Curriculum Assessment

Andrew C. Porter
Vanderbilt University
June, 2004

Curriculum can be divided into the intended, enacted, assessed, and learned curricula. For K-12 education, the intended curriculum is captured most explicitly in state content standards—statements of what every student must know and be able to do by some specified point in time. The enacted curriculum refers to instruction (e.g. what happens in classrooms). The assessed curriculum refers to student achievement tests. States, districts, and the U.S. government test various subjects at various grade levels. Teachers use their own tests to monitor student performance.

In what follows, curriculum assessment is taken to mean measuring the academic content of the intended, enacted, and assessed curricula as well as the content similarities and differences among them. Pedagogy (i.e. how content is taught), while important to explain student learning, is not addressed in this chapter. Further, measuring the learned curriculum is a separate topic in its own right and is not considered here. Clearly, students can and do learn both more and less than the content of the assessed curriculum.

Knowing the content of the enacted curriculum is important because what students are taught is a powerful predictor of student achievement on a test (Gamoran, Porter, Smithson, & White, 1997; McKnight et al., 1987; Rowan, 1998; Schmidt, 1983a, 1983b; Sebring, 1987; Walberg & Schanahan, 1983), and helps explain a portion of the achievement gap between White, Black, and Hispanic students (Porter, 2003). Knowing the content of the intended curriculum is important because the intended curriculum is the content target for the enacted curriculum. Knowing the content of the assessed curriculum is important because student

achievement is measured only for the content assessed. Students may learn important content not on the test and that learning may go unidentified.

There are many important questions for research and practice that can only be answered through curriculum assessment of content. Do teachers teach what is tested? Do teachers teach what is in the textbook? Is the content of the textbook the same as the content of the test? Does the content of what is tested match well the content of the intended curriculum? Is standards-based reform working? Is the content of the enacted curriculum (what is taught) coming into an increasingly better match with the content of the intended curriculum?

In what follows attention is first given to defining content. Having defined content, attention is given to assessing the content of the intended, enacted, and assessed curricula. Where alternative approaches exist, attention is given to each. Having assessments of the content of the intended, enacted and assessed curricula, consideration is given to defining and measuring alignment among them. Next, evidence of the quality of data from the various procedures is summarized. The chapter closes with a section on potential uses of curriculum assessment data in both education research and practice.

But What Is Content?

Is content math, science, English language arts, or social studies? Is knowing just the subject specific enough? What kind of mathematics? Is the content arithmetic, measurement, algebra, geometry, or statistics? Even these are general areas to which entire courses are devoted. What kind of algebra? Is the content functions, matrices, or linear equations? And what are students supposed to know and be able to do in reference to linear equations? Should they be able to distinguish a linear equation from a non-linear equation? Should they be able to solve a linear equation? Should they be able to use a linear equation to solve a story problem? Clearly,

curriculum assessment requires important decisions about what is and is not content as well as how fine grained or precise the distinctions among types of content need to be. To some extent, the best definition of content can be decided on only within the context of a purpose. Even then, determining the best definition requires empirical investigation.

For many in education, academic content is defined as the topics that might or might not be taught, tested or included in content standards. Unfortunately, defining content in terms of topics has proven to be insufficient at least if explaining variance in student achievement is the goal (Gamoran et al, 1997). For example, knowing whether or not a teacher has taught linear equations while providing some useful information, is insufficient. What about linear equations was taught? Were students taught to distinguish a linear equation from a non-linear equation? Were students taught that a linear equation represents a unique line in a two space and how to graph the line? For every topic, content can further be defined according to categories of cognitive demand. In mathematics cognitive demand might distinguish memorize; perform procedures; communicate understanding; solve non-routing problems; conjecture, generalize, prove. In English language arts and reading, topics might be phoneme blending or suffixes, prefixes and root words and cognitive demand might be recall, demonstrate/explain, analyze/investigate, evaluate, generate/create.

"Languages" can and have been developed to define the content of academic subjects. The content language for an academic subject should be exhaustive in its inclusion of all possible types of content, and it should be common in the sense that the same language is used across studies and purposes (Porter, 2002; Schmidt et al., 2001). The terms used in the language to describe types of content should have the same meaning across people and time. The language should be reform-neutral in the sense that the content of any particular curriculum reform can be

well described using one common language. In short the language should be capable of describing the content of any intended, enacted or assessed curriculum.

Content languages have been developed to describe the content of mathematics, science, and English language arts (see http://www.ncrel.org/sec/). The languages are two-dimensional and can be presented in a rectangular matrix with *topics* as rows and *cognitive demands* (sometimes called *performance goals* or *performance expectations*) as columns. A language might have 60 to 100 topics (rows) and 3 to as many as 10 levels of cognitive demand (columns). Content is defined at the intersection of a particular topic (row) and a particular cognitive demand (column)—the cells of the matrix. For example, content might be to use a linear equation (the topic) to solve a novel problem (the cognitive demand).

When defining content (i.e. building a common content language), a number of decisions must be made. What topics and how many topics make up the rows of the matrix? What and how many levels of cognitive demand make up the columns? There is a tension between the desire for fine levels of measurement, on the one hand, and the difficulties involved in making such precise measures on the other. To the extent possible, the words used to describe topics and cognitive demand should be clear and have a single common meaning.

Assessing the Intended Curriculum

Berger, Desimone, Herman, Garet, & Margolin (2002) at the American Institute for Research (AIR) may have been the first to assess the content of content standards. The research team used subject experts to code each unit of the standards documents on topics and cognitive demands. For content analysis of content standards, the most successful approach has been to pick the most specific version of the content standards and analyze the content of each objective, paragraph, or phrase. Each specific part of a content standard documented is given a weight of

1.0. At AIR, three raters independently content analyzed each content standard document. When an objective was judged to represent content in multiple cells of the content language, the weight of 1.0 for that objective was spread evenly across the appropriate cells (e.g. 1/3 for each cell for an objective judged as best represented by three cells in the matrix). Data are averaged cell x cell across experts. Proportions are created for each cell by dividing by the sum of the average weights across all cells in the matrix. The cell proportions sum to 1.0 across rows and columns and indicate the extent to which the content represented by a cell is emphasized in the content standards document. The Council of Chief State School Officers—in collaboration with the Wisconsin Center for Education Research and The North Central Regional Education Laboratory—also has been working with a number of different states to complete content analyses of content standards in math, science, and English language arts (http://www.ccsso.org/projects/surveys_of_enacted_curriculum).

Data representation can proceed in several ways. One of the most powerful displays is in terms of topographical maps that can be created using a variety of charting software, including Excel (Porter, 2002). On the maps, what ordinarily is thought of as north and south represents topics and east and west represents cognitive demand. Shading represents relative content emphasis (percent of total coverage in the legends of Figs. 1 and 2) and is analogous to attitude on a topographical map. The maps clearly show not only what content is emphasized in the content standards, but what content is not (see Figure 1). Data can be displayed at a coarse-grain size such as in Figure 1, or one can look more closely at a particular type of content—for example, number sense and numeration—to get a finer grain picture within that area of content (see Figure 2). Visual comparisons can be made by placing topographical maps next to one another. Figure 1 shows that, at the coarse-grain level, not much content is excluded from the

standards. There are several similarities across states, but some potentially important differences as well. For example, State E places heavy emphasis on solving routine problems involving data analysis and probability. Looking at Figure 2, more differences are revealed by the fine-grained analysis. Here, National Council of Teachers of Mathematics (NCTM) standards are seen as unique in their emphasis on communicating understanding of combinations and permutations.

There is one disturbing property of the maps: topics are interpreted as points on an underlying continuous scale, as are categories of cognitive demand. Of course, there is no underlying continuum for either; each is a nominal scale variable. Thus, while each of the points of intersection between a particular topic or area of topics and a particular cognitive demand is accurately represented in the map, the areas between the points of intersection are meaningless. The charting software has simply smoothed the data as though there were an underlying continuum. Bar graphs can be also used to present the same data, with a bar at each point of intersection between cognitive demand and topic. In experimenting with this approach, however, the results are busy and difficult to interpret. Communication seems to be easier and more accurate using the topographical maps (Porter, 1998).

Others have taken similar approaches to assessing the content of the intended curriculum. All have their origins in the work of the Content Determinants Group at the Institute for Research on Teaching at Michigan State University (Porter et al, 1988) and all are quite similar one to another. Bill Schmidt and his colleagues (Schmidt et al., 2001) used similar procedures to content analyze mathematics and science textbooks from several different countries. Textbooks were divided into lessons; lessons were further divided into blocks (e.g., narrative-related, graphic, exercise). Blocks were then content analyzed by independent experts, and averages were taken. Textbook content was interpreted as each country's intended curriculum. The United

States was found to have a curriculum that covers more content in less depth than higher achieving countries.

Freeman, Kuhs, Porter, Floden, Schmidt, and Schwille (1983) content analyzed elementary school mathematics textbooks and tests. Analyses were limited to the items in the student exercise portions of each lesson. By using a common language to content analyze both tests and textbooks, comparisons were made among the tests, among the textbooks, and between textbooks and tests. Surprisingly, not only was some emphasized content in the textbooks not tested, but some content tested was not covered by the textbooks.

Assessing the Enacted Curriculum

Teachers make content decisions about a) how much time to spend on a school subject, b) what content to cover within that time, c) which students to teach what content, and d) to what standards of achievement (Schwille et al., 1983). In elementary school, for example, is mathematics taught every day, and for how long? By mid-year, one teacher may have spent as much time teaching mathematics as another will have spent by the end of the school year (Porter et al., 1988). In high school, students take courses that meet for a specified period of time for a specified number of days. Some high-school teachers initiate instruction immediately at the beginning of each class, while others spend considerable time getting students to sit down and pay attention (Porter et al., 1993).

Time is only one dimension of content. For example, in first-year algebra, do all teachers teach the same content? For a variety of reasons, the answer is no (Porter, 1998). Teachers may teach what they believe is most important, what they think the students are ready to learn, or what is most enjoyable and easy to teach. There are many factors that can and do influence teacher decisions about what to teach.

Content can differ from one student to the next, even within a class. In elementary school, instruction is sometimes individualized, with students moving at their own pace through a set of tested objectives. Other times, students are grouped according to estimates of ability, with different content taught to different groups.

What about standards of achievement? Teachers monitor student achievement and make decisions about pacing based on their assessments of student achievement. Is new content introduced only as old content is mastered, and if so, mastered by all of the students or some fraction of the students? Pacing decisions are important determinants of the content of the enacted curriculum because a slow pace covers less content. Teachers must negotiate between how much content they would like students to learn and how much content students can learn within the constraints of time, pedagogy, and effort.

Assessing the content of the enacted curriculum is substantially more challenging than assessing the content of the intended curriculum. If one is interested in the content of instruction for only a handful of days, then classroom observations using a common content language are a possibility. But generally, interest is in longer periods of time. What is the content of a student's elementary school experience or, at least, what is the content of a student's experience in a particular course or for a particular school year? For longer periods of time such as these, observations are not feasible.

An alternative approach is to use teacher self-report surveys that can be completed daily (these are typically called *logs*) or retrospectively over longer periods of time (e.g., a week, a month, a semester, or even a full school year). Two competing goods are at play in deciding how frequently to have teachers report on the content of their instruction. The more frequent the reporting, the less burden on a teacher's recall—but the greater the burden on a teacher's time.

Teacher logs have been used to study teacher content decisions in mathematics at the elementary school level (Porter, 1989) and to study the degree to which the content of instruction differs among high-school math and science courses with the same name (e.g., is the content of first-year algebra the same or different from class to class?) (Porter, 1989). As for assessing the content of the intended curriculum, assessing the content of the enacted curriculum begins with a definition of content (e.g. content language). Surveys are designed to record teacher reports of the content of their instruction, for example indicating what topics they taught, for how much time, and for each topic taught, what level of cognitive demand with what emphasis. These responses can be translated into proportions of content emphasis for each cell in the topics-by-cognitive-demand matrix much as described for content analyses of content standards in the previous section on assessing the intended curriculum. Data can be analyzed and displayed as was described for the intended curriculum (Porter & Smithson, 2001).

Over time, procedures for teacher logs have become progressively more structured to facilitate ease of data collection and analysis. In one study, teachers were asked each day to identify up to five areas of mathematics content (topic by cognitive demand) taught that day and with what emphasis (Porter, 1989). Rather than having teachers describe the content for each student in the class, three target students were selected: one believed to be at the 80^{th} percentile, one at the 50^{th}, and one at the 20^{th} percentile of within-class mathematics aptitude. Teachers were provided a catalog of 288 mathematics topics that might be taught. Logs were collected weekly and edited for ambiguities.

Brian Rowan and his colleagues (Rowan, Camburn, & Correnti, 2003; Rowan, Harrison, & Hayes, 2003) developed teacher logs for use in studying elementary school reading, writing, and mathematics (see www.sii.soe.umich.edu). Their logs ask teachers to report on what content

was covered, what students did, what materials were used, and how the teacher interacted with the student. In Rowan's work less emphasis was put on assessing content (the focus of this chapter) and more emphasis was placed on assessing pedagogy. Each log contains 150 questions. A branching strategy is used, with teachers first reporting emphasis placed on each of several curriculum strands. More extensive follow-up questions on pedagogy and content are asked for three focal strands (i.e., word analysis, comprehension, and writing). The Rowan content language does not define content at the intersection of topics and cognitive demand, although a difficulty scale has been formed within strands that may be related to the distinction among levels of cognitive demand.

In the Rowan study, days are sampled with a block of days from the spring, a block from the winter, and a block from the fall for a total of 90 days out of the school year (typically there are 180 days in a school year). Students also are sampled. In each class, eight focus students are randomly selected; the teacher completes a log for a different student each day. The motivation for sampling is to ease teacher burden; when sampling is not random, however, bias can result.

Teacher logs can provide excellent information on the content of the enacted curriculum, but they are expensive and burdensome. Teachers must be recruited to the task of completing the logs. Given sufficient incentives, a high percentage of target teachers can be recruited. But if the study involves a national probability sample of teachers, recruiting a high percentage of the sampled teachers to the task of completing daily logs may be difficult. An alternative is to use teacher self-report surveys, as in logs, but to ask teachers to report less frequently. End-of-year surveys were used in a study of curriculum reform in high-school mathematics and science (Porter et al., 1993). End-of-semester surveys were used in a study of upgrading in high-school mathematics (Gamoran, Porter, Smithson, & White, 1997). For these studies, teachers were

asked to report on the content of their instruction for an extended period of time—a semester or a school year. Days were not sampled and response burden was substantially less than for daily logs. The trade-off was that by surveying teachers less frequently than daily, the challenge to a teacher to accurately remember and report their content teaching practices was greater.

For assessing the content of the enacted curriculum, direct observation is not feasible except in rare cases where interest is in only a handful of instructional periods. One such case is to validate teacher self report (about which more is said in a following section on data quality). The only real distinction between logs and surveys is frequency of data collection; all logs are themselves surveys.

Assessing the Assessed Curriculum

The procedures used to assess the intended curriculum work at least as well for assessing the assessed curriculum. Once again the procedure begins with a common content language. Experts are recruited and trained to use the language and independently perform content analyses. Each item on a test is content analyzed by inferring the content (topic by cognitive demand) required to correctly answer the item. The main challenge to assessing the content of the assessed curriculum is accurately inferring from reading a test item how students will approach that item. In fact, students may differ one from another in the approach they take to answering an item. For example, in mathematics one student may have seen many problems virtually identical to a story problem on a test. For them, the item represents the cognitive demand of performing procedures (similar to a computational problem). Another student may have little familiarity with story problems of the type represented by the item. For them, the item represents the cognitive demand of solve a novel problem. Experts doing the content analyses must make a judgment as to the most likely approach that students will take to answering the

item. As seen in the following section on quality of data, experts make their judgments independently in ways that produce high inter-expert agreement. Still, careful empirical work involving students using think-aloud protocols can be used to investigate the modal student approach and differences among students.

Item score points are evenly distributed across the range of content tested by an item. An average is taken across content analyzers. Using total test score points as the base, proportions of content emphasis are calculated for the cells of the topics-by–cognitive-demand content matrix. Data can be analyzed and displayed in the same way as described when measuring the intended curriculum.

What Is Alignment and How Might It Be Measured?

Once the enacted, intended, and assessed curricula have been assessed, questions can be asked about the extent to which content is similar across them. To the extent content is the same, they are said to be aligned. For example, one might ask to what extent a student achievement test is aligned with a state's content standards. In fact, the No Child Left Behind Act of 2001 (NCLB) requires that each state align assessments to content standards. If the content assessed is exactly the same as the content represented in the standards, alignment is perfect. There are two ways in which alignment can be less than perfect: Content in the standards may not be assessed, and content assessed may not be in the standards.

Figure 3 represents various types of alignment that might be measured. The learned curriculum (achieved) is at the top, with the enacted curriculum (instruction) below. At the district, state, and national levels there is both the assessed curriculum (assessment) and the intended curriculum (standards). Within a level of the school hierarchy—at the state level, for example—one can ask questions about horizontal alignment. Is the state test aligned to the state

standards? Questions also can be asked about alignment across levels of the school hierarchy (i.e. vertical alignment). Is the district test aligned to the state test, or are the district standards aligned to the state standards? Vertical alignment can also explain an important aspect of opportunity to learn. The extent to which the content of the enacted curriculum a student experiences is aligned to the content of the test a student takes (assessed curriculum), the student can be thought of as having had an opportunity to learn (Porter, 1995).

Researchers in education have been interested in one type of alignment or another for decades (Cohen, 1995; Freeman et al., 1983), but recently interest has heightened. In part, this increased interest in alignment is due to findings that the content of the enacted curriculum is a strong predictor of gains in student achievement (Gamoran et al, 1997). Further, the U. S. Department of Education has been monitoring compliance of NCLB's requirement that states have assessments aligned to their content standards. States themselves have sought to provide increasingly better information about the degree to which their assessments are aligned to their content standards. Several reviews of different approaches to measuring alignment have appeared (Ananda, 2003; Bhola, Impara, & Buckendahl, 2003; CCSSO, 2002; Olson, 2003; Rothman, 2003).

One method for measuring alignment of assessments to content standards was developed by Norman Webb (1997, 2002). His procedure has been used for studies of alignment in language arts, mathematics, social studies, and science at the elementary, middle-, and high-school levels. The procedure involves experts' judgments on four criteria related to content agreement between assessments and standards: 1) categorical congruence, 2) depth of knowledge consistency, 3) range of knowledge correspondence, and 4) balance of representation. Webb provides no single overall composite measure of degree of alignment.

According to Webb, categorical congruence is met if there are at least six items measuring the topics represented by a standard. Depth of knowledge consistency asks experts to judge whether items in the assessment are as demanding cognitively as what the students are expected to know and do as stated in the standards. At least half the items corresponding to an objective (within a standard) have to be at or above the level of knowledge of the objective. There are four ordered levels of depth of knowledge. In mathematics, for example, Level 1 is recall, Level 2 is skills/concept, Level 3 is strategic thinking, and Level 4 is extended thinking. Ordered levels of depth of knowledge assumes that when a student demonstrates achievement on a higher "depth of knowledge" they necessarily have achieved on all lower depths of knowledge. Range of knowledge correspondence is met if at least half the objectives for a standard are measured by at least one assessment item. Balance of representation indicates the degree to which one objective within a standard is given more emphasis on the assessment than another. Webb's index of balance of representation is a function of only those objectives for a standard that had one or more items assessing the objective; the index ranges from 0 to 1.0.

Webb's procedures for assessing the alignment of a student achievement test to a state's content standards have been adapted by others. For example, researchers at the Buros Center on Testing used a modified Webb procedure to determine if items from commercially available tests match the Nebraska standards (Impara, 2001; Plake, Buckendahl, and Impara, 2001). Their criterion for alignment was a function of not only an item's assessed level of alignment to the standard, but also of agreement among the teachers doing the content analyses. Herman, Webb, and Zuniga (2002) convened panels to judge alignment between California's Golden State Examination in high-school mathematics and the University of California statement on competencies in mathematics for high-school graduates applying to the university. They did not

use Webb's balance and range criteria, but did add a criterion of *source of challenge* indicating whether an item was difficult not because of its content, but because of some confusing aspect of the item that was irrelevant to the content being assessed.

Karen Wixson et al. (2002) used a modified version of the Webb method to assess the alignment of achievement tests in four states to content standards in elementary reading. Wixson and colleagues dropped Webb's categorical congruence criterion and added *coverage* to indicate the extent to which objectives are represented by at least one assessment item. They also added a criterion, *structure of knowledge comparability*, to indicate the extent to which the philosophy of the standards matched the philosophy underlying the assessments.

There are several aspects of Webb's measure of alignment that are important. First, a set of content standards is a given, and Webb measures how a particular assessment aligns to those content standards. Second, for each of the four dimensions of alignment, Webb sets a criterion for how much alignment is enough. Third, alignment is a function of content, and content is defined by topics and cognitive demand. Cognitive demand is considered as an ordinal scale such that if an item has equal or higher cognitive demand than the standard, alignment is judged to be present. In all cases, content standards dictate the level of detail at which topics and cognitive demands are defined. The more general the standards, the less precise the distinction among types of content. Webb's procedures have been used by many states.

Another frequently used measure of the alignment between assessments and content standards was developed by Porter and his colleagues (Porter & Smithson, 2001; see http://www.ccsso.org/projects/surveys_of_enacted_curriculum). The procedure begins by completing content analyses of a state's content standards and (separately) a state's assessment of student achievement (as described in earlier sections on assessing the intended curriculum and

the assessed curriculum). Again, the basic data are proportions of content emphasis in each cell of the topics-by-cognitive-demand matrix. (See Figure 4 as an illustration.) Alignment is perfect if the proportions for the assessment match cell by cell the proportions for the standards. A conceptually straightforward index of alignment is defined

$$\text{Alignment} = 1.0 - \frac{\sum |x - y|}{2}$$

where x = assessment cell proportions and y = standards cell proportions. The index ranges from 0 (no alignment at all) to 1.0 (perfect alignment). Starting with matrices of proportions, there are other ways that an alignment index can be calculated. One example is the correlation of proportions of content emphasis across cells between the assessment and the standard. For one data set, Porter (2002) found that these two indices of alignment correlate .86. Other quantitative measures of alignment using the basic content-emphasis proportions are possible (Porter, 2002). The alignment measure offered here has been used to describe alignment in many states in the subjects of mathematic, science, and reading/language arts.

Whichever quantitative index of alignment is used, several important properties exist. First, the basic data are a function of a common content language. Thus, all measures of alignment are a function of the same levels of detail in distinctions among topics and cognitive demand. Second, because measuring alignment begins with content analyses using one common content language, once a state's content standards have been content analyzed, data can be used not only to measure the alignment of one state's standards with that state's test, but with other tests and standards as well. Further, if a state should change its test but not its content standards, then the standards do not need to be reanalyzed—only the new test needs to be analyzed.

Third and perhaps most importantly, the procedure is not limited to measuring the alignment of tests to standards. Anything having content that can be described using the common

content language can be aligned with anything else having content that can be described using the common content language. The alignment among content standards across states can be measured. Similarly, alignment across states can be measured for instruction or tests. The procedure is not limited to measures of horizontal alignment; the alignment of instruction to content standards can be measured to assess opportunity to learn. Alignment between instruction and standards can be measured at the individual teacher level, and variance across teachers can be studied. Fourth, the alignment index is symmetric. For example, degree of alignment is a function of both content that is tested but not in the standards, and content that is in the standards but not tested.

What the index does not indicate is how much alignment is enough. Is an alignment of .4 between a state's assessment and its content standards sufficient, or should it be higher? A reasonable but somewhat arbitrary criterion for how much alignment, is enough could be set as Webb has done for each of his four dimensions of alignment but there is no absolute criterion for alignment. Instead, alignment has been judged comparatively. For example, is alignment of a state's test to that state's standards higher than the alignment of the state's test to other state standards? The answer should be yes at least according to NCLB. When a state's assessment is no more aligned to that state's content standards than it is to other states' content standards, alignment probably needs to be improved. At the same time, alignment of a test with content standards for only one form of the test should not be perfect. One form of a test is a sample of the content in the standards. If the state uses a different form of the test each year, as it should, then content analyses could be conducted across multiple forms of the test. As the number of forms increases, the total number of items increases. At a point when sufficient numbers of items are present so that the sample becomes close to being the population, the test should be perfectly

aligned to the content standards. But the specificity of the content standards also limits the degree of alignment. The more vague and general the content standards, the less perfect alignment can be.

Use of a common content language allows for some powerful data displays. For example, if one has content analyzed standards and assessments from each of several states and perhaps national professional standards, the results can be displayed in a standards-by-assessment matrix of alignment values. An example comes from *Moving Standards to the Classroom: The Impact of Goals 2000 Systemic Reform on Instruction* (Berger, Desimone, Herman, Garet, & Margolin, 2002). Table 1 shows the results for four states and the NCTM content standards (Porter, 2002). The main diagonal of the matrix (underlined) reports alignment of each state's test with that same state's content standards. Presumably, a state's assessment would be more highly aligned with its own content standards than with other states' content standards. For these states, however, within-state alignment—on average, .4—is no higher than between state alignment (.39). What might account for the result? One possibility is that state standards are not sufficiently specific to allow an assessment to be tightly aligned. Another possibility is that states need to work harder on getting their assessments aligned to their standards. Of course, if all state content standards were perfectly aligned with one another, there could be no difference in the degree of alignment within state versus between state. But the state content standards in Table 1 are far from perfectly aligned.

Achieve Incorporated, founded in 1997 jointly by state governors and chief executive officers of major corporations, also developed a method for measuring alignment between assessments and content standards (Rothman, Slattery, Vranek, and Resnick 2002). The procedures have been used by several states in mathematics and English language arts.

Achieve measures alignment on six dimensions: 1) content centrality—assesses the match of an item to a standard using four levels of degree of match; 2) performance centrality—matches an item's level of cognitive demand to the standard's level of cognitive demand using four levels to indicate the degree of match; 3) source of challenge—assesses whether item difficulty is primarily a function of content or of some aspect of the item that is irrelevant to the content being assessed; 4) level of challenge—looks to see if as a set the items matches the span of difficulty in cognitive demand found in the content standards (described in narrative); 5) balance—looks at the set of items to judge the balance in content emphasis against the standards (described in narrative); and 6) range—looks at the set of items to judge the extent to which the full amount of content in the standards is represented in the assessment by at least one item.

Achieve uses a team of teachers, curriculum specialists, and subject matter experts to do the content analyses. Content analyses are completed by the group rather than by experts individually.

Project 2061 at the American Association for the Advancement of Science has designed procedures to critique textbooks in mathematics and science from the perspective of alignment with selected standards (Kesidou & Roseman, 2002; see http://www.project2061.org). Again, experts are used to do analyses. Two independent two-member review teams are created; each team is comprised of one experienced teacher and one education researcher. Rather than average results across the two teams, any differences are reconciled by Project 2061 staff. Training of experts to do the analyses is extensive and involves one full week—much greater than the training for any of the procedures previously described.

The Project 2061 procedures consider the alignment between content in the textbook and content in the standards. The procedures also consider alignment between the pedagogical

strategies of the textbook and those implied in the standards. For instruction, they consider 1) providing a sense of purpose, 2) taking account of students' ideas, 3) engaging students with phenomena, 4) developing the use of scientific ideas, 5) promoting student reflection, 6) assessing progress, and 7) enhancing the learning environment. The procedure also assesses the content accuracy of the textbooks. Clearly, the 2061 procedures reach well beyond asking about alignment of content between textbooks and content standards. Even when assessing alignment of content, they not only ask about match in terms of topic by cognitive demand, but also in terms of content coherence and accuracy. Under "coherence," for example, they ask about connections among the ideas treated.

Quality of Data

In curriculum assessment, various approaches have been used to assess the intended, enacted, and assessed curricula and to measure the nature and degree of alignment along them. While somewhat different approaches have been used over the years and by different researchers, there are a number of important common issues concerning the quality of data.

Quality of Common Language

Several of the procedures start with a common language for describing content. Most of the languages are subject specific, and most consist of defining content at the intersection of topics and cognitive demand. Procedures using a common language are dependent upon the quality of the language on which they are based. Does the language make all of the content distinctions that are useful for the purposes to which the language will be used? Are topics described using terms that have common meaning across users of the language? Is the language exhaustive—in other words, does it include all of the content that might be intended, taught, and/or assessed for a particular academic subject at a particular grade level?

While within a subject and grade level languages have evolved over time, and while there are some differences among languages used by investigators, all of the languages have been created through a careful process of studying national professional content standards, state content standards, textbooks, and tests—and through an iterative process of review and revision by content experts including professors, professional educators, and teachers. Languages have sometimes been simplified in recognition of the burden they might otherwise create for respondents when used in a questionnaire, for observers when used in classroom observations, or for content analyzers when used to describe the content of a document. Such simplifications come at a price if they result in a language that glosses over important distinctions among types of content that are useful for the purposes to which the language is to be used. There is no way to ensure the perfect language. With each use, content that needs to be added can be identified, and distinctions among topics and/or among levels of cognitive demand can be sharpened or stated in a way that helps ensure a common meaning. At the same time, there is value in having a language that stays fixed over time. Results of one study can be compared more directly to results from another study, and content analyses of standards need not be redone when tests are revised and new studies of alignment are desired.

Expertise of Content Analysts

Measuring the content of the intended curriculum and/or the assessed curriculum invariably involves using experts to do the content analyses. Even experts need to be trained in the particular procedures (e.g., the use of the content language); thus, the quality of data is a function of the degree of expertise recruited and the quality of training conducted. To make the research replicable, the recruitment and training of experts needs to be carefully described.

Inter-expert Reliability

Most content analysis procedures use multiple experts who complete the content analyses independently. The resulting data are averaged across experts. Assessing the reliability of the summary data is an important part of the research. Reliability is underestimated by interrater reliability; generalizability theory needs to be used to estimate the reliability of the aggregate data. Even when the content language is defined at the intersection of 70-100 topics by five levels of cognitive demand, reliability of aggregate proportions of content emphasis tend to be quite good, not only for assessments, but even for the more challenging task of analyzing textbooks and content standards. In one study, for example, reliability of aggregate data was .7 across just two raters and .8 across four raters for both tests and content standards (Porter, 2002). Increasing the number of raters beyond four results in diminishing returns in terms of increased reliability and is probably not warranted. Often in a set of independent raters there is one expert who completes the task in ways that do not agree with the rest of the group. Eliminating the odd expert increases reliability and, hopefully, validity.

Some of the procedures used to measure alignment of a test to a content standard use experts who operate as a group, not as independent individuals. If one group is used, then no estimate of reliability is possible. To estimate reliability would require multiple groups, each operating independently. Generally, such reliability studies have not been conducted, and are needed.

Target Period of Time when Assessing the Enacted Curriculum

When measuring the enacted curriculum, quality of data issues are more and more complex. One of the decisions that must be made in a study of the content of the enacted curriculum is the time period to be described. While in theory the relevant period of time could be a day, a unit, a semester, a year, or even the elementary school experience, most studies have

focused on a school year. Perhaps the focus on a school year is because students often stay with one teacher for a full school year. Perhaps it is because studies of longer periods of time are extraordinarily difficult to conduct. Studies of shorter periods of time are often of substantially less value if one is interested in understanding gains in student achievement (although evaluating a particular unit of instruction would be an exception).

When a full school year is the focus, the task is to measure all of the content taught during that school year. Observing many classrooms every day for a school year is expensive. If one could take a representative sample of days, that would make observations more manageable. But how many days should be sampled, and should the sampling be random or systematic? Some research exists on the number of days that need to be sampled for classroom observations to yield reliable and valid results (Shavelson & Dempsey-Atwood, 1976). These studies have tended to focus on pedagogical practices that are used on many, if not most, days. Samples of 8 to 12 days are acceptable. The rarer the event, the larger the sample of days necessary. When measuring content at the fine-grain level of topics by cognitive demand, much of the content is fairly rare in the sense of being taught on many, if not most, days. In elementary school, for example, geometry might be taught every day but only for two weeks. Thus, samples of days need to be large and representative, beyond what would be possible for classroom observations.

Students to be Studied When Assessing the Enacted Curriculum

Another sampling issue concerns students within a classroom. Are all students taught the same content or are different students taught different content? The amount of content variability across students within a class is an empirical question. Different approaches have been taken to assess within-class variability; one is to have teachers describe the content of instruction for each of several target students each day (Porter, 1989). Another approach has been to have a teacher

describe the content of instruction for a different target student each day (Rowan, Camburn, & Correnti, 2003). The latter approach is less burdensome, but confounds day with student.

Frequency of Data Collection When Assessing the Enacted Curriculum

Teachers may be surveyed daily, once a week, once a month, once a semester, or once a year. An end-of-year survey requires a teacher to remember the content for the entire school year. A daily log asks the teacher to remember the content for that particular day. Certainly the quality of data from daily logs is better than the quality of data from end-of-year surveys, everything else being equal. Keeping everything else equal requires that logs be kept every day for the period of time under investigation. If logs are kept for a sample of days, then studies must be conducted to show that the sample results are accurate descriptions of the full year results.

Some studies have been conducted to address the issue of how well end-of-year survey data agrees with data collected through daily logs aggregated across the full school year (Porter et al., 1993). In turn, daily logs have been compared to data from classroom observations for selected days during which both data exist. Agreement between observations and logs was high, with correlations of .6 to .8. Agreement was lowest for cognitive demand, but in that study there were nine levels of cognitive demand rather than the more typical five. Agreement between daily logs aggregated to the full school year and end-of-year surveys was similar, with correlations generally in the range of .6 to .8. Agreement between logs and end-of-year surveys were lower for content rarely taught and content that was not the focus of the course under investigation.

The Possibility of Bias When Assessing the Enacted Curriculum

There is, of course, possibility for observer bias or respondent bias. When observers are used, they should be carefully trained to complete their observations in a way that is common across observers and valid to the content language being used. If the research involves

treatments, the observer should be blind as to who receives treatment and who does not. When surveys—including logs—are conducted, anonymity should be insured. It is important for teachers to understand that there are no high stakes consequences for them or their students in relation to their responses. Generally, surveys should ask questions that clearly indicate what is wanted in behavioral terms, not evaluative. How much and what type of content was taught should be asked, not whether the content taught what was intended, was good, or was aligned with the state content standards. To the extent possible, the survey should make clear the nature of the distinctions respondents are to make. For cognitive demand, it is useful to provide elaboration of each level of cognitive demand. In mathematics, for example, the cognitive demand "perform procedures, solve routine problems" can be elaborated as "do computations, make observations, take measurements, compare, and develop fluency." Unfortunately, such elaborations are only helpful when they do not greatly increase response burden.

A number of investigations of the validity of survey data for reporting instructional practice have been completed. Most of this work has been done in mathematics. The findings are that survey data are excellent for describing quantity—what content is taught, and for how long—but not good for describing quality (Burstein et al., 1995; Herman, Klein, and Abedi, 2000; Mayer, 1999; McCaffrey et al., 2001; Spillane and Zeuli, 1999).

Completeness of the Data Set When Assessing the Enacted Curriculum

Another dimension of the quality of measures of the enacted curriculum concerns the completeness of the data set. If surveys are used, what is the response rate? For those who respond, how complete are their responses across all of the questions asked? Generally, if surveys are carefully conducted, incentives provided, and follow-ups rigorously pursued,

response rates can exceed 75% (Porter, 2003). If the survey is well formatted and not too long, most respondents complete all of the questions.

Gains in Student Achievement as a Criterion for Validity

One criterion for judging the quality of the data for measures of the enacted curriculum is to use the data to predict gains in student achievement. If prediction is good, then the quality of the enacted curriculum data must be good as well. If the quality of prediction is not good, then interpretation of the results is ambiguous. Perhaps the delivery of the content was so poor that even though the content was taught, it was not learned. Generally, however, studies that have used measures of the content of the enacted curriculum to predict gains in student achievement have found strong correlations. For example, Gamoran et al. (1997) found that end-of-semester surveys aggregated to the school year were good predictors of gains in student achievement for first-year high-school mathematics (correlation of approximately .4). Importantly, when the data were collapsed to measure just cognitive demand or just topics, correlations dropped to near zero (Porter, 2002). At least, this study strongly suggests that simplifying a content language to either just cognitive demand or just topic would be a huge mistake if one is interested in collecting data that predict gains in student achievement.

What Are the Uses of Curriculum Assessment Data?

Some uses of curriculum assessment data were mentioned at the beginning of the chapter. Now, with a better understanding of the types of measures and data available, more thorough attention can be given to uses, not only in research but in the practice of education as well.

The Intended Curriculum

Bill Schmidt and his colleagues (Schmidt, 2001) used data from content analyses of textbooks to describe the intended curriculum for countries participating in the Third

International Mathematics and Science Study (TIMSS). They found that textbooks in the United States cover many more topics in less depth than those in higher achieving countries. With the current emphasis on state content standards, one could use content analyses to better understand what content is to be taught and what content is not to be taught. Content standards are linear—primarily text—and are not particularly analytic. Content analyses make the content messages of content standards clearer and easier to understand. Content analysis results could be used to better communicate the intentions of content standards to teachers. One finding has been that the standards are very broad and general (Porter, 2002). Perhaps the content languages could be used to build better and more focused content standards. A committee of experts might be convened, given the content language, and asked to reach consensus on what content (topics by cognitive demand) is to be taught and what content is not to be taught. To force focus, experts might be given an upper bound on the number of topics by cognitive demand that can be included in the content standards for a given subject at a given grade.

The Enacted Curriculum

The earliest work on assessing the enacted curriculum was done to create a dependent variable for use in research on teachers' content decisions (Porter et al., 1988). Rowan and his colleagues (Correnti, Rowan, & Camburn, 2003) use measures of the enacted curriculum as an intervening variable in their research on the effects on achievement of models of comprehensive whole school reform. Measures of the enacted curriculum can also be used to investigate the quality of implementation of a new curriculum (Porter et al., 1993). When states increased the number of math and science credits required for high-school graduation, some hypothesized that the influx of new and weaker students in courses would result in the watering down of course content. A study was conducted to see if the enacted curriculum in courses experiencing a large

influx of new students was different from courses where there was not such a large influx. The answer was generally no (Porter et al., 1993).

The Assessed Curriculum

When student achievement tests are constructed, items are written against test specifications detailing the content to be tested. The items on the test are to serve as a representative sample of the domain of content. Content standards specify the domain. Test construction requires careful attention to building test forms that are well aligned to the content standards. Where tests are used for student accountability the law requires that students have an adequate opportunity to learn the content tested; the enacted curriculum must be aligned to the assessed curriculum (Debra P. v. Turlington, 1981).

Measures of Alignment

As was stated earlier, the No Child Left Behind Act requires states receiving Title I funds to align their student achievement tests with their content standards. Alignment of instruction to textbooks can be used to study the effect of textbooks on the content of instruction (Freeman & Porter, 1989). Alignment of instruction to assessments can be used to study the effects of assessments on the content of instruction. Alignment of instruction to content standards can be used to assess the effects of standards-based reform (Porter & Smithson, 2001). If standards-based reform is successful, then over time instruction should become increasingly more aligned with content standards.

The alignment of instruction to assessments can be used as a control variable in research on teacher pedagogical practices. Clearly, what is taught is an important determinant of student achievement, but so are the pedagogical strategies describing how the content is taught. By using

the alignment of the content of instruction to the assessment as a control variable, research on pedagogical practices has greater precision.

Alignment can also be used to study the coherence of a state or district's policy system. Policies include content standards, assessments, professional development, and curriculum materials. Alignment results can be displayed in a policy-by-policy content matrix. The greater the coherence of the policy system, the larger the alignment values in the off-diagonal elements. Where alignment is low, adjustments can be made.

The above are just a few of the potentially many uses of curriculum assessment data. You may think of others. Perhaps your own research will use curriculum assessment data in new and powerful ways that shed light on the effects of education policies and practices and how they might be strengthened.

Reference List

Ananda, S. (2003). *Rethinking issues of alignment under No Child Left Behind.* San Francisco: WestEd.

Berger, A., Desimone, L., Herman, R., Garet, M., & Margolin, J. (2002). *Content of state standards and the alignment of state assessments with state standards.* Washington, DC: U.S. Department of Education.

Bhola, D.S., Impara, J.C., & Buckendahl, C.W. (2003). Aligning tests with states' content standards: Methods and issues. *Educational Measurement: Issues and Practice, 22*(3), 21-29.

Burstein, L., McDonnell, L., Van Winkle, J., Ormseth, T., Mirocha, J., & Guiton, G. (1995). *Validating national curriculum indicators.* Santa Monica, CA: RAND.

Cohen, S.A. (1995). Instructional alignment. In Lorin W. Anderson (Ed.), *International encyclopedia of teaching and teacher education* (2nd ed.). New York: Pergamon.

Correnti, R., Rowan, B., & Camburn, E. (2003, April). *School reform programs, literacy practices in 3rd grade classrooms, and instructional effects on student achievement: Notes from the first year of a study of instructional improvement.* Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Council of Chief State School Officers (2002, September). *Models for alignment analysis and assistance to states.* Washington, DC: Author.

Debra P. v. Turlington, 474 F. Supp 244 (M.D. Fla. 1979); affirmed in part 644 F. 2d 397 (5th Cir., 1981).

Freeman, D.J., Kuhs, T.M., Porter, A.C., Floden, R.E., Schmidt, W.H., & Schwille, J.R. (1983). Do textbooks and tests define a national curriculum in elementary school mathematics? *Elementary School Journal, 83,* 501-513. (Reprinted in *The Education Digest,* 1984, March, 47-49.)

Freeman, D.T., & Porter, A.C. (1989). Do textbooks dictate the content of mathematics instruction in elementary schools? *American Educational Research Journal 26*(3), 403-421. (Also Research Series No. 189, East Lansing: Michigan State University, Institute for Research on Teaching.)

Gamoran, A., Porter, A.C., Smithson, J., & White, P.A. (1997, Winter). Upgrading high school mathematics instruction: Improving learning opportunities for low-achieving, low-income youth. *Educational Evaluation and Policy Analysis, 19*(4), 325-338.

Herman, J.L., Klein, D.C.D., & Abedi, J. (2000). Assessing students' opportunity to learn: Teacher and student perspectives. *Educational Measurement: Issues and Practice, 19*(4), 16-24.

Herman, J.L., Webb, N., & Zuniga, S. (2002, April). *Alignment and college admissions: The match of expectations, assessments, and educator perspectives.* Paper presented at the annual meeting of the American Education Research Association, New Orleans, LA.

Impara, J.C. (2001, April). *Alignment: One element of an assessment's instructional utility.* Paper presented at the 2001 annual meeting of the National Council on Measurement in Education, Seattle, WA.

Kesidou, S., & Roseman, J.E. (2002). How well do middle school science programs measure up? Findings from Project 2061's curriculum review [Electronic version.]. *Journal of Research in Science Teaching, 39*(6), 522-549.

Mayer, D.P. (1999). Measuring instructional practice: Can policymakers trust survey data? *Educational Evaluation and Policy Analysis, 21*(1), 29-45.

McCaffrey, D.F., Hamilton, L.S., Stecher, B.M., Klein, S.P., Bugliari, D., & Robyn, A. (2001). Interactions among instructional practices, curriculum, and student achievement: The case of standards-based high school mathematics. *Journal for Research in Mathematics Education, 22*(5), 493-517.

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

Olson, L. (2003, Spring). Standards and tests: Keeping them aligned. *Research Points: Essential Information for Education Policy, 1*(1).

Plake, B.S., Buckendahl, C.W., & Impara, J.C. (2001, June). *A comparison of publishers' and teachers' perspectives on the alignment of norm-referenced tests to Nebraska's language-arts content standards.* Paper presented at the annual large-scale assessment conference of the council of chief state school officers, Snowbird, UT.

Porter, A.C. (1989). A curriculum out of balance: The case of elementary school mathematics. *Educational Researcher 18*(5), 9-15. (Also Research Series No. 191, East Lansing: Michigan State University, Institute for Research on Teaching.)

Porter, A.C. (1995). The uses and misuses of opportunity-to-learn standards. *Educational Researcher 24*(1), 21-27.

Porter, A.C. (1998). The effects of upgrading policies on high school mathematics and science. In D. Ravitch (Ed.), *Brookings papers on education policy 1998* (pp. 123-172). Washington, DC: Brookings Institution Press.

Porter, A.C. (1998, October). *Curriculum reform and measuring what is taught: Measuring the quality of education processes.* Paper presented at the annual meeting of the Association for Public Policy Analysis and Management, New York City, NY.

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

Porter, A.C. (2003, October). *Prospects for school reform and closing the achievement gap.* Paper presented at the Educational Testing Service Invitational Conference, New York City, NY.

Porter, A., Floden, R., Freeman, D., Schmidt, W., & Schille, J. (1988). Content determinants in elementary school mathematics. In D.A. Grouws & T.J. Cooney (Eds.), *Perspectives on research on effective mathematical teaching* (pp. 96-113). Hillsdale, NJ: Lawrence Erlbaum Associates. (Also Research Series 179, East Lansing: Michigan State University, Institute for Research on Teaching.)

Porter, A.C., Kirst, M.W., Osthoff, E.J., Smithson, J.S., & Schneider, S.A. (1993). *Reform up close: An analysis of high school mathematics and science classrooms.* (Final report to the National Science Foundation on Grant No. SPA-8953446 to the Consortium for Policy Research in Education.) Madison: University of Wisconsin-Madison, Wisconsin Center for Education Research.

Porter, A.C., & Smithson, J.L. (2001). Are content standards being implemented in the classroom? A methodology and some tentative answers. In S.H. Fuhrman (Ed.), *From the capitol to the classroom: Standards-based reform in the states—One hundredth yearbook of the National Society for the Study of Education, Part II* (pp. 60-80). Chicago: University of Chicago Press.

Rothman, R. (2003). *Imperfect matches: The alignment of standards and tests.* Manuscript in preparation, National Research Council.

Rothman, R., Slattery, J.B., Vranek, J.L., & Resnick, L.B. (2002). *Benchmarking and alignment of standards and testing.* (CSE technical report No. 566.) Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Rowan, B., Camburn, E., & Correnti, R. (2003). *Using teacher logs to measure the enacted curriculum in large-scale surveys: Insights from the study of instructional improvement.* Revised version of paper presented at the annual meeting of the American Educational Research Association, April 2002, New Orleans, LA.

Rowan, B., Harrison, D.M., & Hayes, A. (2003). *Using instructional logs to study elementary school mathematics: A close look at curriculum and teaching in the early grades.* Ann Arbor: University of Michigan, School of Education.

Schmidt, W., McKnight, C., Houang, R., Wang, H.C., Wiley, D., Cogan, L., et al. (2001). *Why schools matter: A cross-national comparison of curriculum and learning.* San Francisco: Jossey Bass.

Shavelson, R.J., & Dempsey-Atwood, N. (1976). Generalizability of measures of teaching behavior. *Review of Educational Research, 46,* 553-611.

Spillane, J.P., & Zeuli, J.S. (1999). Reform and teaching: Exploring patterns of practice in the context of national and state mathematics reforms. *Educational Evaluation and Policy Analysis, 21*(1), 1-27.

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

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

Wixson, K.K., Fisk, M.C., Dutro, E., & McDaniel, J. (2002). *The alignment of state standards and assessments in elementary reading.* CIERA report #3-024. Ann Arbor: University of Michigan, School of Education, Center for the Improvement of Early Reading Achievement.

Table 1

Alignment of Assessments With Standards: Seventh-Grade Math—Goals 2000 Study

Standard	Assessment			
State	B	D	E	F
B	.37	.39	.37	.45
D	.35	.37	.36	.40
E	.36	.33	.43	.31
F	.32	.35	.30	.41
NCTM	.34	.40	.33	.47

Note. Average within-state alignment = .40; average between-state alignment = .39; average state-test-to-NCTM alignment = .39.

Figure 1. Seventh-grade standards.

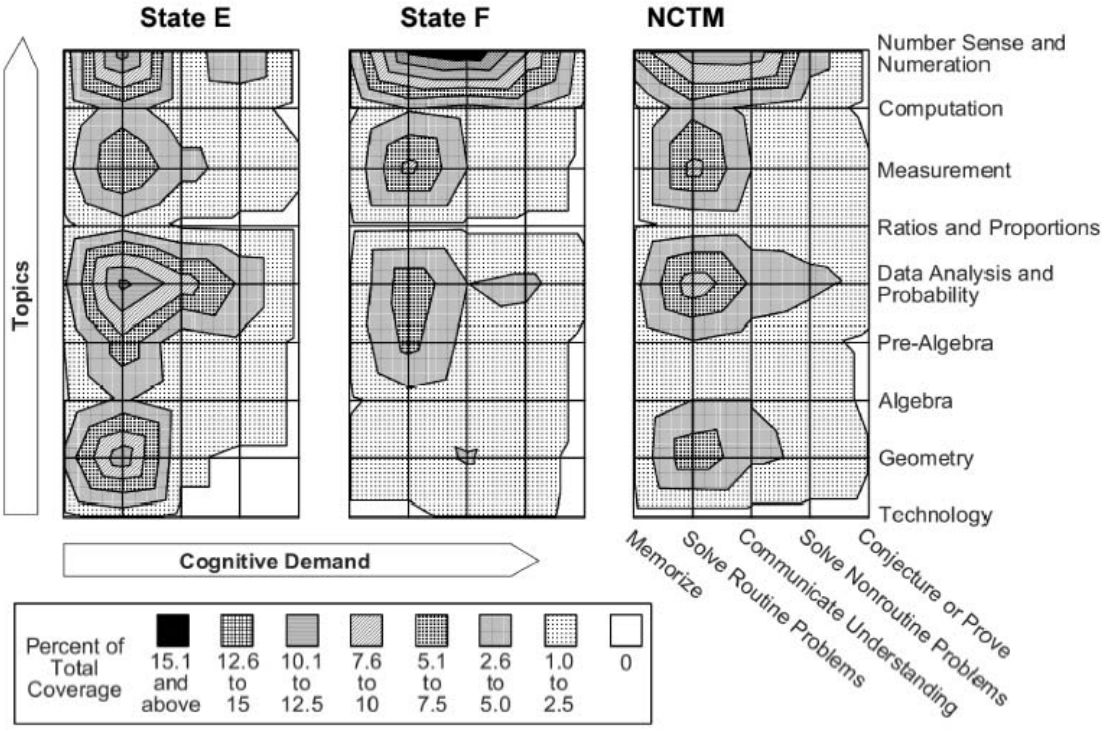

Figure 2. Seventh-grade standards: Close view.

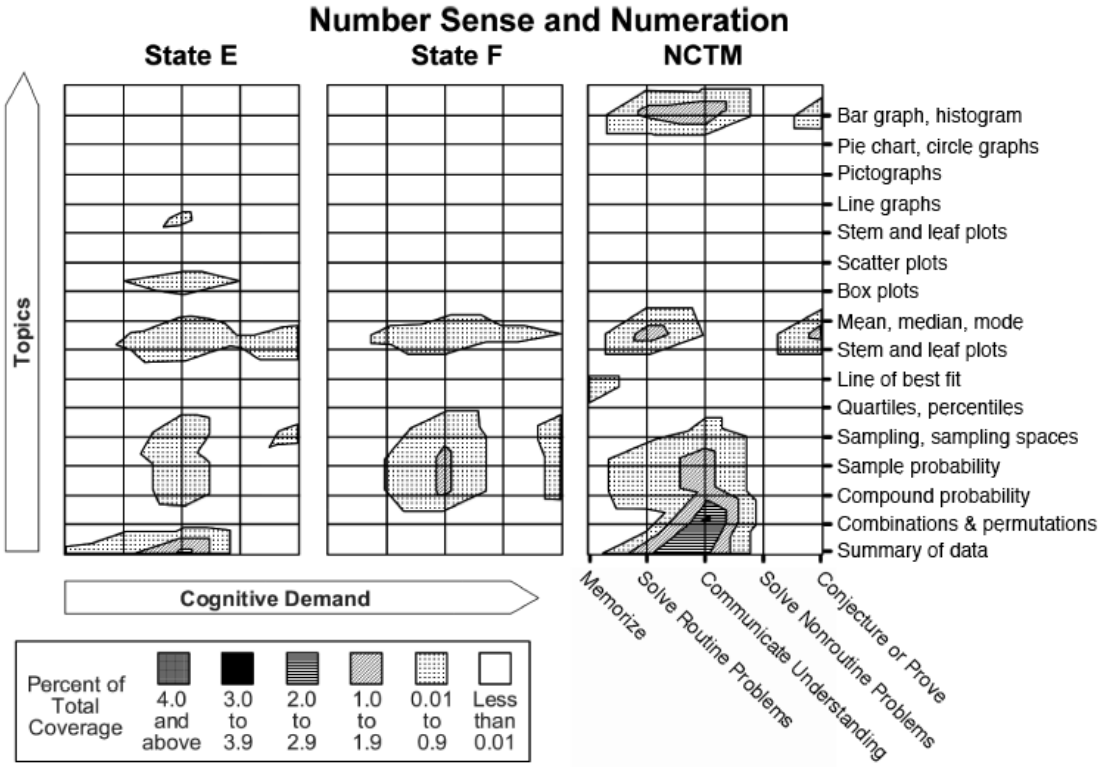

Figure 3. Vertical and horizontal alignment.

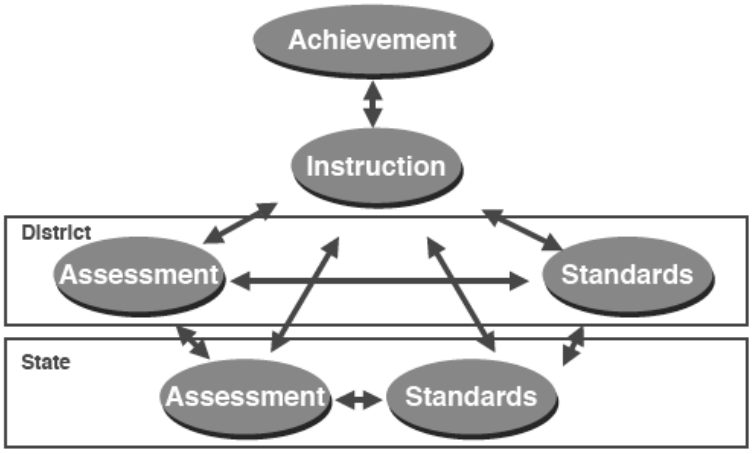

Figure 4. Example matrices to measure alignment.