to be equitable and fair.

Equity and Fairness Criteria	
Unit of Comparison	
Expectations	Levels of knowledge that all or different groups of students are expected to achieve by specific times (e.g. content knowledge understanding by all students by end of grade 12, by college-intending students by grade 12 . . .).
Assessment	Degree to which the assessment system affords students a reasonable opportunity to demonstrate the full level of knowledge as expected for all students or for their specific group.
Scale of Agreement	
Full	Students are afforded a fair and reasonable opportunity to demonstrate the full level of knowledge expected for all students. Assessment practices are such that variation of assessment results are only a variation in the attainment of expectations and free from being

	influenced by culture, ethnicity, gender, or any other irrelevant factor.
Acceptable	Assessment practices are appropriate to measure the attainment of expectations for all or any designated group of students while minimizing culture, ethnicity, or gender bias.
Insufficient	Important judgments are made on students' attainment of expectations based on biased or limited assessment practices that do not afford a student or group of students a reasonable opportunity to demonstrate what they know and can do.

4 - Pedagogical Implications

Classroom practice greatly influences what students learn. 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.

Judging the pedagogical implications of expectations and assessments requires more than simple content analysis. Any such review must attempt to gauge the likely implications on pedagogy. Meaningful analyses have been done by directly asking teachers how they interpret expectations and assessments, and how their classroom practices fit with them (Cohen, 1990; Romberg, Zarinnia, & Williams, 1990).

Of course, the true test is what happens in the classroom. For example, educators are now paying increased attention to the importance of involving students in scientific inquiry, hands-on learning, and more “authentic” instruction (Newmann, Secada, & Wehlage, 1995). Assessments that suggest a more passive type of instruction would be less aligned with those expectations. Likewise, expectations that indicate students are to develop abilities to perform scientific inquiry through actively constructing ideas and explanations (NRC, 1996) will lack full alignment with assessments that are solely based on an assumption that students have memorized the canonical ideas and explanations. Alignment is achieved when the instructional practices and materials implied by expectations and those implied by assessments are consistent.

Critical elements to be considered in judging alignment include:

A - Engagement of students and effective classroom practices. Traditional forms of student assessment and the constraints imposed by limits on time and other resources may place an inordinate influence on the superficial acquisition of skills and facts or any one area in relationship to others. In this way, education systems can gravitate toward readily measured outcomes, instead of more complex but also more desirable outcomes, such as students being able to investigate, create models, or otherwise demonstrate deeper content knowledge (McCarthy, 1994).

Expectations and assessments need to work together to provide consistent messages to teachers, administrators, and others about the goals of learning activities. For example: A preliminary draft

of statewide academic standards for Illinois indicated that students should learn and contribute productively both as individuals and as members of groups (Illinois Academic Standards Project, 1996). This was defined in the draft as an important skill that will greatly determine the success of students later in life. But if no part of the assessment system produces evidence of whether students are contributing productively as members of groups, then teachers would receive conflicting messages about how much classroom time should be spent having students work in teams.

Engagement of Students and Effective Classroom Practices Criteria	
Unit of Comparison	
Expectations	Range of instructional practices most likely for all students to achieve the full extent of expectations.
Assessment	Range of instructional practices most likely for all students to develop the knowledge and experiences to perform satisfactorily on assessments.
Scale of Agreement	
Full	Instructional practices most likely to have students fully achieve expectations are the same as the instructional practices most likely to have students adequately demonstrate their attainment of these expectations on available assessments.
Acceptable	Students are not disadvantaged by instructional practices required for them to do well on assessments in achieving the full extent of the expectations.
Insufficient	For students to do well on assessments forces teachers to give undue emphases to instructional practices that inhibit students' learning content to the full extent as expressed by the expectations.

B - Use of technology, materials, and tools. Technology, materials and tools are vital to knowing and "doing" mathematics and science today. Students should develop skill and confidence using tools such as calculators and computers in their everyday lives (National Council of Teachers of Mathematics, 1991). In science they should increase their repertoire of tools and techniques and improve their skills in measurement, calculation, and communications (American Association for the Advancement of Science, 1993).

The draft of the *Texas Essential Knowledge and Skills* for science (Texas Education Agency, 1996) stratified expectations as basic understanding, knowledge and skills, and performance descriptions. One basic understanding for middle grades was that change occurs and can be observed and measured. A knowledge and skills goal for this basic understanding was for

students to know that theoretical work done by early scientists has revolutionized modern science. This goal was delineated further by performance descriptions including:

> The sixth-grade student, using resources including books, periodicals, videos, technology, and experts, identifies the contributions of different scientists from different cultures at different times (p. 27).

Fully assessing students' attainment of this performance expectation requires determining whether students can use these resources to learn what different scientists from different cultures have done. Lessening the requirement on an assessment by having students respond to questions rather than actually having them use materials could result in lowering the priority for students to use multiple resources in doing research. Such an assessment would not be aligned with the expectation and could influence teachers to give less attention to students meeting the full intent of the expected performance.

Expectations and assessments need to send students consistent messages about technology and how it is related to what they are expected to learn. If standards indicate that students should learn to routinely use calculators or computers, for example, then the curriculum should provide adequate opportunity for students to use them in this manner. To be aligned, assessments should allow students to use calculators and computers effectively to derive correct answers.

Use of Technology, Materials, and Tools Criteria	
Unit of Comparison	
Expectations	The technology, materials, and tools students need to be skilled in using to achieve the full extent of expectations.
Assessment	The technology, materials, and tools students need to be skilled in using to perform satisfactorily on assessments.
Scale of Agreement	
Full	Adequate performance on assessments require students to be accomplished in using the full range of technology, materials, and tools as intended by the expectations.
Acceptable	Adequate performance on assessments require students to be accomplished in using a sampling of technology, materials, and tools as intended by the expectations.
Insufficient	Students are prohibited on assessments from using technology, materials, or tools that students are expected to become accomplished in using.

5 - System Applicability

Although expectations and assessments 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 important and useful mathematics and science. Expectations and assessments are in alignment based on system applicability if those who have a stake in the education system, those who are to implement the system, and those who are to be held accountable are able to understand these documents, see how they are related, and believe they are attainable. Public review and discussion of expectations has been used by many states to gain public buy-in of expectations. Having assessment as an open process (NCTM, 1995) can further public understanding and acceptance of how attainment of expectations will be measured. Ongoing review of how the expectations and assessments are working as a system and as intended will help assure system applicability.

System Applicability Criteria	
Unit of Comparison	
Expectations	Degree all important system stakeholders understand, accept, and value expectations; and the degree they think expectations are attainable.
Assessment	Degree all important system stakeholders understand, accept, and valued assessments; and the degree they think assessments measure important knowledge and skills.
Scale of Agreement	
Full	The public, teachers, students, and others within the system view expectations and assessments as closely linked, acceptable, attainable, and important.
Acceptable	The public, teachers, students, and others within the system do not fully understand the link between expectations and assessments, but are favorable towards them and are willing to support students' attainment of them.
Insufficient	The public, teachers, students, and others see little relationship between expectations and assessments and give weight to one over the other.

Conclusions

These five categories are intended to be a comprehensive set for judging the alignment between expectations and assessments. Each general category and all subcategories are important in ascertaining the coherence of a system—the degree that assessments and expectations converge to direct and measure student learning. In practice, to reach full agreement between expectations and assessments on all criteria is extremely difficult. Tradeoffs will need to be made because real constraints exist on any education system such as resources, finances, time, and legal authority. Decisions on what tradeoffs are to be made among these criteria or on what level of compliance will be acceptable should be made in full awareness of potential consequences.

How decisions on tradeoffs and lessening compliance will be made will depend on a number of factors. The assessment of content knowledge in depth sometimes prohibits assessment of the full range of content. Sampling of content is generally required on most forms of assessment. The burden falls on those in the system to justify what is acceptable and reasonable within the existing context. In any case, assessment should be thought of very broadly including system-wide, local, and classroom assessments. The assessment of the attainment of some expectations to the depth and breadth required most likely will be the responsibility of classroom teachers.

The underlying principle is to have a high degree of match between what students are expected to know and what information is gathered on students' knowledge. If the criteria are not fully met, this should be done in full awareness of what action is being taken and to what degree alignment is weakened.

Above all else, when judging the alignment within a system and using these criteria to consider the relationship of expectations and assessments, a sense of reality needs to be maintained. The available resources, the amount of time required, legislative mandates, and other factors will influence how well alignment can be determined, and how practical it is to make such determinations.

Alignment of expectations and assessments is a key underlying principle of systemic and standards-based reform. Establishing alignment among policy elements is an early activity for improving the potential for realizing significant reform. Those working to build aligned systems should not think too narrowly about the task. The criteria presented here demonstrate that a number of factors can be considered in judging alignment among policy elements. These can be studied in several alternative and potentially complementary ways.

In approaching reform, the consideration of alignment cannot come too soon. And just as educators need to remain vigilant to assure that expectations, assessments, and instructional practices are current, they also will need to review the alignment among these major policy elements as new policies are instituted, new administrative rules are imposed, and system needs are changed.

References

Adams Twelve Five Star Schools. (1995). *Science curriculum framework grades 9-12*. Northglenn, CO: Author.

American Association for the Advancement of Science (Project 2061). (1993). *Benchmarks for science literacy: Science for all Americans*. New York: Oxford University Press.

American Association for the Advancement of Science. (1995). *Summary comparison of content between the November draft of National Science Education Standards and Benchmarks for science literacy: Science for all Americans*. Washington, DC: Author.

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, NJ: Prentice Hall.

Baxter, G. P., Shavelson, R. J., Herman, S. J., Brown, K. A., & Valadez, J. R. (1993). Mathematics performance assessment: Technical quality and diverse student impact. *Journal for Research in Mathematics Education, 24*(3), 190-216.

Blank, R. K., & Pechman, E. M. (1995). *State curriculum frameworks in mathematics and science: How are they changing across the states?* Washington, DC: Council of Chief State School Officers.

Bybee, R. W. (1995). Achieving scientific literacy. *The Science Teacher*, *62*(10): 28-33.

Chubin, D. E. (in press). Systemic evaluation and evidence of education reform. In D. Bartels & J. O. Sandler (Eds.), *Implementing science education reform: Are we making an impact*? Washington, DC: American Association for the Advancement of Science.

Clements, M. A. (1980). Analyzing children's errors on written mathematical tasks. *Educational Studies in Mathematics*, *11*, 1-21.

Cohen, D. K. (1990). A revolution in one classroom: The case of Mrs. Oublier. *Educational Evaluation and Policy Analysis*, *12*(3), 327-345.

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

Council of Chief State School Officers. (1996). *Key state education policies on K-12 education: Content standards, graduation, teacher licenses, time and attendance*. Washington, DC: Author.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 443-507). Washington, DC: American Council on Education.

Ferrini-Mundy, J. & Johnson, L. (1996). Building the case for standards-based reform in mathematics education. In Bartels, S. (Ed.), *Implementing science education reform: Are we making an impact? AAAS Yearbook (36)*. Washington, DC: American Association for the Advancement of Science.

Gay, S., & Thomas, M. (1993). Just because they got it right, does it mean they know it? In N. L. Webb & A. F. Coxford (Eds.), *Assessment in the mathematics classroom*. *1993 Yearbook* (pp. 130-134). Reston, VA: National Council of Teachers of Mathematics.

Harmon, M. (1991). Fairness in testing: Are science education assessments biased? In G. Kulm & S. M. Malcom (Eds.), *Science assessment in the service of reform* (pp. 31-54). Washington, DC: American Association for the Advancement of Science.

Hiebert, J., & Carpenter, T. P. (1992). Learning and teaching with understanding. In D. A. Grouws (Ed.), *Handbook of research on mathematics teaching and learning: A project of the National Council of Teachers of Mathematics* (pp. 65-97). New York: Macmillan.

Humphrey, D. C., & Shields, P. M. (1996). *A review of mathematics and science curriculum frameworks*. Menlo Park, CA: SRI International.

Illinois Academic Standards Project. (1996). *Preliminary draft: Illinois academic standards for public review and comment, English language arts and mathematics, Volume One, State goals 1-10.* Springfield, IL: Author.

Laguarda, K. G., Breckenridge, J. S., Hightower, M. M., & Adelman, N. E. (1994). *Assessment programs in the Statewide Systemic Initiatives (SSI) states: Using student achievement data to evaluate the SSI*. Washington, DC: Policy Studies Associates.

Madaus, G. F. (1983). *The courts, validity, and minimum competency testing*. Boston: Kluwer-Nijhoff.

Mathematical Sciences Education Board. (1993). *Measuring what counts. A conceptual guide for mathematics assessment*. Washington, DC: National Academy Press.

McCarthy, C. (1994). Being there—A mathematics collaborative and the challenge of teaching mathematics in the urban classroom. In N. L. Webb & T. A. Romberg (Eds.), *Reforming mathematics education in American's cities*. *The Urban Mathematics Collaborative Project* (pp. 173-195). New York: Teachers College Press.

McKnight, C., Britton, E. D., Valverde, G. A., & Schmidt, W. H. (1992a). *Survey of mathematics and science opportunities: Research report series No. 42: Document analysis manual.* East Lansing: Michigan State University, Third International Mathematics and Science Study.

McKnight, C., Britton, E. D., Valverde, G. A., & Schmidt, W. H. (1992b). *Survey of mathematics and science opportunities: Research report series No. 43: In-depth topic trace mapping, draft 4.* East Lansing: Michigan State University, Third International Mathematics and Science Study.

McLeod, D. B. (1992). Research on affect in mathematics education: A reconceptualization. In D. A. Grouws (Ed.), *Handbook of research on mathematics teaching and learning: A project of the National Council of Teachers of Mathematics* (pp. 575-596). New York: Macmillan.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 13-103). New York: Macmillan.

Messick, S. (1994, March). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, *23*(2), 13-23.

Moss, P. A. (1992). Shifting conceptions of validity in educational measurement: Implications for performance assessment. *Review of Educational Research*, *62*(3), 229-258.

National Academy of Sciences. (1997). *Preparing for the 21st century. The education imperative*. Washington, DC: Author. (http://www2.nas.edu/21st)

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

National Council of Teachers of Mathematics. (1991). *Professional standards for teaching mathematics*. Reston, VA: Author.

National Council of Teachers of Mathematics. (1995). *Assessment standards for school mathematics.* Reston, VA: Author.

National Research Council. (1996). *National science education standards*. Washington, DC: National Academy Press.
Newmann, F. M. (1993). Beyond common sense in educational restructuring: The issues of content and linkage. *Educational Researcher*, *22*(2), pp. 4-13, 22.
Newmann, F. M., Secada, W. G., & Wehlage, G. G. (1995). *A guide to authentic instruction and assessment: Vision, standards, and scoring*. Madison, WI: Center on Organization and Restructuring of Schools.
Office of Technology Assessment. (1988). *Technology and the American transition*. Washington, DC: U.S. Government Printing Office.
Porter, A. C. (1995). *Developing opportunity-to-learn indicators of the content of instruction: Progress report*. Madison, WI: Wisconsin Center for Education Research.
Resnick, L. B., & Resnick, D. P. (1992). Assessing the thinking curriculum: New tools for educational reform. In B. R. Gifford & M. C. O'Connor (Eds.), *Changing assessments: Alternative views of attitude, achievement and instruction* (pp. 37-75). Norwell, MA: Kluwer.
Robitaille, D., Schmidt, W. H., Raizen, S., McKnight, C., Britton, E., & Nicol, C. (1993). *The third international mathematics and science study; Monograph No. 1: Curriculum frameworks for mathematics and science*. Vancouver, Canada: Pacific Educational Press.
Roeber, E. D. (1996). *Review of the Oregon content and performance standards. A report of the National Standards Review Team prepared for the Oregon Department of Education*. Salem: Oregon Department of Education.
Romberg, T. A., & Carpenter, T. P. (1986). Research on teaching and learning mathematics: Two disciplines of scientific inquiry. In M. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed., pp. 850-873). New York: Macmillan.
Romberg, T. A., Zarinnia, E. A., & Williams, S. (1990). *Mandated school mathematics testing in the United States: A survey of state mathematics supervisors.* Madison, WI: National Center for Research in Mathematical Sciences Education.
Rosenstein, J. G., Caldwell, J. H., & Crown, W. D. (1996). *New Jersey Mathematics Curriculum Framework*: *A collaborative effort of the New Jersey Mathematics Coalition and the New Jersey Department of Education*. New Brunswick, NJ: The New Jersey Mathematics Coalition, Rutgers, The State University of New Jersey.
Santos, M., Driscoll, M., & Briars, D. (1993). The classroom assessment in a mathematics network. In N. L. Webb, & A. F. Coxford (Eds.), *Assessment in the mathematics classroom*. *1993 Yearbook* (pp. 220-228). Reston, VA: National Council of Teachers of Mathematics.
Schmidt, W. H., & McKnight, C. (1995). Surveying educational opportunity in mathematics and science: An international perspective. *Educational Evaluation and Policy Analysis*, *17*(3), 337-353.
South Carolina State Department of Education. (1993). *South Carolina curriculum frameworks*. Columbia, SC: Author.
Stein, M. K., Grover, B. W., & Henningsen, M. (1996). Building student capacity for mathematical thinking and reasoning: An analysis of mathematical tasks used in reform classrooms. *American Educational Research Journal, 33*(2), 455-488.
Texas Education Agency. (1996). *Science: Texas essential knowledge and skills*. Austin, TX: Author.

U.S. Congress, House. (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.

Van den Heuvel-Panhuizen, M. (1996). *Assessment and realistic mathematics education*. Utrecht, The Netherlands: Freudenthal Institute.

Virginia Board of Education. (1995). *Standards of learning for Virginia public schools.* Richmond, VA: Author.

Webb, N. L. (1993). Mathematics education reform in California. In *OECD Documents: Proceedings of a conference*. *Science and Mathematics Education in the United States: Eight innovations* (pp. 117-142). France: OECD.

Zucker, A. A., Shields, P. M., Adelman, N., & Powell, J. (1995). *Evaluation of NSF's Statewide Systemic Initiatives (SSI) program: Second year report, Cross-cutting themes*. Menlo Park, CA: SRI International.

Glossary

Assessment Activity - a specific sample of questions that will elicit a sample of student behavior.

Assessment Framework - a list of expectations that will be assessed.

Assessment Instrument - a purposeful collection of assessment activities intended to measure one or more concepts or a range of knowledge.

Assessment Report - the summary (numerical or qualitative) of student performance on an assessment instrument. May be reported at the student, classroom, school, district, state, or other levels.

Assessment Specifications - specific aspects, limits, and boundary conditions on the domain of knowledge being assessed to guide the selection and development of assessment activities.

Assessment Standards - a series of criteria for judging the adequacy of an assessment of student achievement.

Benchmarks - more specific levels of desired performance at various grade levels or ranges of grades. Benchmarks also is used as another term for performance standards, or more rarely, content standards.

Blueprint - an overall description of the assessment to be created. This will include what will be assessed, how the assessment will be created and validated, anticipated reports of results, and anticipated uses of results. The document may also have sample assessment exercises and sample scoring rubrics.

Curriculum Framework - a document that specifies what students should know and to be able to do for grade levels or grade ranges and the organization of the curriculum and instruction.

Performance Standard - defines levels of quality of student performance on an overall assessment instrument. Typically, two or three levels of desired overall performance is defined.

Scoring Rubric - a set of rules for assigning different levels of performance, ordered by quality, to student responses on open-ended assessment activities.

Appendix

Task Force Participants

Andrew Porter, Task Force chair, Director, Wisconsin Center for Education Research
C. Averett, Student Assessment, North Carolina Department of Education
Rolf Blank, Education Indicators, Council of Chief State School Officers
Joyce Krumtinger, Mathematics Education, Illinois Department of Education
Mozell Lang, Science, Michigan Department of Education
Donna Long, Mathematics Education, Indiana Department of Education
Megan Martin, Science Education/Assessment, California Department of Education
Senta Raizen, Director, National Center for Improving Science Education
Doris Redfield, Student Assessment, Virginia Department of Education
Ed Reidy, Deputy Commissioner, Kentucky Department of Education
Ed Roeber, Student Assessment, Council of Chief State School Officers
Walter Secada, Mathematics Education, Wisconsin Center for Education Research
Sharif Shakrani, Assessment Division, National Center for Education Statistics
Linda Sinclair, Science Education, South Carolina Department of Education
Larry Suter, Program Officer, National Science Foundation
William Tate, Mathematics Education, Wisconsin Center for Education Research
Roger Trent, Student Assessment, Ohio Department of Education Research
Norman Webb, Mathematics Education/Assessment, Wisconsin Center for Education Research
David Wiley, Education and Social Policy, Northwestern University