Alignment as a Teacher Variable

Andrew C. Porter
Department of Leadership, Policy, and Organizations
Vanderbilt University

John Smithson
Wisconsin Center for Education Research
University of Wisconsin, Madison

Rolf Blank
Council of Chief State School Officers
Washington, District of Columbia

Timothy Zeidner
Department of Leadership, Policy, and Organizations
Vanderbilt University

With the exception of the procedures developed by Porter and colleagues (Porter, 2002), other methods of defining and measuring alignment are essentially limited to alignment between tests and standards. Porter's procedures have been generalized to investigating the alignment between content standards, tests, textbooks, and even classroom instruction as experienced by teachers and students. They have also led to the development of tools that hold promise for innovative uses in both research and the improvement of education practice. The purpose here is to define and illustrate innovative uses of two of the tools from Porter's work, content maps and a quantitative index of the degree of alignment. The intention is to put these tools in the hands of education researchers.

With the passage of the No Child Left Behind (NCLB) Act in 2001, the concept of alignment and how to measure alignment has taken center stage. The NCLB Act requires that states have challenging content standards for academic subjects and

---

Correspondence should be addressed to Andrew C. Porter, Department of Leadership, Policy, and Organizations, Vanderbilt University, Peabody No. 59, Nashville, TN 37203. E-mail: andy.porter@vanderbilt.edu

aligned tests of student achievement. Tests aligned to content standards is the law if a state is to receive Title I funds. However, the concept of alignment is not limited to student achievement tests and content standards, and it is not limited to state-level policies and procedures. As shown in Figure 1, there are issues of horizontal alignment (e.g., within a level of government) and vertical alignment (between levels of government and into schools and classrooms).

With the recent focus on alignment between student achievement tests and content standards, several new and promising procedures for defining and measuring alignment have been developed and used. Some are discussed in this special issue, and several reviews of different approaches to measuring alignment have appeared (Ananda, 2003; Bhola, Impara, & Buckendahl, 2003; Olson, 2003; Rothman, 2003). Perhaps the most frequently used approach was developed by Norman Webb (1997, 2002) and is discussed in this issue. Several people have modified Webb's procedures to measure alignment (Herman, Webb, & Zuniga, 2002; Impara, 2001; Plake, Buckendahl, & Impara, 2001; Wixson, Fisk, Dutro, & McDaniel, 2002). Achieve, Inc. developed a method for measuring alignment as well (Rothman, Slattery, Vranek, & Resnick, 2002), and Project 2061 at the American Association for the Advancement of Science designed and used procedures to investigate the alignment between textbooks and standards for both mathematics and science (Kesidou & Roseman, 2002; see http://www.project2061.org).

With the exception of the procedures developed by Porter and colleagues (Porter, 2002), the various methods of defining and measuring alignment are limited to alignment between tests and standards or, in the case of Project 2061, textbooks and standards. Porter's procedures have been generalized to investigating the alignment between content standards, tests, textbooks, and even classroom instruction as experienced by teachers and students. They have also led to the devel-
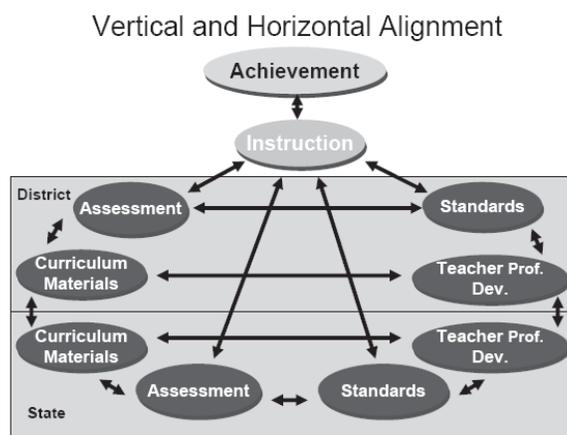

FIGURE 1   Vertical and horizontal alignment.

opment of tools that hold promise for innovative uses in both research and the improvement of education practice.

The purpose here is to define and illustrate innovative uses of two of the tools from Porter's work, content maps and a quantitative index of the degree of alignment. The intention is to put the tools in the hands of education researchers who will use them in the ways illustrated here and in ways that draw on, but extend, what follows. First, the two tools are defined. Next their use is illustrated (a) to make state-by-state comparisons and (b) for the quantitative index use as a dependent variable in a randomized experiment that compares the effects of a professional development program for middle school teachers of mathematics and science to a no-treatment control. The professional development program uses the content maps and index of alignment to empower teachers to reflect on their instruction and undertake changes they deem appropriate.

Before going further, one caveat is appropriate. Alignment is only good for education if the target for alignment is of sufficient quality. There are arguments about what students should know and be able to do, reflected in the reading and mathematics wars. These disagreements are largely matters of values, making it difficult to know who is right and who is wrong, or even what right and wrong might mean. In the case of what students should know and be able to do, with the possible exception of English language arts, there is much more important knowledge, skills, and applications that students might learn than they could possibly learn. Many believe that focusing on a few important areas of content so that students come to a level of mastery is a better strategy than a curriculum that attempts to cover a wide range of content and is, therefore, necessarily shallow in expectations for student mastery (Schmidt et al., 2001). In short, an aligned instructional guidance system resulting in aligned classroom instruction should produce better results on an aligned student achievement test. Standards-based reform, if focused on the wrong content, could have a disastrous result—just as standards-based reform focused on the right content may have a positive result. With that in mind, Gamoran, Porter, Smithson, and White (1997) demonstrated that the degree of alignment between a teacher's instruction and a test of student achievement in high school mathematics accounted for approximately 25% of the variance among teachers.

CREATING A QUANTITATIVE INDEX OF ALIGNMENT

Two-Dimensional Languages for Describing Content

The basic concept behind the quantitative index of alignment, which is the focus of this article, is a two-dimensional language for describing content. One dimension of the language is topics (e.g., within a general content area of number sense/properties/relations, a specific topic might be place value). The other dimension of the

language is categories of cognitive demand (sometimes called expectations for students). Two-dimensional languages have been created for mathematics, science, and English language arts (see www.SECsurvey.org). In middle school mathematics, topics are grouped into the following general areas: (a) number sense/properties/relations, (b) measurement, (c) data analysis/probability/statistics, (d) algebraic concepts, (e) geometric concepts, and (f) instructional technology. Within each of these general areas of content are more specific topics. In mathematics, there are 84 specific topics within the six general content areas. The five categories of cognitive demand are (a) memorize facts, definitions, formulas; (b) perform procedures; (c) demonstrate understanding of mathematical ideas; (d) conjecture, generalize, prove; and (e) solve nonroutine problems, make connections (definitions are given in Figure 2). The content language for mathematics can be represented by a two-dimensional content matrix with 84 rows (specific topics) and 5 columns (categories of cognitive demand).

Mathematics was the initial language created (Porter, Schmidt, Floden, & Freeman, 1978). The topics and categories of cognitive demand were generated through analyzing K–12 mathematics textbooks, standardized tests, state and dis-

Memorize Facts/ Definitions/ Formulas

Recite basic mathematics facts
Recall mathematics terms & definitions
Recall formulas and computational procedures

Perform Procedures

Use numbers to count, order, denote
Do computational procedures or algorithms
Follow procedures/instructions
Solve equations/formulas/routine word problems
Organize or display data
Read or produce graphs and tables
Execute geometric constructions

Demonstrate Understanding of Mathematical Ideas

Communicate mathematical ideas
Use representations to model mathematical ideas
Explain findings and results from data analysis strategies
Develop/explain relationships between concepts
Show or explain relationships between models, diagrams, and/or other representations

Conjecture/ Generalize/ Prove

Determine the truth of a mathematical pattern or proposition
Write formal or informal proofs
Recognize, generate or create patterns
Find a mathematical rule to generate a pattern or number sequence
Make and investigate mathematical conjectures
Identify faulty arguments or misrepresentations of data
Reason inductively or deductively

Solve Non-routine Problems/ Make Connections

Apply and adapt a variety of appropriate strategies to solve non-routine problems
Apply mathematics in contexts outside of mathematics
Analyze data, recognize patterns
Synthesize content and ideas from several sources

FIGURE 2   Definition of cognitive demand.

trict content standards, and national professional standards and by interviewing teachers about the content of their instruction. The goal is to have a language that captures all of the content that might be taught. One of the key decisions is at what level of detail should content be described? The goal is to describe content (a) at a level of detail that teachers believe is important and (b) about which they make decisions as to whether to teach the content or not. At the same time, the language is to be of sufficient detail to capture all-important distinctions among items on an achievement test or among specific objectives in content standards. Obviously, there is no one perfect language. At the same time, a good language is one that operates at a level of detail that people find useful and uses labels for topics and categories of cognitive demand that communicate commonly understood distinctions. Over time and through use, the three content languages have evolved through adjustments in the list of specific topics, their labels, and in the number and labels of cognitive demand. More than 30 states have used the languages to investigate alignment between content standards and student achievement tests. Many schools and districts have used the languages to describe to teachers their instructional practices (see www.SEConline.org).

The content matrix is the starting point for defining a quantitative measure of alignment. If one wishes to measure the degree of alignment between a state's test and that state's content standards in mathematics at a specific grade level, the test and content standards must be content analyzed by experts in mathematics. Each expert independently content analyzes each item on the test (or objective in the standards). The expert makes judgments as to what cells (intersections of topic and category of cognitive demand) in the matrix best represent the content being tested by the item (or content described by the objective). Some items (objectives) are very specific and fall within a single cell. Other items (objectives) are more complex and require students to demonstrate knowledge for more than one topic and/or more than one category of cognitive demand. If the item yields a single score point and is placed in three cells, then for that content analyst the score point is equally distributed across the three cells identified. If an item is worth three points, those three points are equally distributed across the three cells identified by the content analyst. For content standards, each objective is weighted equally (although this is, of course, an assumption, which may not fully reflect what the authors of the content standards had in mind). Some objectives in content standards are quite broad and inclusive, requiring many cells in the content matrix to represent the content of the objective.

Once a content analyst has completed the task for a test (or content standard), that analyst's basic data are reduced to cell-by-cell proportions by dividing the cumulative weight in a cell across all items (or all objectives) by the total number of score points for the test (or total number of objectives content analyzed for the standards). Reducing the data to proportions puts the results in a common metric that allows comparison across different tests and standards. Results are averaged

cell by cell across content analyst to create a two-dimensional matrix of proportions, which sum across rows and columns to 1.0. The proportions in each cell represent the best estimate of the relative emphasis of the content for that cell on the test (or in the standards). In similar fashion, other documents can be content analyzed, for example, textbooks or the content of a professional development experience, if that experience is well represented through materials.

To measure the content of instruction, again the two-dimensional content language is the starting point. Teachers are asked to report on the content of their instruction through a survey. They are asked to first read through all of the specific topics that they might or might not have taught. For those topics taught, the teacher gives an estimate of the amount of time spent on that topic for the period of instruction being described: 0 = none, not covered; 1 = light coverage (less than one class/lesson); 2 = moderate coverage (one to five classes/lessons); and 3 = sustained coverage (more than five classes/lessons). Next, the teacher is asked, for each topic taught, to indicate the degree of emphasis for each level of cognitive demand: 0 = no emphasis (not an expectation for this topic), 1 = slight emphasis (less than 25% of the time on this topic), 2 = moderate emphasis (25% to 33% of the time on this topic), and 3 = sustained emphasis (more than 33% of the time on this topic). Obviously this task is not an easy one for the respondent. At the same time, teachers work through the list of topics quickly and identify many for which they provided no instruction.

The typical target for description is a school year. The shorter the period of instruction being described, the better the description from the teacher. Teachers can be surveyed weekly, once a semester, or at the end of the school year. When surveying more frequently than at the end of the school year, results are aggregated to the school year. Response time obviously varies across teachers, but the once-at-the-end-of-the-year task can usually be completed within 40 min to an hour.

Once the data are in hand for a teacher, they are reduced to cell-by-cell proportions in the content matrix in which, again, the proportions sum to 1.0 over rows and columns. If one wishes to describe a group of teachers, averages across teachers cell by cell creates a group content matrix. What is being measured is not the amount of time a teacher spends on mathematics but within that time the relative emphasis of particular types of content represented by the intersection of topics (rows) and categories of cognitive demand (columns). The focus is on the content of instruction not the pedagogy. Good instruction requires appropriate content well delivered. Unfortunately, good pedagogy for inappropriate content is still bad instruction.

Content Maps

Content matrices of proportions indicating relative emphasis of types of content can be displayed in the form of content maps. A content map is analogous to a top-

ographical map in which the topics are analogous to north and south and the cognitive demands are analogous to east and west. The contours of the map indicate relative content emphasis, with considerable emphasis of a particular type of content analogous to a mountain and no emphasis on other content analogous to sea level. Maps exist at a coarse grain level, depicting relative content emphasis for general areas of content topics and cognitive demand (see Figure 3), and at a fine grain level, depicting relative content emphasis within a content area across specific topics and cognitive demand (see Figure 4). The data for Figures 3 and 4, which illustrate the use of content maps and the alignment index, are for a sixth-grade teacher of mathematics in Miami-Dade (one of the teachers in the experimental study described later). The teacher emphasized performing procedures especially for length/perimeter and area/volume within number sense and relations. Algebraic concepts were not taught. Alternatively, the data could be displayed in a three-dimensional bar chart, but the result is a picture that is busy and difficult to read.
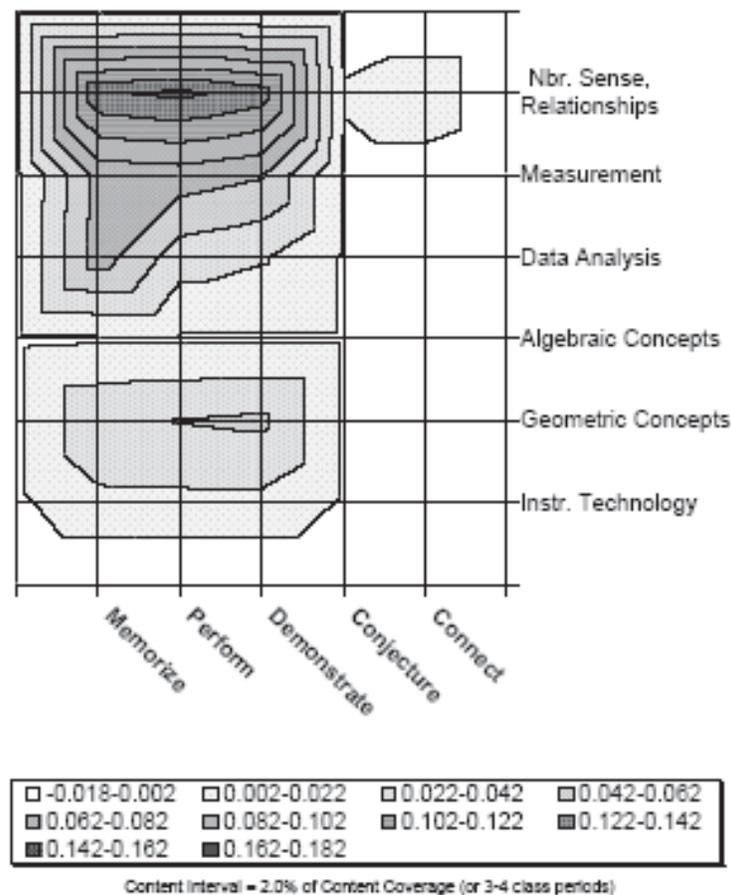

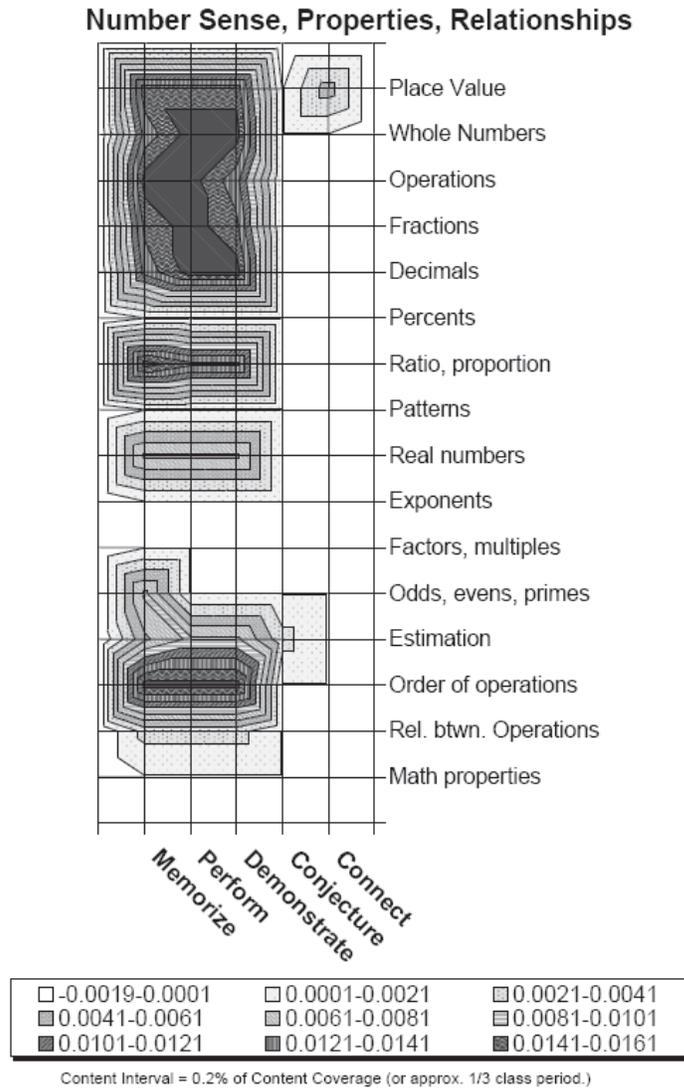FIGURE 3   Miami-Dade Grade 6 teacher coarse grain map.

Number Sense, Properties, Relationships

	-0.0019-0.0001		0.0001-0.0021		0.0021-0.0041
	0.0041-0.0061		0.0061-0.0081		0.0081-0.0101
	0.0101-0.0121		0.0121-0.0141		0.0141-0.0161

Content Interval = 0.2% of Content Coverage (or approx. 1/3 class period.)

FIGURE 4   Miami-Dade Grade 6 teacher fine grain map.

The dimensions of the content matrix are nominal scale and the maps assume continuous (at least ordinal) north/south and east/west scales. Although every intersection of a row and column on the content map is faithful to the basic content matrix of proportions, the smoothing in between points is arbitrary. Nevertheless, the expression "a picture is worth a thousand words" may apply to a content map. The viewer quickly sees what content was emphasized and what content was not. Content maps can be used to display the relative content emphasis of a test, content standards, curriculum materials, professional development, or (the focus here) a teacher's or group of teachers' instructional practices concerning what content is taught and to what degree of emphasis.

Quantitative Index of Alignment

Placing two content maps side by side provides a graphic illustration of areas of agreement and disagreement in content emphasis. For example, content maps might be compared to see the nature of alignment between a state's assessment of student achievement in fourth-grade mathematics and the state's content standards for fourth-grade mathematics. In addition to such graphic displays of content agreement and disagreement, a quantitative measure of the degree of content alignment is also possible.

As for content maps, the basic data for calculating a quantitative measure of alignment are content matrices of proportions. Alignment can be calculated between two teachers, between a teacher and a test, between a teacher and a standard, between a test and a standard, and so forth. The basic question for alignment is the extent to which the pattern of proportions in one content matrix mirrors the pattern of proportions in another content matrix. The alignment index is defined as follows:

$$ | \quad | $$

In this equation x denotes cell proportions in one matrix and y denotes cell proportions in the other matrix. The index can take on values from zero (no alignment) to 1.0 (perfect alignment).[1] Clearly the larger the value of the index, the better the alignment. Because the first step in calculating the index is to put relative content emphasis at the cell level in a common metric of proportions, the interpretation is not exactly the proportion of content in common between, as an example, one teacher and another. The alignment between the two teachers could be perfect, with one teacher spending twice as much time on mathematics as the other. Still, the relative emphasis of content at the cell level could be identical for the two teachers.

One way to think about interpreting the values of the index is normatively. If for most states the alignment of student achievement tests to content standards is .4 and for one state the alignment is .6, that state's alignment would be quite high. Normative interpretation is illustrated in a following section. There are other quantitative indexes of alignment that can be calculated from the basic data of two content matrices of proportions, but all the alternative indexes suggested thus far are nearly perfectly correlated with the index defined previously (Porter, 2002).

Content maps help interpretation of the value of an alignment index. For example, if the degree of alignment between a teacher's instruction and the content standards for that instruction is .4, inspection of the two content maps quickly reveals the nature of the alignment by identifying areas of content for which there is agreement and

---

[1]Conceptually, the index is the sum of cell-by-cell intersects.

areas of content for which there is not. Clearly, there are many quite different pairs of content maps that yield .4 alignment. Similarly, the alignment index helps interpretation of a pair of content maps by describing the overall degree of agreement.

USES OF ALIGNMENT INDEXES

State-by-State Comparisons

One particularly useful display of indexes of alignment is in the form of a matrix analogous to a correlation matrix. For example, one might ask: To what extent is math instruction at eighth grade the same in one state as in another? To the extent that there is alignment between states in the content of instruction, the conclusion would be that there exists, de facto, a national curriculum in eighth-grade mathematics. To the extent that there is not alignment, the conclusion would be that each state has its own unique system, resulting in eighth-grade mathematics instruction at least somewhat unique from one state to the next.

The basic data for such an analysis would start with a state probability sample of teachers for each of several states. The teachers would complete the survey (as described previously). The data from the survey would be analyzed to produce a content matrix of proportions indicating relative content emphasis for each teacher. The content matrices could be averaged cell by cell across teachers, resulting in an average content matrix for each state. From these results, an alignment index could be calculated to compare each state to each other state. The results could be reported in a state-by-state matrix of alignment indexes. Table 1 presents the results from a 10-state study conducted by the CCSSO and

TABLE 1
Alignment of Instruction With Instruction:
Eighth-Grade Mathematics—SCASS Study

State	H	J	K	L	E	O	G	I	M	N
H										
J	.73									
K	.59	.66								
L	.56	.64	.67							
E	.65	.71	.78	.70						
O	.71	.80	.63	.65	.70					
G	.71	.81	.66	.67	.71	.84				
I	.73	.82	.63	.66	.68	.79	.80			
M	.68	.77	.61	.62	.66	.73	.76	.79		
N	.62	.69	.58	.61	.62	.71	.70	.67	.65	

Note. Average alignment = .69. SCASS = State Collaborative on Assessment and Student Standards.

their State Collaborative on Assessment and Student Standards project. Data were collected in the spring of 1999. Unfortunately the teacher samples were not probability samples, but the data provide an illustration of the concept of a matrix presentation of alignment indexes (Blank, Porter, & Smithson, 2001). As can be seen, there is considerable alignment from one state to the next, although the alignment is not perfect, with an average alignment index of .69. As seen later, the alignment of a teacher's instruction to content standards and student achievement testing is substantially less (on average around .3).

The concept of matrices of alignment indexes can be generalized to compare the degree of alignment among states on the content emphases of their tests, content standards, and the like. Matrices of alignment indexes can also compare state tests to state content standards (where the entries in the main diagonal represent within-state alignment). Such a matrix allows exploration of the extent to which a state's test is more aligned with that state's standards than it is aligned with other states' standards. The off-diagonal elements in the matrix provide a baseline for interpreting the extent to which a state's test is especially well aligned with that state's content standards. To the extent that there are uniquenesses across states in their content standards and to the extent that each state has a test well aligned to its own content standards, the alignment indexes in the main diagonal would be large relative to the off-diagonal entries. Alignment can also be assessed between a state's test and the National Assessment of Education Progress test. Similarly, the alignment between state standards and national professional standards, for example, the National Council of Teachers of Mathematics content standards, can be investigated. These types of analyses are described and illustrated in Porter (2002).

Index of Alignment as a Variable

Because the index of alignment can be defined at the level of an individual teacher, the index can be used as a continuous variable in research on teaching. For example, in the process product days of research on teaching (Brophy & Good, 1986), the basic paradigm was to measure student achievement in the fall and again in the spring and observe teacher pedagogical practices during the year. Pedagogical practices were then used to predict gains in student achievement. Unfortunately, the content of instruction was ignored. Some results were found that have been useful, but generally the results were not strong or easily replicated. Perhaps, had the content alignment of each teacher's instruction to the test been used as a control variable in the analyses, the error variance would have been sufficiently reduced to reveal more and stronger relationships between pedagogical strategies and gains in student achievement. Arguably, among school-controlled factors, the degree of alignment of the content of instruction to the test used to measure student achievement is the strongest predictor of gains in student achievement (Gamoran et al., 1997).

The focus here is on using the index of alignment as a dependent variable in assessing the effects of a professional development intervention for middle school teachers of mathematics and science (described later). The study also illustrates that the index of alignment and the content maps can be used to facilitate teacher reflection on their practice. Porter (2002) provides other uses of a quantitative index of alignment as a variable, not only in research on teaching but in studies of effects of state and district instructional guidance policies as well.

TESTING THE EFFECTS OF A PROFESSIONAL DEVELOPMENT PROGRAM

During the period of spring 2001 through spring 2003, we conducted a place-based randomized trial to test the effects on instructional practices of a mathematics/science professional development program for middle school teachers (Grades 6–8; Porter, Blank, Smithson, & Osthoff, 2005). The basic design was to randomly assign schools to treatment or control conditions in each of five districts: (a) Charlotte-Mecklenburg, (b) Chicago, (c) Miami-Dade, (d) Philadelphia, and (e) Winston-Salem. Initially, there were 28 treatment and 27 control schools. Treatment schools formed five- to seven-member mathematics and science leadership teams comprised of one administrator, mathematics and science department chairs, and other math and science teachers in the school. The intervention was a training of trainers model in the sense that school leadership team members received the professional development directly and they in turn worked with the other teachers in their schools to fully implement the intervention. The independent variable that drove the design was the treatment/control contrast. The dependent variables were several, but the ones of interest here are the degree to which teacher instruction was aligned to various possible targets for instruction: the state student achievement test and the state content standards.

The treatment did not specify what particular changes in instructional practices were desired; rather, the treatment was one of empowerment. Teachers were provided data about their instructional practices and the practices of their colleagues in the form of content maps and the extent to which instructional practices were or were not aligned to student achievement tests and content standards through use of the alignment index. Each school with leadership from its team could set reform goals specific to that school. For example, educators in a school might look at their instructional practices and decide they need better grade-to-grade alignment in the content of science being taught. Alternatively, they might decide they want more emphasis on data analysis and probability in their mathematics instruction. They could also decide to focus on pedagogical strategies, although that is not the focus here. One hypothesis was that when provided information on the degree to which instruction was aligned to assessments and standards through alignment indexes

and content maps, the teachers would decide to work on improving the alignment of their instruction in ways that made sense to them. Perhaps they would set the content standards as a target for aligning instruction, or they might see the student achievement test as a better target for instructional alignment (to the extent that the content standards and tests were not perfectly aligned).

The treatment intervention consisted of the following components:

1. Baseline teacher surveys of instructional practice to generate content maps and indexes of alignment.
2. Two-day professional workshop for the school leadership teams that came together district by district.
3. Follow-up technical assistance in the schools.
4. A professional development follow-up workshop.
5. Evaluation of progress and refocusing of assistance at the end of Year 1.
6. A posttest survey.

The professional development was based on the work of Schmoker (2002) and Fullan (2002) on how to use student achievement data for improving instructional practices and the CCSSO work on Surveys of the Enacted Curriculum. Thus, the data-based decision-making professional development taught teachers not only how to use student achievement data but also how to use data from content maps and indexes of instructional alignment. Each treatment school received a customized school-based report at the beginning of the intervention containing approximately 150 pages of school-specific data organized into charts, content maps, and alignment indexes for the teachers in the school (see Figures 3 and 4 for an example of the content maps of one teacher in the study). The intervention taught members of each school leadership team not only how to interpret data on student achievement, pedagogy, the content of instruction, and alignment but also how to work with their colleagues to empower them with the same knowledge and skills (Blank, 2004).

As for most school-based professional development interventions, implementation varied from school to school and even district to district. Response rates from teachers on the surveys were approximately 75% at baseline and again at posttest, although response rates varied considerably from school to school. Attendance at the professional development sessions also varied and, of course, there was teacher mobility resulting in teachers at baseline not present at posttest and vice versa. In Charlotte-Mecklenburg, for example, there was a reorganization of schools, students, and teachers due to a Federal Court desegregation order, resulting in the reassignment of more than two thirds of the students and staff during the course of the study. Because the focus here is to illustrate the use of the alignment index as a dependent variable, further description of the intervention and its implementation is not necessary but can be found in Porter et al. (2005).

SAMPLE AND RESULTS

The basic design for the intervention study had teachers nested within schools and schools nested within treatment/control by district, with district crossed with treatment/control. The dependent variable was gains in alignment of instruction to various possible targets for instruction (e.g., student achievement tests, content standards) from baseline to the end of the intervention. Data were analyzed using hierarchical linear modeling, with treatment/control and district as fixed effects and teachers (Level 1) and schools (Level 2) as random effects. Although schools were randomly assigned to treatment/control conditions, making the design a true experiment, using gains as the dependent variable served to improve precision by reducing the error variance and increasing statistical power.

There was attrition throughout the course of the study. Four schools dropped out before random assignment. An additional three treatment schools dropped out within the first few months. Although response rates from teachers were approximately 75%, the longitudinal nature of the design meant that some teachers who responded at baseline did not at posttest and vice versa. The reorganization in Charlotte-Mecklenburg resulted in attrition, as did the considerable teacher mobility in and out of the schools in all districts during the course of the study.

In general, there were more student achievement tests and content standards in mathematics than in science. Thus, more measures of alignment for more of the research sites were possible in mathematics than in science. For the two North Carolina districts in the study, no science assessment for any grade level was available. For these respondents, no alignment measure for an assessment was calculated. For mathematics in North Carolina, grade-specific assessments and standards for Grades 6, 7, and 8 (the focus grades for the study) were available.

When possible, alignment targets (e.g., tests and standards) were selected for the specific grade level for which instructional content was reported. If a subject and grade-specific target for the relevant state or district was not available, the next higher grade-specific target for which data were available was used. If no target for a higher grade level was available, the closest available grade level assessment or standard from the relevant district or state was used. Of those mathematics teachers for whom alignment measures could be calculated (across Year 1 and Year 3 data results), 95% of assessment alignment measures and 62% of standards alignment measures were based on targets that matched the grade level for which mathematics instruction was reported. For science, 69% of assessment and 63% of standards alignment measures were based on targets that matched the grade level for which science instruction was reported.

For the longitudinal sample with complete data, there were 11 treatment and 11 control schools for mathematics and for science standards but only 7 treatment and 8 control schools for science tests. Mathematics is tested more frequently than science. Similarly, there were fewer teachers in science than in mathematics.

If one includes all of the data available at baseline and posttest, the sample size increases substantially (see Table 2). It was decided to impute data by regressing alignment at baseline on alignment at post, and alignment at post on alignment at baseline using the complete data longitudinal samples. The correlations, based on the longitudinal complete data sample, were quite high: (a) math test alignment .58, (b) math standard alignment .67, (c) science test alignment .78, and (d) science standard alignment .76. For math, there were 18 treatment schools with 138 teachers and 17 control schools with 142 teachers. For science instruction aligned with the test, there were 14 treatment schools with 95 teachers and 14 control schools with 92 teachers. For standards, the sample size increased to 19 treatment schools with 117 teachers and 19 control schools with 126 teachers.

Tables of means for the three types of data sets are found in Tables 3, 4, and 5. In Table 3, the means are reported for the complete data longitudinal sample. Alignment was higher for math than science and higher for standards than tests. Gains in alignment are also reported in the three tables. Gains in alignment to the math test were essentially zero for both treatment and control. Gains in alignment for the math standards were essentially zero for the control but about .03 for the treatment group (equivalent to approximately .5 standard deviations). In science, the gains were similar for treatment and control and smaller than for the math standard.

In math, the mean alignment for test was approximately .20; for standards, the mean alignment was approximately .30. Standard deviations based on teachers ranged from .04 to .08. In science, means were approximately .15 for tests and .18 for standards. Standard deviations based on teachers ranged from .03 to .05. Mean gains in alignment were close to zero, with standard deviations based on teachers ranging from approximately .03 to .08, depending on the data set and conditions. The distributions across teachers of both post alignment and gains in alignment were, in all cases, approximately normal with an occasional

TABLE 2
Sample Size for All Teachers and Schools

	Treatment Test		Control Test		Treatment Standard		Control Standard	
	Year 1	Year 3	Year 1	Year 3	Year 1	Year 3	Year 1	Year 3
Math								
Schools	18	13	15	15	18	13	15	15
Teachers	98	70	90	86	98	70	90	86
Science								
Schools	14	9	14	10	19	12	19	15
Teachers	65	53	67	50	80	62	92	62

TABLE 3
Alignment Means Based on the Complete Data Longitudinal Sample

	Math Test		Math Standard		Science Test		Science Standard	
	Treatment	Control	Treatment	Control	Treatment	Control	Treatment	Control
Year 1	0.204	0.207	0.302	0.304	0.156	0.146	0.180	0.168
Year 3	0.189	0.191	0.328	0.309	0.170	0.165	0.197	0.185
Gains	−0.014	−0.015	0.025	0.005	0.014	0.018	0.017	0.017

TABLE 4
Alignment Means Based on the Imputed Sample

	Math Test		Math Standard		Science Test		Science Standard	
	Treatment	Control	Treatment	Control	Treatment	Control	Treatment	Control
Year 1	0.208	0.220	0.335	0.322	0.150	0.134	0.187	0.169
Year 3	0.194	0.193	0.361	0.324	0.168	0.154	0.200	0.187
Gains	−0.014	−0.027	0.026	0.003	0.018	0.020	0.013	0.018

TABLE 5

Alignment Means Based on All Teachers and Schools

| | Math Test | | Math Standard | | Science Test | | Science Standard | |
	Treatment	Control	Treatment	Control	Treatment	Control	Treatment	Control
Year 1	0.200	0.201	0.301	0.322	0.162	0.146	0.177	0.167
Year 3	0.195	0.197	0.314	0.297	0.165	0.162	0.204	0.183
Gains	−0.005	−0.005	0.013	−0.024	0.003	0.016	0.027	0.015

outlier here or there. In no case was a teacher's alignment greater than .5, and the largest values of alignment (which approached a .5) were for math teachers' alignment to standards.

In the hierarchical linear modeling analyses, district-by-treatment/control interactions were entered into the model but did not increase the percent of variance accounted for and were dropped from further analyses. Consistent with the means reported in Table 3, Table 4, and Table 5, the only significant treatment effect was for increases in alignment of mathematics instruction to the mathematics standards for the imputed sample ($p = .012$). The gain in alignment for the treatment group was .02 greater than the control group on alignment of instruction to math content standards (for an effect size of .36). Results reported in the column labeled *complete* are based on data aggregated to the school level and then analyzed in a treatment/control-by-district analysis. Because gains are defined at the school level (as opposed to the teacher level), the precision is less.

The results are encouraging but not definitive. There should be no surprise that alignment is higher for math than for science. Mathematics receives more attention than science from states and districts; math is tested more frequently, math is a part of the NCLB Act, and math is a core academic subject. Some might be pleased to see that the treatment appears to focus teachers' instructional attention more on standards than on tests. An effect size of .36 is by most standards worth paying attention to.

Why did the intervention not have uniform and robust effects on instructional alignment? One possible explanation is that the intervention was one of teacher and school empowerment with no specific direction as to how instruction should change. We hypothesized that teachers might attempt to improve the alignment of their instruction, and apparently some—at least in mathematics—did and with some success. Further, interventions in urban middle schools are difficult. States take over districts (e.g., Philadelphia), districts reorganize (e.g., Charlotte-Mecklenburg), and teachers are mobile for a variety of reasons. Even if there were greater teacher stability, the challenges of implementation of a training of trainers model are substantial. Given all of the challenges, the results are encouraging. Further, a great deal was learned about how to strengthen the intervention and how to improve implementation. Not surprisingly, we learned that both district and school commitment and leadership is key to successful implementation; achieving district and school commitment must be a priority for any replication study.

There are other dependent variables to be analyzed, but they are not the focus of this discussion. We did not have good data on gains in student achievement and were not optimistic that student achievement could be changed in such a brief period of time. Hopefully, the effects of the intervention will continue to strengthen in the schools, and student achievement will improve. In addition to the promising results for math instruction alignment to standards, we have anecdotal reports from

a number of schools that attribute improvement in student achievement to the professional development intervention.

SUMMARY AND CONCLUSIONS

The central point of this article is that a quantitative measure of alignment can be defined at the individual teacher level and used as an important variable in research on teaching. The quantitative index was defined and detailed descriptions provided of the data on which the index is based. The quantitative measure of alignment is useful for judging alignment among and between individual teachers, groups of teachers, standards, student achievement tests, curriculum materials, and anything else that can be content analyzed. Content maps were also defined and illustrated; content maps provide "pictures" of teachers' content emphases. A demonstration was provided of how content maps can be used to explore the meaning of a particular value of the alignment index. Similarly, matrices analogous to correlation matrices were illustrated for displaying measures of alignment to investigate patterns of alignment between and within states for instruction, student achievement tests, and content standards. Although the illustrations were for mathematics, the tools are available for science and English language arts as well.

Results from a place-based randomized trials test of a professional development intervention for middle school math and science was used to illustrate how the quantitative index of alignment can be used as a dependent variable in investigating the significance of a treatment versus control contrast. Importantly, the distributions of the alignment index at a particular point in time for teachers and gains in alignment between two points in time for teachers were found to be approximately normal (with an occasional outlier here or there).

There are a number of issues one might raise about the reliability and validity of the content maps, the index of alignment, and the data on which both are based. First and most important, the data for measuring the content of instruction is based on teacher self-report collected through surveys. There are two potential problems with such data: (a) teachers may not complete the survey and (b) teachers who complete the survey may not provide accurate information. Both are real problems that require work on the part of the researcher. Experience has shown that, with effort, teacher response rates to the survey can be as high as 75% for a national probability sample (Garet, Porter, Desimone, Birman, & Yoon, 2001). In studies such as the one reported here, in schools where there was a strong commitment to the study, the response rate was 100% (21 schools had 90% or greater response rates on the baseline measure).

As for the validity of the data from teachers who do complete the survey, there has been much recent research that offers support. Most surveys have been criti-

cized for not being able to distinguish true variation (Burstein et al., 1995). Recent investigation has found that survey measures like the one used here are effective in describing and distinguishing among different types of teaching practices, especially when researchers use composite indicators rather than single-item indicators (Mayer, 1999). Several studies have shown that anonymous teachers' self-reports on their teaching are highly correlated with classroom observations and teacher logs and that one-time surveys that ask teachers questions about the content and strategies they emphasize are quite valid and reliable in measuring teachers' instruction (Mullens, 1995; Mullens & Gayler, 1999; Mullens & Kasprzyk, 1996; Schmidt, McKnight, & Raizen, 1997; Shavelson, Webb, & Burstein, 1986; Smithson & Porter, 1994). The problem of teachers being inclined to answer surveys in socially desirable ways has been shown to be less of a problem with anonymous surveys and when the results will not be used for accountability (as was the case here) than in focus groups or interviews in which teachers are in a more public forum (Aquilino, 1994, 1998; Dillman & Tarnai, 1991; Fowler, 1995). Further, when survey questions do not seek judgments of quality but rather an accounting of behaviors, such as is the case with the survey of content, social bias is less and the validity and reliability of teacher self-report data can be quite high (Mullens & Gayler, 1999).

Clearly, survey data can vary in quality. Every effort must be made to ensure a high response rate. At the same time, every effort must be made to ensure that respondents understand the task and feel comfortable completing the task. Especially, respondents must be assured that the data will not be used in any way that might have negative repercussions for the respondent. In addition, there are issues of respondent memory. The more frequently a survey is conducted, the less the challenge to a respondent's memory. For the type of survey used here, agreement between daily logs aggregated to a full school year and end-of-year surveys was quite good, with correlations in the range of .6 to .8 (Porter, Kirst, Osthoff, Smithson, & Schneider, 1993).

The content analyses can suffer from the inability of content analysts to report the same data when independently content analyzing a student achievement test or a state's content standards. Using generalizability theory, the reliability at the content matrix cell level of average proportions across two analysts has been found to be approximately .70, and across four analysts approximately .8 (Porter, 2002). For these analyses, the data are represented as a matrix of all cells of the content matrix by content analysts. Increasing the number of analysts beyond four yields diminished returns on increased reliability. Surprisingly, the reliability for describing relative emphasis of content for tests was no different than for content standards for the data sets available (results for English language arts are not available).

The challenges of producing reliable and valid data are many and continuing. Future studies might investigate strategies that result in greater response rates. Future work might also investigate ways in which the survey format and language

could be modified to more effectively and accurately communicate to each respondent the nature of the task and how it is to be completed.

The number of different types of uses of the quantitative index of alignment for research on teaching is just beginning to be explored. In this article, the index served as a dependent variable in a professional development intervention study. The index could be a dependent variable in any investigation of the effects on the alignment of teachers' instructional content practices to any one of a number of different targets (e.g., standards, tests, curriculum materials). These studies could be either experimental, as illustrated here, or studies of natural variation.

The goal of standards-based reform is to create an aligned instructional guidance system that results in classroom instruction aligned to challenging state content standards. The tools described here can play two roles in such investigations. First, the index can be used to measure the extent to which the instructional guidance system is truly well aligned and, through use of the content maps, to provide formative feedback as to how alignment might be improved. Second, the alignment index can serve as a dependent variable to measure any changes in the alignment of instruction over time or to investigate natural variation among states that have more or less well-aligned instructional guidance systems.

Alignment can also serve as an intervening variable in investigations of the effects of teaching practices on gains in student achievement. Similarly, the quantitative index of alignment can be used as a control variable to decrease the error variance and improve precision in investigations of the influence of teachers' pedagogical practices on gains in student achievement. It is left to the readers to explore additional uses of these tools to measure and describe the content of teacher's instructional practices and how they are aligned from one teacher to another or to various external targets (e.g., content standards).

There is considerable room for further work in the area of tools for measuring and describing alignment of instruction to content standards, student achievement tests, and the like. For example, the tools described here are limited to the content of instruction and one approach to defining content. What might be the value of adding measures of pedagogy to measures of content? Are there other ways to define content that would be more powerful? For example, in addition to the two dimensions of topics and cognitive demand, might there be a third dimension of how content is presented? Clearly, we need to know more about how to interpret values of the quantitative measure of alignment. As described previously, one approach is normative. Is the value of the index observed larger than what one might ordinarily expect? There may be other interpretations along the lines of a squared correlation coefficient being the percent of variance accounted. What is the random sampling distribution of the quantitative index of alignment? Is it possible to build procedures for putting confidence intervals

around an estimate of alignment or around the difference between two estimates of alignment to allow hypothesis testing? For the content maps, are there alternative graphic displays that are easier to accurately interpret by teachers, by researchers, or by both? These are just a few of the issues that lie ahead.

ACKNOWLEDGMENTS

REFERENCES

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

Aquilino, W. S. (1994). Interview mode effects in drug surveys. *Public Opinion Quarterly, 58,* 210–240.

Aquilino, W. S. (1998). Effects of interview mode on measuring depression in younger adults. *Journal of Official Statistics, 14,* 15–30.

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.

Blank, R. K. (2004). *Data on Enacted Curriculum study: Summary of findings.* Washington, DC: Council of Chief State School Officers.

Blank, R. K., Porter, A. C., & Smithson, J. L. (2001). *New tools for analyzing teaching, curriculum and standards in mathematics and science* (Report from Survey of Enacted Curriculum Project; National Science Foundation REC98-03080). Washington, DC: Council of Chief State School Officers.

Brophy, J. E., & Good, T. (1986). Teacher behavior and student achievement. In M. C. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed., pp. 328–375). New York: Macmillan.

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

Dillman, D. A., & Tarnai, J. (1991). Mode effects of cognitively designed recall questions: A comparison of answers to telephone and mail surveys. In P. N. Beimer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz, & S. Sudman (Eds.), *Measurement errors in surveys* (pp. 367–393). New York: Wiley.

Fowler, F. J., Jr. (1995). Improving survey questions: Design and evaluation. *Applied social research methods series* (Vol. 38). Thousand Oaks, CA: Sage.

Fullan, M. (2002). The change leader. *Educational Leadership, 59*(8), 16–20.

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,* 325–338.

Garet, M. S., Porter, A. C., Desimone, L., Birman, B. F., & Yoon, K. S. (2001, Winter). What makes professional development effective? Results from a national sample of teachers. *American Educational Research Journal 38,* 915–945.

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,* 522–549.

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

Mullens, J. (1995). *Classroom instructional processes: A review of existing measurement approaches and their applicability for the teacher follow-up survey* (NCES 95-15). Washington, DC: National Center for Education Statistics.

Mullens, J. E., & Gayler, K. (1999). *Measuring classroom instructional processes: Using survey and case study field test results to improve item construction* (NCES 1999-08). Washington, DC: National Center for Education Statistics.

Mullens, J., & Kasprzyk, D. (1996). Using qualitative methods to validate quantitative survey instruments. In *1996 Proceedings of the Section on Survey Research Methods* (pp. 638–643). Alexandria, VA: American Statistical Association.

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

Porter, A. C., Blank, R., Smithson, J., & Osthoff, E. (2005, May). Place-based randomized trials to test the effects on instructional practices of a mathematics/science professional development program for teachers. *Annals of the American Academy of Political and Social Science, 599,* 147–175.

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., Schmidt, W. H., Floden, R. E., & Freeman, D. J. (1978). Practical significance in program evaluation. *American Educational Research Journal, 15,* 529–539.

Rothman, R. (2003). *Imperfect matches: The alignment of standards and tests.* Unpublished paper commissioned by the National Research Council, Committee on Test Design for K–12 Science Achievement, Washington, DC.

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

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

Schmidt, W. H., McKnight, C. C., & Raizen, S. A. (1997). *A splintered vision: An investigation of U.S. science and mathematics education.* Boston: Kluwer.

Schmoker, M. (2002). Up and away. *Journal of Staff Development, 23*(2), 10–13.

Shavelson, R. J., Webb, N. M., & Burstein, L. (1986). Measurement of teaching. In M. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed., pp. 50–91). Washington, DC: American Educational Research Association.

Smithson, J. L., & Porter, A. C. (1994). *Measuring classroom practice: Lessons learned from efforts to describe the enacted curriculum—the Reform Up Close study* (CPRE Research Rep. Series No. 31). Madison: University of Wisconsin, Consortium for Policy Research in Education.

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 Rep. No. 3-024). Ann Arbor: University of Michigan, School of Education, Center for the Improvement of Early Reading Achievement.