American Educational Research Journal
http://aerj.aera.net

How Well Aligned Are State Assessments of Student Achievement With State Content Standards?

Morgan S. Polikoff, Andrew C. Porter and John Smithson

Am Educ Res J published online 16 June 2011
DOI: 10.3102/0002831211410684

The online version of this article can be found at:
http://aer.sagepub.com/content/early/2011/06/15/0002831211410684

Published on behalf of

American Educational Research Association

and

http://www.sagepublications.com

American Educational Research Journal
Month XXXX, Vol. XX, No. X, pp. 1–31
DOI: 10.3102/0002831211410684

How Well Aligned Are State Assessments of Student Achievement With State Content Standards?

Morgan S. Polikoff
University of Southern California
Andrew C. Porter
University of Pennsylvania
John Smithson
University of Wisconsin-Madison

Coherence is the core principle underlying standards-based educational reforms. Assessments aligned with content standards are designed to guide instruction and raise achievement. The authors investigate the coherence of standards-based reform's key instruments using the Surveys of Enacted Curriculum. Analyzing 138 standards-assessment pairs spread across grades and the three No Child Left Behind tested subjects, the authors find that roughly half of standards content is tested on the corresponding test and roughly half of test content corresponds to the standards. A moderate proportion of test content is at the wrong level of cognitive demand as compared to the corresponding standards, and vice versa. Between 17% and 27% of content on a typical test covers topics not mentioned in the corresponding standards. Policy and research implications are discussed.

Keywords: measurement, accountability, education policy, assessment, curriculum, high stakes testing

---

Morgan Polikoff is an assistant professor of K-12 policy and leadership at the University of Southern California's Rossier School of Education, Waite Phillips Hall 904D, Los Angeles, CA, 90089; e-mail: *polikoff@usc.edu*. He studies the design of standards-based reform policies and their effects on teachers' instruction and student outcomes.

Andrew C. Porter is dean of the Graduate School of Education, University of Pennsylvania. He is an applied statistician and psychometrician who studies the measurement of education leadership, student achievement testing, curriculum policies and their effects, and professional development for teachers.

John Smithson is a research associate at the Wisconsin Center for Education Research at the University of Wisconsin-Madison. His work has focused on developing indicators of classroom practice and instructional content.

The primary rationale underlying standards-based reform, the most prominent K-12 education policy of the past 20 years, is coherence. In the earliest description of Smith and O'Day's (1991) systemic reform, instructional coherence was identified as a necessary component for wide-scale educational change. In that vision, instructional coherence referred to the creation of rigorous curriculum frameworks in the core academic subjects and the support of those frameworks through mutually reinforcing sources, such as aligned curriculum materials, pre- and in-service teacher training designed to support the frameworks, and aligned assessments of student learning to provide information about school progress toward student mastery of those frameworks (Clune, 1993; Smith & O'Day, 1991). Instructional coherence was thought to focus schools' and teachers' attention on the key content students were to know and be able to do and drive wide-scale instructional improvement. Well before high-stakes accountability became a part of standards-based reform, there was instructional coherence at the core.

Twenty years on since systemic reform and the state systemic initiatives, instructional coherence remains an important part of standards-based reform in its current incarnation, the No Child Left Behind Act (NCLB). The text of NCLB echoes the framework for systemic reform laid out in the early 1990s, claiming that improving student achievement and ensuring access to a high-quality education for all will be accomplished first and foremost through "ensuring that high-quality academic assessments, accountability systems, teacher preparation and training, curriculum, and instructional materials are aligned with challenging State academic standards so that students, teachers, parents, and administrators can measure progress against common expectations for student academic achievement" (No Child Left Behind Act of 2001, 2002, pp. 1439-1440). NCLB mentions alignment dozens of times, specifically focusing on the alignment of assessments with content standards. States are required to show to the Department of Education that their assessments are aligned with the content specified in their standards, and they are also required to assist local entities in identifying curricula that support those standards. Clearly, the coherence of the system remains of utmost importance in the vision of standards-based reform under NCLB.

Given the fundamental role of coherence in the theory of change underlying standards-based reform, the literature on the extent to which standards-based reform has resulted in the level of coherence called for in the initial theory is thin. To be sure, there are a number of investigations of the alignment of standards and assessments within individual states, required by law under NCLB. These studies generally use Webb's (2005, 2006) alignment procedure and find that the assessments are well aligned to the standards. However, there are few cross-state studies that compare the content of assessments with their target standards, and the studies that exist are dated in

the context of the increasing involvement of the federal government in state educational policy (Porter, 2002; Resnick, Rothman, Slattery, & Vranek, 2003; Webb, 1999, 2002; Wixson, Fisk, Dutro, & McDaniel, 2002). Furthermore, the studies generally provide limited information about the detailed nature of the alignment or misalignment across states, grades, and academic subjects. Such information is of fundamental importance in developing an understanding of the extent to which standards-based reform has been implemented as intended and, therefore, in interpreting results about the effects of standards-based reform on teachers and students.

Against this backdrop, the purpose of this study is to investigate the coherence of state content standards and assessments. Data from content analyses of grades 3-12 standards and assessments in states in English language arts and/or reading (ELAR), mathematics, and science (a total of 138 pairs of documents) are used to address the following research questions:

Research Question 1: To what extent are state assessments of student achievement under No Child Left Behind aligned with state content standards?
Research Question 2: To the extent that there is misalignment between assessments and standards, what is the nature of that misalignment?

The results shed important light on the implementation of standards-based reform in the observed states and the ways in which the policy could be improved in terms of its design and effectiveness.

Background

The Rationale for Alignment

As originally articulated, the need for a coherent, systemic approach to educational reform was based on a number of observations about U.S. schools. Most importantly, the federalist system of education in the United States had created a fragmented system (O'Day & Smith, 1993; Smith & O'Day, 1991), with incoherent curricula (Cohen, 1989; Goodlad, 1984) and poor quality teacher preparation and professional development (Smith & O'Day, 1991). The system was fragmented not only in terms of differences in expectations across sites, but also incoherence in external pressures faced by teachers within sites—especially the lack of relationship between tested outcomes and the curriculum (Cohen & Spillane, 1992). This fragmentation was a function not only of the Constitution's delegation of educational authority to the states, but also the states' historic devolvement of most educational functions to districts.

According to the theory, this fragmentation helped ensure that educational reforms rarely, if ever, achieved their target of change (Cuban, 1984; Tyack & Cuban, 1995). When change did come, it was not often of the type envisioned by policymakers. Change was often additive rather than

substitutive, and policies were often heavily adapted or modified during implementation in ways that were often contrary to the original intent (McLaughlin, 1990; Tyack & Cuban, 1995). Policymakers blamed educators for these failures, arguing that teacher capacity for broad school improvement was weak. In contrast, educators argued that the reforms were passed down without consideration for the varied needs of schools (Tyack & Cuban, 1995). In short, by the late 1980s, educational researchers had come to realize that the standard model of policy implementation would not achieve desired results in improving schools throughout the nation's 10,000+ districts. Prominent researchers and policymakers believed a more coherent approach was needed.

The vision of systemic reform was a combination of top-down and bottom-up approaches designed to impact classroom teaching and learning (Smith & O'Day, 1991). The focus was on the state as the primary level of policymaking. Smith and O'Day (1991) outlined three facets of systemic reform: a unifying vision, a coherent system of instructional guidance, and a restructured school governance system. The first step in the theory of change was establishing instructional coherence. Smith and O'Day recognized that teaching and learning were the core of the school's work. They advocated the creation of deep curriculum frameworks (now more commonly called content standards) in the core content areas. The frameworks would provide guidance by establishing a set of core content expected of all students. Along with frameworks, they recommended the adoption of high-quality curricular materials aligned to the frameworks. To support teachers, they advocated coherent pre- and in-service professional development in content and pedagogy, guided by performance assessment of prospective teachers to ensure program quality. As the final piece of instructional guidance, they encouraged the creation of high-quality student assessments aligned with achievement goals for the purposes of providing information on school progress. Properly designed accountability measures for students or schools might provide additional incentives for improvement, but in the original vision the decisions on whether to use test results for accountability purposes would be left to states.

The original theory also included a focus on unifying vision and restructured school governance. In terms of goals, Smith and O'Day (1991) advocated a focus on the full distribution of students with explicit, measurable targets based on educational outcomes. In terms of school governance, Smith and O'Day supported increased school-level control over selection and development of personnel, the sharing of authority and responsibility within the school, and the capital resources to reach the goals. With the three components in place, the theory suggested that (a) the quality of instruction would improve, based on the introduction of the content frameworks and the emphasis on teacher control over implementation of the standards; (b) teacher collaboration would improve, as some of the responsibility formerly

vested in the school and district leadership would be devolved to teachers; (c) teacher pre- and in-service training would improve with the new focus on coherence and capacity building; and (d) the assessment and accountability system would encourage schools and teachers to focus on all students. The proposed reforms were supported by influential researchers and policymakers (Clune, 1993; Jennings, 1998; Ravitch, 1995) and were the model for state reforms in the 1990s, such as the National Science Foundation's Statewide Systemic Initiatives, begun in 1992 (Clune, 2001).

As systemic reform evolved during the 1990s, the focus on goals and instructional coherence remained, but the school governance reforms did not, with control instead shifting to states and even the federal government (Chatterji, 2002; Cohen, 1995). Nevertheless, standards-based reform, systemic reform's successor, intensified the focus on instructional coherence through the creation of content standards and the alignment of state educational policies in support of those standards. Foremost among the educational policies demanding alignment with standards are the state assessments of student achievement used for measuring school and district progress toward meeting the goal of improved student proficiency in the core academic subjects. Alignment of standards and assessments is demanded by the law (No Child Left Behind Act of 2001, 2002) and is critical for ensuring the validity of the inferences made from assessment results (Kane, 2008).

Measuring the Alignment of Standards and Assessments

There are three primary methods for establishing the coherence (alignment) of standards and assessments (for a complete review, see Martone & Sireci, 2009; Porter, 2006). The most widely used by states is Webb's (2002, 2007) alignment procedure. In the Webb procedure, panelists code the depth-of-knowledge (DOK) of a set of standards, goals, objectives, and assessment items. Here, standards are the broadest, with goals underneath standards and objectives underneath goals. Alignment is judged on four measures. Categorical Concurrence indicates the extent to which the assessment and the standards contain the same content, with a focus on the extent to which each macro-level standard is adequately covered by the test content. DOK consistency indicates the degree to which the assessment tests content at or above the specified level of cognitive challenge in the standards. Range of Knowledge Correspondence indicates the extent to which the breadth of knowledge required by the standards and assessments is comparable, focusing on the proportion of objectives in the standards that are tested. Finally, Balance of Representation indicates the degree to which the test's coverage of the standards is balanced across objectives. There is a criterion for each of the four measures, and a test and standards are deemed to be in alignment if all four criteria are met (Webb, 2007).

Reports are made for individual alignment analyses; many of these reports are available online (e.g., Webb, 2005, 2006). The reports generally indicate that standards and assessments are aligned and, where they are not, make suggestions for strategies to improve alignment.

The Achieve procedure (Resnick et al., 2003) has been designed to measure the alignment of standards and assessments as well as the degree of challenge of the assessment. Again, analysts consider the content of standards/objectives and assessment items, rating the match on four scales. Content Centrality indicates the extent to which the test item's content matches the objective's content well, partially, or not at all. Performance Centrality indicates the extent to which the test item's cognitive demand level matches the level specified in the objective well, partially, or not at all. Items are also rated on Challenge, with a focus on Source of Challenge—the extent to which the level of challenge of the item is due to construct irrelevant sources—and Level of Challenge—the extent to which the item set has a range of difficulty that is both matched to the level of difficulty of the standards and appropriate for the target students. Finally, the analysts rate the assessment-standard pair on Balance and Range, indicating the extent to which the assessment covers a wide range of content from the standards and the extent to which emphasis is balanced across topics. Again, summary measures are made on each criterion; also, narrative discussions of alignment are presented.

The third prominent method for estimating alignment uses data from the Surveys of Enacted Curriculum (Porter, 2002). Unlike the Webb and Achieve approaches, the Surveys of Enacted Curriculum (SEC) approach does not rely on direct comparison of assessments or assessment items with objectives or standards. Instead, trained content analysts first map the standards and assessments onto a common framework—a content taxonomy, developed by subject matter experts. The taxonomies define content at the intersection of topics and cognitive demands. Analysts place assessment items and objectives from standards documents into the taxonomies, and the documents are then represented as matrices of proportions, where the proportion in each cell (topic and cognitive demand) indicates the proportion of total content in the document that emphasizes that particular combination of topic and cognitive demand. Then, the matrices for standards and assessments are compared, cell by cell, and an alignment index is calculated, indicating the proportion of content in common. There are neither subscales indicating different pieces of alignment, nor is there an absolute criterion indicating adequate alignment. The SEC procedure is discussed in more detail in the method section.

A recent review of the three approaches (Martone & Sireci, 2009) found strengths in each approach. The Webb approach was seen to provide the strongest quantitative information for evaluating alignment on multiple criteria. In contrast, the Achieve method provides the most useful narrative

summary of alignment, and it also discusses the challenge of the assessment. The SEC approach was applauded for its applicability to instructional issues—it is the only tool of the three that can be used to compare the alignment of instruction with standards or assessments. However, it was seen to provide a less detailed evaluation of alignment. The research presented here uses the SEC procedure and supplements the alignment index with other measures of the agreement and disagreement between standards and assessments. In addition to allowing us to address important, policy-relevant questions about the alignment of standards and assessments, these descriptive components address some of the criticisms of Martone and Sireci (2009) and expand the applicability of the SEC approach for investigating alignment.

Previous Investigations of the Alignment of Standards and Assessments

There are a few studies of within-state standards-test alignment that consider multiple states. An early study of three states in science and four in mathematics was conducted by Norman Webb (1999) using his alignment procedure. The results revealed varying degrees of alignment across states, subjects, and alignment indicators. In particular, a large proportion of the assessments tested material that was of lower cognitive demand than the material identified in the standards. Also, the state assessments failed to cover 50% or more of the content in the content standards, falling short of the range-of-knowledge criterion. Overall, the results suggested poor to moderate alignment depending on the particular criterion, with few recognizable patterns across content areas or grades. A later study updated the results and expanded the analyses to ELAR and social studies, with mainly similar findings (Webb, 2002).

In a 2002 study, Porter reported on the alignment of standards and assessments in mathematics among four states and with the National Council of Teachers of Mathematics (NCTM; 2000) standards. Using the Surveys of Enacted Curriculum content coding procedures, Porter found that the average within-state standard-test alignment was .40. Surprisingly, this value was not appreciably higher than the average standards-test alignment between states (.39) or the average state assessment–NCTM standards alignment (.39). The conclusion was that there was low to moderate alignment between state standards and their assessments in mathematics.

Using the Achieve alignment method, a team of researchers (Resnick et al., 2003) examined the assessments of five states in ELAR and mathematics. Across grade levels, they generally found that individual items mapped quite well to standards—there was little information being tested that was not included in state standards. Looking at the tests as a whole, however, they found that the tests often covered a small proportion of the content in standards (as low as 27%), leaving large proportions of state standards completely

untested. Furthermore, global ratings of the level of challenge of the assessments indicated that with few exceptions, the tests were inappropriately easy as compared to the standards.

Finally, a study by Wixson and colleagues (2002) consisted of surveys of all 50 states and in-depth analyses of 4 states in ELAR. The surveys found that more than 90% of states reported their assessments were aligned to their standards (not surprising, since this is what the law requires). Their analysis, based on a modified Webb framework, found varying degrees of alignment, with especially poor alignment in the states with the largest number of macro-level standards (often called content strands). Furthermore, they noted a tendency for the cognitive complexity of items to be somewhat lower than the cognitive complexity identified for objectives. This finding held in 3 of the 4 states. In short, the reported alignment was higher than the alignment as coded by experts.

It is difficult to report on and compare the previous literature on standards-assessment alignment because each study uses a different framework and different states. In general, however, there is agreement across these few studies that no matter the alignment method used, the average alignment of standards with assessments is moderate. Both the Resnick et al. (2003) and Webb (1999, 2002) studies found that large proportions of the content standards were not tested. They also found that the cognitive demand of assessment items was often lower than that of the corresponding content standards. The Wixson et al. (2002) study confirmed poor alignment in ELAR, with major discrepancies based on cognitive complexity. There are several limitations to the studies, however. First, the data are generally old, predating NCLB. Second, each used samples of five states or fewer. Third, none examined all three NCLB content areas.

In what follows, we use an alignment procedure with documented reliability to provide a systematic account of the coherence of state content standards and assessments since the existence of NCLB for all three subjects addressed in NCLB for 19 states. The results allow the most comprehensive and detailed accounting to date of the implementation of standards-based reform as reflected in the two lead policy instruments of content standards and assessments. From the results, lessons are drawn for interpreting the effects of standards-based reform and possible ways to strengthen the state policies.

Method and Data

To address the research questions, we use data from the Surveys of Enacted Curriculum. Through state participation in the Council of Chief State School Officers (CCSSO) State Collaborative on Assessment and Student Standards (SCASS), a 31-state database of SEC alignment studies exists, which allowed secondary data analysis.

The SEC procedure originated with the work of scholars at Michigan State University who sought to identify the factors influencing the content decisions of elementary school mathematics teachers (e.g., Porter, Floden, Freeman, Schmidt, & Schwille, 1988). In that line of research, a three-dimensional language for describing the content of instruction was used, with the dimensions representing specific topics, cognitive demand levels, and modes of presentation. In later work, the language was generalized to science and ELAR, the "mode of presentation" dimension was dropped, and the other dimensions were modified. The result is a two-dimensional framework of topics by cognitive demands. The framework has been used widely used to examine the content of standards, assessments, and teachers' instructional practices.

Each subject is divided into general areas: 16 for mathematics, 14 for ELAR, and 27 for science. The general areas for each subject are listed in Appendix A in the online version of the journal. Each general area is divided into between 4 and 19 specific topics, for a total of 217 specific topics in mathematics, 163 in ELAR, and 211 in science. For each specific topic, there are five levels of cognitive demand. These differ by subject area, and their definitions are presented in Appendix B in the online journal. For instance, in mathematics the cognitive demand levels are memorize, perform procedures, demonstrate understanding, conjecture/generalize/prove, and solve novel problems/make connections. The content languages can be used to analyze content standards, assessments, curriculum materials, and instructional practices.

The documents are coded by trained analysts. For training, a sample set of items is coded by each analyst. The sample codes are discussed by the group to establish common understandings of the coding framework and clear up general confusions. Next, the actual coding begins, with each analyst working independently. Items/objectives can be flagged for discussion, and analysts can reconvene to address their concerns. The final judgment of correct placement of an item/objective in the framework is left up to the individual coder. For assessments, the content analyst reviews each test item and decides what intersection of specific topics by cognitive demands (cells) are represented. Because some assessment items tap multiple cells in the framework, up to three cells are allowed to represent a single item. For standards, the most specific statements, often called objectives, are coded. Each objective is allowed to represent up to six cells.

Each document (standard or assessment) is analyzed by three to five content analysts; content analysts are common across some but not all documents analyzed (e.g., a reviewer analyzes the full set of eighth-grade mathematics items for a particular state but may or may not analyze the items for any other states). For each analyst, the coding data are converted to proportions of total test/standard content, with a sum of one over all cells. Each item is weighted based on the number of points it is worth on the

assessment; for instance if a 3-point item is placed in two cells, each cell receives 1.5 points. For standards, each objective is weighted equally. While it would be possible to devise alternative weighting strategies for objectives, such as weighting by some measure of "importance," such strategies would require a well reasoned and accepted measure of each objective's importance relative to each other objective's importance. For most state standards documents, objectives are generally presented as lists with no indication of the degree of importance, so the measure of importance would need to be based on some hypothetical proxy for author-intended importance (e.g., word count). The exception to this rule is for those standards documents that identify "focal" or "power" standards (e.g., Kentucky's Core Content for Assessment), but only certain documents contain such indications. Given these challenges with identifying an alternative weighting scheme for objectives, we decided to weight all objectives equally. A benefit of this approach is that it is objective and replicable. Once the matrix of proportions has been calculated for each coder, the proportions in each cell are averaged across content analysts. The data are stored at the Wisconsin Center for Educational Research (WCER) at the University of Wisconsin-Madison.

Data

While there are a total of 31 states with either standards or assessments in the WCER database, this study's sample was constrained to states where a pair of documents (content standards and an assessment) from at least one grade was in the database. See Table 1 for a listing of states in the sample and years when content analyses were done. There are 19 states included in the analyses—11 for ELAR, 14 for mathematics, and 9 for science. Some states have multiple pairs of documents (e.g., Indiana in ELAR has document pairs in each grade 3-12), with a total of 138 pairs of documents across the three subjects—53 in ELAR, 62 in mathematics, and 23 in science. All assessment data are based on the analysis of a single form of a state's assessment. Often, the single form contains a large proportion of common "core items" that are contained in all test forms. While the content analyses took place throughout the 2000s, more than half of the documents included in the analyses presented here were still in use as of the 2008-2009 school year. Aside from Table 1, the state names are not presented with the results because of confidentiality agreements with states.

The quality of the content analysis data has been investigated with a generalizability theory *d*-study. Using generalizability coefficients for raters by cells to estimate reliability, Porter, Polikoff, Zeidner, and Smithson (2008) investigated results for two states' ELAR and mathematics for content standards and assessments drawn from the same data set as analyzed here. For ELAR, reliabilities averaged .74 across grades and states for three coders.

Table 1
List of Standards and Assessments by State and Grade

	English		Mathematics		Science	
State	Grades	Years	Grades	Years	Grades	Years
ME	3-8	2005-2006			4, 8, 11	2004
OK	3-5, 8, 10	2003-2008[a]	3-8 Alg1, Alg2, Geo	2004-2008[a]	5, 8, 10	2004-2005[a]
KS	3-8, HS	2003-2006[a]	3-8, HS	2004-2006[a]		
OH	3-8, 10	2003-2006[a]	3-10	2002-2006[a]	5, 8, 10	2005-2007[a]
IN	3-11	2004-2007[a]	3-8, Alg1	2002-2005[a]	5, 7, 10	2006[a]
NY	3-8, HS	2005[a]	8	2003		
WI	4, 8	2003[a]	4, 8	2002[a]	4, 8	2002[a]
UT	4-6	2008[a]				
MT	4, 8	2005[a]	3-8, 10	2005-2008[a]	4, 8	2007-2009[a]
NC	5, 8, 9	2003				
VA	6, 11	2008[a]				
FL			3, 5, 8, 10	2003-2004	5, 8	2003
NH			3, 6	2002-2005	10	2004
ID			3, 4, 7, 8	2004-2006[a]		
OR			3-8, 10	2003-2006[a]	5, 8	2004-2005[a]
MS			4, 8	2004		
TX			6, 8	2003		
WV			8	2002		
IL					4, 7	2003-2006[a]

[a]Indicates the target standards were still in use as of the 2008-2009 school year.

For mathematics, three-coder reliabilities averaged .78 across grades and states. The more analysts there were for the document, the higher the generalizability coefficient was. In an earlier study, Porter (2002) calculated equally high generalizability coefficients for mathematics. All data used here are based on three to five content analysts.

The cell proportions can be converted to an alignment index (Porter 2002), which indicates the extent to which two documents have the same content messages (i.e., the extent to which the cell proportions are equal to each other across two documents). The index is

$$\text{alignment} = 1 - \Sigma(|x_i - y_i|) / 2$$

where x_i indicates the cell proportion in cell i for document x and y_i indicates the cell proportion in cell i for document y. The index ranges from 0 to 1, with 1 indicating perfect alignment (100% content in common). The alignment index is defined at the cell level, such that content in any two documents is aligned only if it represents the same topics and cognitive demands in the same proportions. Thus, the alignment index can also be

interpreted as the proportion of each document's content that is in common with the other document's content.

It would also be possible to define alignment in other ways using SEC data, such as at the marginal for topic or cognitive demand. There are two important reasons for defining alignment at the cell level (i.e., intersection of specific topics and levels of cognitive demand). First, the frameworks were developed over time with input from teachers and content experts, and teachers describe their content decisions at the intersection of specific topics and levels of cognitive demand (Porter, Kirst, Osthoff, Smithson, & Schneider, 1993). Second, previous research indicates that teacher content coverage is predictive of value-added to student achievement when coverage is defined at the intersection of specific topics and levels of cognitive demand but not when defined at the topic or cognitive demand level alone (Gamoran, Porter, Smithson, & White, 1997).

Content maps can be used to illustrate the nature of alignment or misalignment. The content maps are generated using Microsoft Excel and resemble topographical maps where specific topics are displayed like lines of latitude and cognitive demands like lines of longitude. Excel assumes that the topics and cognitive demands are continuous variables though both are nominal. The content maps are therefore correct only at the intersections of topics and cognitive demands. The content maps provide a visual record of the content contained in the particular standards document or assessment that can be used to compare the content of standards or assessments within or between states. We use content maps rather than other graphic displays because experience has shown they are easier to read and understand.

The analyses used to address the two research questions are descriptive analyses of SEC data. For Question 1, alignment indices are calculated for each available standard-assessment pair. Where possible, these alignment indices are compared with theoretical maximum indices and also with one another to illustrate the typical degree of standard-test alignment across states. For Question 2, SEC data are used to calculate several indices that describe the nature of nonperfect alignment. The results indicate not only the degree of test-standard alignment within and between states, but also the typical nature of misalignment.

Results

Research Question 1

The first research question asks the extent to which standards and assessments under NCLB are aligned with one another. The results are presented in Table 2. In ELAR, the overall average test-standards alignment index is .19, indicating that the average within–grade level test-standards

Table 2
Alignment of State Standards and Assessment by Grade and Subject

Grade	*N*	*M*	*SD*	Minimum	Maximum
ELAR					
3	6	.18	.04	.13	.22
4	9	.19	.08	.05	.36
5	8	.15	.04	.07	.18
6	7	.18	.06	.09	.25
7	5	.22	.07	.13	.29
8	9	.17	.07	.10	.29
9-12	9	.23	.06	.10	.30
Total	53	.19	.06	.05	.36
Mathematics					
3	8	.29	.09	.19	.44
4	8	.32	.08	.20	.43
5	6	.30	.07	.21	.37
6	8	.25	.09	.11	.38
7	7	.21	.06	.12	.29
8	13	.27	.10	.05	.42
9-12	10	.30	.15	.01	.47
Total	62	.27	.10	.01	.47
Science					
4-5	9	.26	.04	.19	.33
7-8	9	.24	.07	.16	.35
9-12	5	.29	.07	.22	.37
Total	23	.26	.06	.16	.37

Note. Index is traditional alignment index (Porter, 2002) indicating proportion of content in common. ELAR = English language arts and/or reading.

pair share 19% of their content. In other words, 19% of the standards' content is in perfect proportional agreement at the topic-by-cognitive demand level with 19% of the assessments' content. The indices range from .05 to .36 across states and grades, with the maximum in fourth grade for State AA and the minimum for State M in Grade 4. There are small and inconsistent differences in average alignment across grades for ELAR.

The average alignment indices are somewhat higher in mathematics (.27) and science (.26) than in ELAR. Mathematics, in addition to having a higher average alignment, has more spread, the highest maximum, and the lowest minimum, with alignment indices ranging from .47 for Algebra I in State K to .01 for Grade 10 in State AA. In contrast, the narrowest range of alignment indices is in science, with an average of .26, a minimum of .16 for State I Grade 8, and a maximum of .37 for State K biology. In short, the alignment of state standards with assessments of student achievement is typically in the range .20 to .30. However, there are some states with alignment

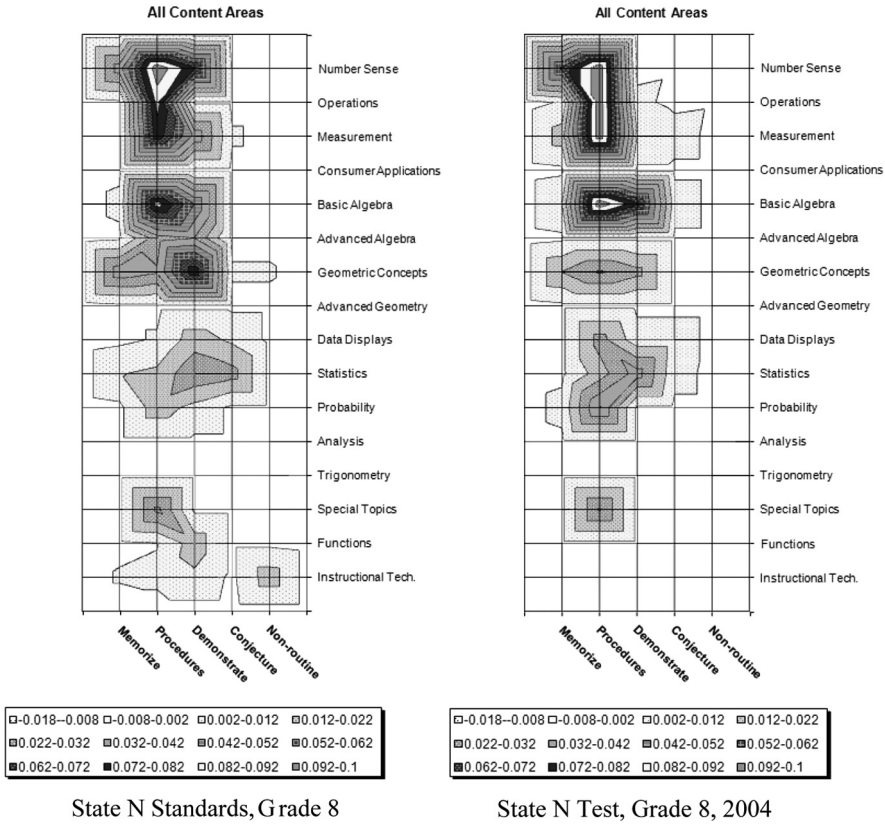

Figure 1. **Comparison of standards and assessment for "average" aligned state in mathematics, coarse-grained maps, alignment = .27.**

indices below .10 and above .35, as well as standard deviations between .06 and .10, depending on subject and grade, indicating some degree of variability in test-standards alignment across states.

What does it mean to have an average alignment of .27 for mathematics, .26 for science, and .19 for ELAR? There are at least two ways to provide some clarity on the meaning of the alignment index. One is to present the content maps for a "typical" standards-assessment pair with an alignment index near the average. These maps give a visual representation of the sources of alignment and misalignment. Example maps are presented in Figures 1 and 2. Both maps compare the content of standards and assessments for a pair of documents at the average level of alignment in mathematics—.27.

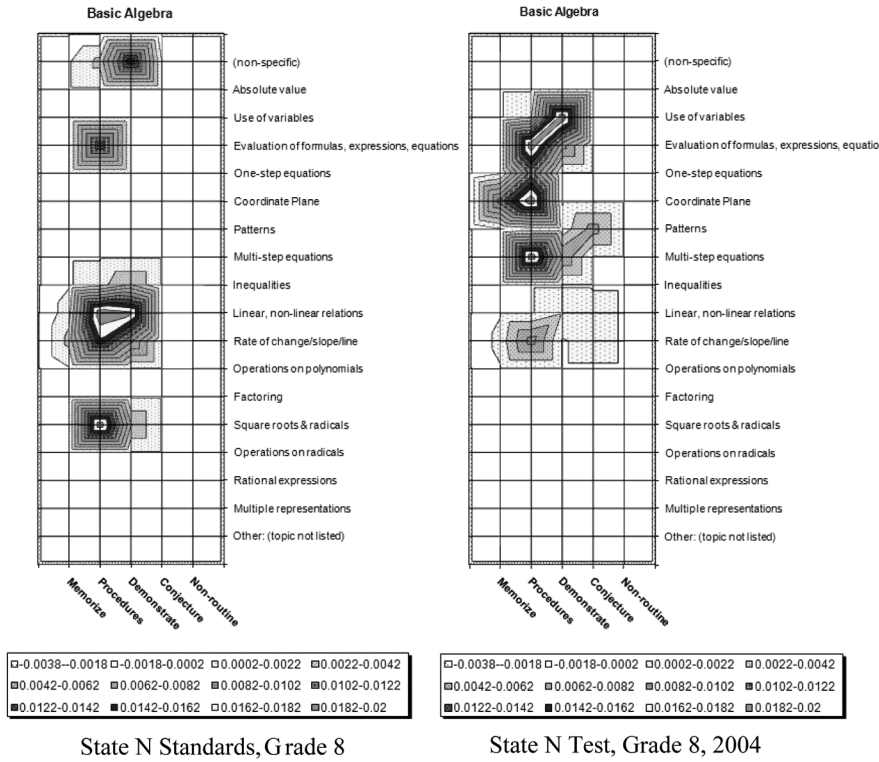

Figure 2. **Comparison of standards and assessment for ''average'' aligned state in mathematics, fine-grained maps for Basic Algebra, alignment = .27.**

The state is State N, eighth grade. The map in Figure 1 is a coarse-grained map—the topics on the y-axis are the broad content areas on the SEC math survey (e.g., number sense, measurement). These maps give a picture of agreement at the more general level of content specificity; thus, because alignment is calculated at the fine-grained level of detail, these maps may overstate alignment. Figure 1 illustrates that both the test and standards have major focuses on Number Sense, Operations, Measurement, and Basic Algebra, all at the level of Perform Procedures, and moderate emphasis on Number Sense and Geometric Concepts at the level of Demonstrate Understanding. The test mirrors the emphases on Perform Procedures for each content area, but it does not emphasize Geometric Concepts as much as the standards do, nor does it capture Number Sense at the level of Demonstrate Understanding. In general, however, the two maps in

Figure 1 suggest that the standards and assessment, even though aligned only .27, have substantial overlap.

It is important, therefore, to examine the finer-grained maps in order to see alignment at the level at which it is calculated. Figure 2 shows fine-grained maps for Basic Algebra, a topic covered substantially in both standards and assessments in State N Grade 8 mathematics. Here it is possible to see the differences between the content in the standards and the content that is assessed. The major focuses of the standards are Evaluation of Formulas/Expressions/Equations, Linear/Non-Linear Relations, Rate of Change/Slope/Line, and Square Roots & Radicals, all at the level of Perform Procedures, as well as Linear/Non-Linear Relations at the level of Demonstrate Understanding. Of these five focal cells, only Evaluation of Formulas/Expressions/Equations at the level of Perform Procedures is also a focal topic on the assessment. In contrast, the assessment tests students on Multi-Step Equations and Coordinate Planes at the level of Perform Procedures and Use of Variables at the level of Demonstrate Understanding, topics that are not included at all in the standards document. From this pair of maps, it is clear that the source of misalignment for this particular state and grade in mathematics is not merely that the test samples from the domain of the standards, but also that the test assesses students on content that is not specified in the standards. Space does not allow showing maps of all pairs of documents in the database. However, it should be noted that the pair shown in Figures 1 and 2 is an average pair for mathematics, with an alignment of .27—there are many pairs of documents with alignments much lower and much higher than this figure.

A second way to interpret values of the alignment index is against maximum possible values (Barghaus, Porter, & Polikoff, 2010). While it is true that the theoretical maximum alignment value is defined to be 1.00, the actual maximum alignment is lower. This is because there are a finite number of test items, content analysts, and SEC cells allowed per test item. For each "target" in the database (e.g., set of content standards), it is possible to calculate a maximum possible alignment given the number of items on the test, the typical number of cells each item is placed into, the number of content analysts, and the rate of agreement among content analysts. Briefly, the process is as follows, beginning with the content analysis of the standards as the target and the number of items on the test as the constraint. If there are x items on the test, each item counts for 1/x of the test's content. The first (hypothetical) item is chosen to maximally align to the target standards (e.g., on a 20-item test, the item is chosen that would have the greatest alignment index were an alignment analysis done between the target and the 5% of test content represented by the item). Then the proportion of the standards content that is aligned with the first item is subtracted from the content analysis of the standards and a second (hypothetical) item is chosen, again to maximally align with the remaining standards content. This

process is completed x times until there is an x-item test, each item selected to maximally align to the target standards. If hypothetical items are used, the test is "maximally aligned," and the alignment index comparing the test to the standards is the maximum possible alignment index. If a real item bank is used, we have constructed the maximally aligned assessment given the available items. For more details on the procedure, see Barghaus and colleagues (2010).

As an example, State AC's fourth-grade ELAR test and standards are aligned .17, slightly below the ELAR average of .19. State AC's multiple-choice test has 61 items. Assuming *n* content analysts with complete agreement on each test item as to placement in the SEC framework and placement in exactly one SEC cell per item (the number of content analysts is inconsequential if there is complete agreement), the maximum possible alignment of a test to State AC's fourth-grade ELAR standards is .69. Complete agreement and one SEC cell per item are the most conservative assumptions; hence, this is the lowest maximum alignment possible for this test. In this case, the renormalized alignment index would be .19/.69, or .28. However, these assumptions are unreasonably conservative, as few items are placed in only one cell by each of four raters, and fewer still have complete agreement on the target for each rater. If the assumptions are relaxed so that the raters place the items in two cells each or three cells each but still have complete agreement, the maximum alignments go to .83 or .85, respectively. The maximums increase further still if complete agreement is not required or if there are more items on the test. In the case of shorter tests, such as State I's fourth-grade ELAR assessment, which has 36 items, the maximum alignments are .53, .69, and .78 for content analysts with complete agreement and placement into one, two, and three cells, respectively. Certainly, the maximum alignment indices are never equal to 1.00, but under realistic assumptions about interrater agreement and the number of cells per item, it is clear that the actual alignment indices for tests and standards are far from the maximums possible.

Research Question 2

Having identified that the alignment of tests to standards is far from the maximum across subjects and grades, the second research question asks about the nature of the misalignment. To address this question, we break out the misaligned content for standards and assessments into three categories each. First, we treat the standards as the target and calculate the proportion of test content that falls in each of the following mutually exclusive categories:

1. For each cell that is included in both the standards and the assessments, the proportion of test content on each cell up to but not exceeding the proportion

of standards content on the corresponding cell. The sum of the proportions on the test for these cells equals the alignment index.
2. For each cell that is included in both the standards and the assessments, the surplus proportion of test content on each cell relative to the proportion of standards content on that cell. The sum of these proportions is the amount of content that is *over-tested* relative to the standards.
3. Test content on topics that are also included in the standards, but where the cognitive demand level of the topics differs between the assessments and the standards.
4. Test content on topics that are not at all included in the standards. This could also be called the proportion of test content that is *completely misaligned* with the standards.

Each of the four components is a proportion of total test content. The sum of the proportions across the four components is 1.0.

Next, we treat the assessments as the target, similar to Webb's range-of-knowledge criterion, and calculate the proportion of standards content in each of the following categories:

1. For each cell that is included in both the standards and the assessments, the proportion of standards content on each cell up to but not exceeding the proportion of test content on the corresponding cell. The sum of the proportions on the standards for these cells equals the alignment index—the same as Index 1 previously described.
2. For each cell that is included in both the standards and the assessments, the surplus proportion of standards content on each cell relative to the proportion of test content on that cell. The sum of these proportions is the amount of standards content that is *under-tested* relative to the test.
3. Standards content on topics that are also included on the assessment, but where the cognitive demand level of the topics differs between the standards and assessment.
4. Standards content on topics that are not at all included on the assessment. This could also be called the proportion of standards content that is *completely misaligned* with the test.

Each of the four components is represented as a proportion of total standards content. The sum of the proportions across the four components is 1.0.

To illustrate how these proportions are calculated, consider the example standards and assessment data shown in Figure 3. The values in each cell are proportions of total content on each document in each SEC cell. For this standards-assessment pair, the value of Index 1 (the alignment index) would be .42 (.25 from Adding Fractions level B, .17 from Adding Fractions level C), the value of Index 2 would be .13 (from Adding Fractions level C), the value of Index 3 would be .40 (from Adding Fractions level D), and the value of

Standards	Level B	Level C	Level D	Level E	Level F
Adding fractions	0.33	0.17	0	0.46	0
Subtracting fractions	0	0	0	0	0
Multiplying fractions	0.04	0	0	0	0

Assessment	Level B	Level C	Level D	Level E	Level F
Adding fractions	0.25	0.30	0.40	0	0
Subtracting fractions	0.05	0	0	0	0
Multiplying fractions	0	0	0	0	0

Figure 3. **Example standards and assessment data.**

Index 4 would be .05 (from Subtracting Fractions). For the standards, the values of the four indices would be .42, .08, .46, and .04, respectively.

Nature of misaligned test content. Table 3 categorizes the typical test's content by whether it is aligned to the standards or not and, if not, what the nature is of the misalignment. The leftmost columns with proportions are the alignment indices, which are the same as in Table 2. Moving right, it is apparent that the largest single type of misalignment in ELAR is due to over-testing. The same is true in mathematics and science, but to a smaller extent. The proportion of test content that over-tests cells that are covered in the standards is 38% in ELAR, 31% in mathematics, and 28% in science. The value of this proportion ranges from 76% in State I Grade 6 ELAR to 0% in State AA Grade 6 mathematics. There are no apparent patterns across grades. When the values for over-testing are added to the values for alignment, it indicates that 54% to 57% of test content on average, depending on subject, is on SEC cells that are also included in the content standards.

The next two columns to the right indicate the proportion of test content on topics from the standards, but at the wrong levels of cognitive demand. The proportion is smallest in mathematics—just 16% of test content is on topics from the standards but at the wrong level of cognitive demand. The values are 26% in both ELAR and science. While most of the proportions are similar across grades within subjects, the proportion of test content at the wrong level of cognitive demand is higher in Grades 9 through 12 ELAR (36%) than in all other Grades 3 through 8 (24%).

The final two columns indicate the proportion of test content that is completely misaligned with standards (i.e., covering topics not at all included in the standards). The average proportion is 17% for ELAR, 27% for mathematics, and 20% for science. The value of this index ranges from a low of .01 for State Y Grade 7 ELAR to a high of .98 for State AA Grade

Table 3
Nature of Misaligned Test Content (Proportion of Test Content That Is . . .)

		Aligned With Standards		Over-Tested as Compared With Standards		At a Different Level of Cognitive Demand Than in the Standards		Completely Misaligned With Standards	
Grade	*N*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
ELAR									
3	6	.18	.04	.40	.17	.19	.10	.23	.14
4	9	.19	.08	.34	.16	.26	.12	.20	.16
5	8	.15	.04	.43	.17	.24	.14	.18	.17
6	7	.18	.06	.42	.20	.24	.11	.16	.16
7	5	.22	.07	.38	.24	.28	.10	.12	.14
8	9	.17	.07	.41	.18	.24	.13	.17	.17
9-12	9	.23	.06	.31	.08	.36	.17	.10	.10
Total	53	.19	.06	.38	.17	.26	.14	.17	.15
Current	44	.20	.06	.42	.15	.26	.12	.12	.10
Mathematics									
3	8	.29	.09	.32	.16	.17	.08	.22	.21
4	9	.32	.08	.31	.10	.18	.09	.21	.17
5	7	.30	.07	.33	.16	.16	.05	.23	.15
6	8	.25	.09	.28	.14	.18	.13	.30	.23
7	7	.21	.06	.24	.15	.16	.09	.40	.18
8	13	.27	.10	.28	.15	.16	.11	.29	.22
9-12	10	.30	.15	.39	.10	.10	.07	.25	.27
Total	62	.27	.10	.31	.14	.16	.09	.27	.21
Current	50	.27	.10	.29	.15	.15	.09	.29	.22
Science									
4-5	9	.26	.04	.31	.15	.22	.09	.21	.15
7-8	9	.24	.07	.26	.12	.31	.14	.19	.15
9-12	5	.29	.07	.25	.15	.26	.13	.21	.19
Total	23	.26	.06	.28	.13	.26	.12	.20	.15
Current	17	.25	.06	.26	.14	.26	.08	.23	.16

Note. Values are proportions of total test content. Proportions may not add to 1 within rows due to rounding. Rows labeled *Current* includes standards that are in place in 2008-2009. ELAR = English language arts and/or reading.

10 mathematics. Clearly there is a wide range in the proportion of test content that is completely misaligned with standards content. While the average proportions are not large, it is important to remember that these values, when added with the proportions representing testing at the wrong level of cognitive demand, comprise an average of approximately 45% of the content on the typical test. Furthermore, when alignment is defined in the way that has been shown to be most predictive of student learning gains (i.e., perfect agreement at the topic-by-cognitive demand level as in Gamoran

Table 4
Nature of Misaligned Standards Content (Proportion of Standards Content That Is . . .)

		Aligned With Tests		Under-Tested on The Tests		Tested at a Different Level of Cognitive Demand		Completely Misaligned With Tests	
Grade	*N*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
ELAR									
3	6	.18	.04	.11	.05	.13	.06	.58	.08
4	9	.19	.08	.12	.10	.20	.14	.48	.14
5	8	.15	.04	.07	.03	.17	.10	.60	.09
6	7	.18	.06	.12	.08	.19	.10	.51	.11
7	5	.22	.07	.19	.14	.15	.07	.43	.18
8	9	.17	.07	.14	.12	.12	.05	.56	.14
9-12	9	.23	.06	.16	.09	.18	.08	.44	.13
Total	53	.19	.06	.13	.09	.17	.09	.52	.14
Current	44	.20	.06	.12	.09	.16	.06	.52	.14
Mathematics									
3	8	.29	.09	.21	.14	.20	.12	.30	.07
4	9	.32	.08	.28	.05	.18	.06	.24	.07
5	7	.30	.07	.26	.10	.18	.07	.28	.12
6	8	.25	.09	.18	.07	.19	.09	.38	.12
7	7	.21	.06	.18	.12	.16	.09	.45	.15
8	13	.27	.10	.18	.09	.15	.08	.40	.17
9-12	10	.30	.15	.23	.12	.17	.05	.31	.22
Total	62	.27	.10	.21	.10	.17	.08	.34	.15
Current	50	.27	.10	.22	.10	.17	.07	.33	.16
Science									
4-5	9	.26	.04	.24	.12	.29	.13	.21	.09
7-8	9	.24	.07	.27	.13	.23	.06	.26	.11
9-12	5	.29	.07	.28	.18	.23	.10	.21	.17
Total	23	.26	.06	.26	.13	.25	.10	.23	.11
Current	17	.25	.06	.31	.12	.24	.07	.20	.11

Note. Values are proportions of total standards content. Proportions may not add to 1 within rows due to rounding. Rows labeled *Current* includes only standards that are in place in 2008-2009. ELAR = English language arts and/or reading.

et al., 1997), the proportions of misalignment are closer to 75% to 80%. In either of these metrics, moderate to large proportions of the typical state test assess content that is not represented correspondingly in state standards. It is only when cognitive demand levels are ignored that agreement increases to around .8.

Nature of misaligned standards content. Table 4 categorizes the content in state standards by whether it is aligned to the corresponding tests. As in

Table 3, the leftmost columns repeat the traditional alignment index. The next pair of columns indicates the proportion of standards content that is under-tested on the assessments. These proportions are lowest in ELAR (13%) and are somewhat larger in mathematics (21%) and science (26%). In short, while 19% to 27% of the content in state standards is in perfect alignment with state tests, an additional 13% to 26% of standards content, depending on subjects, is tested but not enough relative to the emphasis the content receives in the standards. When the proportion indicating perfect alignment is added to the proportion indicating under-testing, we see that 32% of content in the typical ELAR standards corresponds to SEC cells that are at least marginally tested, as compared to 48% in mathematics and 52% in science. These proportions are somewhat lower, especially in ELAR, than the analogous proportions for state tests, which were each between 54% and 57%. Thus, the proportion of standards content that is tested is typically smaller than the proportion of test content that is included in the standards.

The next two columns indicate the proportions of standards content that is on topics that are also tested, but at the wrong levels of cognitive demand. Approximately one-sixth of the content in ELAR (17%) and mathematics (17%) standards is tested at the wrong level of cognitive demand. By comparison, one-quarter (25%) of the content in science standards is tested at the wrong level of cognitive demand. There are no apparent patterns across grades as to the proportion of standards content that is tested at the wrong level of cognitive demand.

The final two columns indicate the proportion of standards content on topics that are not tested at all. Of course, part of these proportions could be attributed to variation in content coverage of the assessments across years. Unfortunately, we do not have data for multiple forms of the same assessment, so the results presented here indicate the results when state standards are compared with an individual test form. There are large differences across subjects, with values of 52% for ELAR, 34% for mathematics, and 23% for science. These proportions can also be compared with the corresponding proportions from Table 3. The comparison illustrates that in ELAR and mathematics, the proportion of standards content on topics that are not tested is larger than the proportion of test content on topics that are not included in the standards. In contrast, in science, approximately 20% of both standards and test content is misaligned on topic. Another way to examine the size of misalignment is to add these proportions to the proportions for right topics/wrong cognitive demands; those values are 69% in ELAR, 51% in mathematics, and 48% in science.

In general, half the content in mathematics and science standards and two-thirds in ELAR is misaligned with test content, meaning that the test does not assess substantial proportions of the topics specified in the standards. One reason for this finding may be that states want to report scores at the subscale level; that is, assessments may include a few more items on

Table 5

Cognitive Demand Marginal Proportions for Tests and Standards in Each Subject

Grade	*N*	Level B		Level C		Level D		Level E		Level F	
		Standard	Test	Standard	Test	Standard	Test	Standard	Test	Standard	Test
ELAR											
3	6	.17	.40	.31	.23	.34	.11	.13	.23	.05	.03
4	9	.12	.33	.30	.18	.35	.17	.16	.29	.07	.04
5	8	.17	.37	.19	.15	.29	.11	.22	.33	.13	.04
6	7	.12	.31	.25	.23	.37	.11	.19	.33	.07	.02
7	5	.10	.24	.23	.34	.37	.16	.22	.22	.09	.03
8	9	.12	.30	.21	.17	.35	.14	.21	.34	.11	.05
9-12	9	.03	.23	.13	.24	.45	.22	.27	.27	.11	.04
Total	53	.12	.31	.23	.21	.36	.15	.20	.29	.09	.04
Mathematics											
3	8	.19	.18	.48	.60	.27	.14	.05	.07	.01	.01
4	9	.13	.25	.58	.52	.21	.16	.06	.06	.02	.02
5	7	.12	.21	.54	.63	.23	.11	.06	.04	.05	.01
6	8	.15	.16	.46	.62	.24	.15	.11	.07	.04	.01
7	7	.14	.13	.46	.68	.23	.12	.13	.05	.05	.03
8	13	.12	.12	.48	.63	.23	.16	.10	.07	.07	.02
9-12	10	.10	.10	.57	.76	.19	.08	.10	.05	.05	.01
Total	62	.13	.16	.51	.63	.23	.13	.09	.06	.04	.01
Science											
4-5	9	.25	.51	.23	.09	.39	.14	.11	.21	.03	.05
7-8	9	.21	.43	.18	.10	.43	.16	.14	.26	.05	.06
9-12	5	.22	.39	.15	.15	.41	.17	.16	.20	.06	.10
Total	23	.23	.45	.19	.10	.41	.15	.13	.23	.04	.07

Note. Values are proportions of total content. Proportions may not add to 1 within rows due to rounding.

algebra than is merited by algebra's proportional representation in the standards because otherwise there would not be enough algebra items to report a reliable score for that subscale. Nevertheless, while there is no set standard for how much of the standards content should be tested on any particular test, there is certainly room for improvement, insofar as the 17% of standards content in mathematics and ELAR and 25% in science that is currently tested at the wrong levels of cognitive demand could be redirected toward increased alignment without affecting test length. And except for in ELAR, the content analysis data suggest that a general lack of test space is not the cause of the high proportion of untested content. This is evidenced by a comparison of the total number of SEC cells covered in the standards and assessments. In mathematics, the standards cover an average of 113 SEC cells, only slightly more than the 97 for the tests. In science, the numbers

are even closer—119 for the standards, 117 for the tests. In contrast, in ELAR the average standards document covers 177 cells as compared to 79 for the tests. Clearly, the standards in ELAR are more diffuse than in mathematics and science—they cover more SEC cells—but the tests are more focused. This appears to be a contributing factor to the greater misalignment for ELAR.

Distribution of cognitive demand for standards and tests. Finally, it is possible to examine the distribution of emphasis across levels of cognitive demand. The marginal proportions of total content at each of the five levels of cognitive demand for both standards and assessments are compared in Table 5. Because these are marginal proportions, they indicate a much greater degree of alignment than was found based on the alignment indices. However, tests and standards that are 100% aligned at the marginal for cognitive demand might cover dramatically different topics and thus be quite poorly aligned at the level that predicts value-added to student achievement. Previous research has concluded that state assessments tend to focus more on procedural and other lower level cognitive demand material than is true in the standards (e.g., Resnick et al., 2003; Webb, 1999).

The first pattern that emerges from results in Table 5 is that for mathematics tests and standards and science standards, the two highest levels of cognitive demand are not emphasized much. ELAR standards and assessments and science assessments place considerably more emphasis on the two highest levels of cognitive demand, primarily because all place greater emphasis on cognitive demand level E. For the two highest levels of cognitive demand, the difference in emphasis for test versus standards is not large. For science and ELAR, the test places a greater emphasis on the two highest levels of cognitive demand (30% in science, 33% in ELAR) than do the standards (17% in science, 29% in ELAR). For mathematics, the standards place a higher emphasis on the two highest levels of cognitive demand (13% for standards and 7% for tests). Only the difference in science is large, however.

Memorization does generally show less emphasis in the standards and more on the tests, especially for ELAR and science, though in mathematics the difference in emphasis on memorization between test and standards is trivial. The second level of cognitive demand is emphasized more on the test than the standards in mathematics, more on the standards than the test in science, and roughly the same between standards and test in ELAR. The third level of cognitive demand uniformly finds greater emphasis in the standards than on the test. In ELAR, 36% of the standards versus 15% of the test is on generate/create/demonstrate. In mathematics, 23% of the standards versus 13% of the test is on demonstrate understanding and for science, 41% of the standards but only 15% of the test is on communicate understanding. Whether this third level of cognitive demand should be considered higher or lower level cognitive demand is debatable. Overall, the

results of the cognitive demand analyses indicate that the aggregate cognitive demand levels of the test is on average somewhat lower than the cognitive demand level of the assessments for ELAR and mathematics while there is no difference for science. Even for ELAR and mathematics, the differences were not as large as were expected, given the results of previous studies.

Summary and Conclusions

State content standards and corresponding assessments of student achievement are the foundations upon which the current system of standards-based accountability in U.S. education is built. In the initial formulation of standards-based reform, as well as in the No Child Left Behind law that codified standards-based reform in federal law, the coherence of standards and assessments was seen as paramount to ensuring the validity of the interpretations made from student test scores. There have been a number of alignment procedures developed to measure the extent to which the standards and assessments are, in fact, coherent. One of them, the Surveys of Enacted Curriculum, allows for the comparison of the content of standards and assessments on a common content language. Data from SEC content analyses of standards and assessments in 19 states were used here to investigate whether state standards and assessments under NCLB are in fact aligned and, to the extent that they are not, what the nature of the misalignment is.

Results indicate that the answer to "are standards and assessments aligned with one another?" depends on the definition of alignment. Clearly, when alignment is defined in the way that is most predictive of value-added to student achievement (Gamoran et al., 1997), the answer is no. Average alignment indices for state standards and assessments were below .30 in mathematics and science and below .20 in ELAR. No alignment index was greater than .50 for any state, grade, or subject analyzed. When the definition of alignment was relaxed to account for content on the assessment that tested topics and cognitive demand levels mentioned at all in the standards (or to account for content in the standards that was tested at all), alignment values jumped to approximately .50 or higher in all cases but one. For ELAR assessments, mathematics standards and assessments, and science standards and assessments, roughly half of their total content is on SEC cells that are at least mentioned in their corresponding document. The same was not true for ELAR standards; only one-third of the content in ELAR standards is tested at all. In any case, a significant and sometimes large proportion of misalignment is due to over- or under-testing content relative to its proportion in the standards.

More complete misalignment is when the standards and assessments do not agree on the cognitive demand levels, topics, or both to be tested. Across subjects and grades, this form of incoherence comprised moderate to large

proportions of the total content targeted. Substantial proportions of the topics specified in the standards were not tested at all, especially in ELAR. Perhaps more alarmingly, roughly a quarter of the content on typical tests was not reflective of topics that were mentioned at all in their corresponding grade-level standards.

One hypothesis for alignment levels far from their respective maximums is that assessments measure some content from previous grades, in addition to their specified grade. This might be expected for grade-level specific tests, because proficiency assessments must measure a range of ability levels. To investigate this possibility, we conducted an additional analysis, where we first aggregated the standards and assessments in mathematics and ELAR across Grades 4 through 8. Then we computed alignment indices on the aggregated standards and assessments. For the five states in ELAR where such data were available, the average alignment of aggregated standards with aggregated assessments was .28, as compared to the grade-specific alignment of .18. For the six states in mathematics where such data were available, the average alignment of aggregated standards with aggregated assessments was .43, as compared to the grade-specific alignment of .26. These analyses supported the hypothesis that there is a nontrivial degree of testing standards content at the wrong grades. While perhaps not as problematic as testing content that is never included in the standards at any grade, it is nonetheless problematic for grade-specific tests to assess students on content that is not specified in the corresponding grade-specific standards. This kind of misalignment sends confusing messages to teachers about the content to teach in their particular grade and could contribute to the repetitiveness of the U.S. curriculum in the core subjects.

Another hypothesis for the finding of far below maximum alignment is that certain content is systematically over-tested or under-tested across states and possibly grades, perhaps because some content is more difficult to assess. To investigate this possibility, for each pair of documents, we subtracted the matrix representing the assessment from the matrix representing the standards, leaving us with a matrix indicating the amount of over- or under-testing, cell by cell. Then we averaged across states within grades and subjects, determined the most over-tested and under-tested SEC cells in each content area and grade, and compared these averages across grades, as well as comparing individual states with the averages. We found a high degree of consistency in what was over- or under-tested across grades and states. For instance, in ELAR, at least one cell representing the topic of Comprehension of main ideas, key concepts, and sequences of events was among the most or second most over-tested cells at all grades. Of the 53 pairs of documents in ELAR, 44 had at least one cell in their top five over-tested that was in the topic of Comprehension of main ideas, key concepts, and sequences of events. An example of an under-tested topic was Nature of scientific inquiry/method; a cell from this topic was the most under-tested at

all grades and was in the top five most under-tested cells in 12 of 23 document pairs. Similar degrees of consistency in over- and under-tested topics and SEC cells were found in all subjects, suggesting that systematic under- or over-coverage of particular topics may be driving a portion of the misalignment.

A third hypothesis as to the modest levels of alignment is that the standards in our database that were content analyzed earlier in period 2002-2009 "dragged down" the averages. That is, as NCLB rolled out, states became more attentive to issues of alignment, and thus, alignment indices from more recent standards and assessments would be higher than alignment from older standards and assessments. To investigate this possibility, we calculated the mean for each of the four indices for standards and assessments for *only those standards and assessments that were still in use as of the 2008-2009 year*; the results are presented in Tables 3 and 4 in the bottom row of each section. There are no substantial differences between the means from older documents and the means from newer documents, with the possible exception of for the comparison of ELAR tests to standards. For that comparison, there is approximately 5% less completely misaligned content on the tests relative to the standards for the more recent documents as compared to the overall average. In short, there is little evidence that the alignment of tests and standards has improved markedly over the study period.

The results described here suggest a number of obvious ways to improve alignment, and the notion of coherence suggests that alignment should be improved. Each of the three approaches described in the following should be especially appropriate as standards including the new Common Core State Standards begin to reach a steady, mature state. One approach is to eliminate all complete misalignment on the tests—those test items that test topics that are simply not covered in the standards at all. While most tests have relatively small proportions of content in complete misalignment with standards, few have no such content, and some have at least half of test content in complete misalignment. States and test developers/vendors could come together and eliminate the roughly 20% of test content on material not in the standards, replacing it with material from the standards.

To deal with complete misalignment on the standards, a second solution would be to build test forms that sample from the domain of the standards, such that all content in the standards is tested, if not on every form or in every year, then across forms and years in proportion to its relative importance in the standards. One approach that could be used is an algorithm created by Barghaus and colleagues (2010), which allows test developers to maximize alignment to a target, given a pool of content-analyzed items. The intention should be to encourage teachers to focus their instruction on the standards, which represent the domain of intended content. Forms across years should increasingly approach the standards domain. There

may be psychometric tradeoffs in such an approach, such as form equivalency issues, that would need careful investigation.

A third approach is to focus on the sources of misalignment identified in this article in improving the alignment of test items to standards. For instance, it is apparent that in all subjects and grades, roughly 20% to 30% of test content is on topics in the standards but at different levels of cognitive demand than is found in the standards. Test developers should work harder to match the level of cognitive demand of items to the ways in which topics are specified in the standards.

In addition to these implications for test developers and state assessment leaders, this work should inform policymakers. The results clearly indicate that the standards and assessments in the observed states are not as well aligned as they could or were intended to be. When thinking about the effects of standards-based reform on teacher practice and student achievement, the level of coherence should be taken into account. It may be difficult to evaluate the impact of standards and aligned assessments, because few states appear to have created such a system. Better examples of well-aligned standards-based reform state systems are needed than were found here to accurately evaluate the impact of standards and aligned assessments. Unfortunately, policymakers may see modest results of standards-based reform and interpret those results to mean that the reform does not work. Rather, it appears at least as likely that standards-based reform has not been implemented as designed and that the lack of implementation could be a major contributor to any finding of no or small effects. Furthermore, as policymakers consider implementing teacher merit pay systems based in part on value-added to student achievement, they should be sure that the value-added assessments are appropriate representations of the content teachers were supposed to teach. To assign teacher effectiveness ratings on the basis of content teachers were not supposed to teach would be unfair and would undermine the legitimacy of the system.

Finally, these results present many opportunities for needed future research. Certainly, one important avenue is to investigate the alignment methodologies themselves. No study to date has directly compared the three primary alignment indices, and such a study would provide a major contribution to the understanding of alignment. More work must be done on methods of developing aligned assessments. Barghaus and colleagues' (2010) work represents a good first step, and it would be useful to see an application of this or other alignment methodologies in the creation of actual state student achievement tests. On a related point, it would be useful to have guidelines as to just how well aligned standards and assessments should be. Here, again, the work of Barghaus and colleagues is instructive. The goal should be maximizing alignment given test length and number of raters. Furthermore, alignment should approach 1.0 between assessments aggregated across years and standards.

Even if the coherence of standards-based reform is improved, and tests are well aligned with standards, there will be questions about the extent to which those tests accurately reflect the quality and content of instruction received. Instructional sensitivity is an important reemerging issue in the development of criterion-referenced tests (Polikoff, 2010; Popham, 2007), and research investigating the interplay between alignment and sensitivity would further the field. Furthermore, it is important to investigate the effects of standards and assessments on teachers' instruction. Few if any studies have investigated the ways in which teachers have modified the content of their instruction without simply asking teachers whether their instruction was aligned (e.g., Hamilton & Berends, 2006; Koretz, Barron, Mitchell, & Stecher, 1996; Koretz, Mitchell, Barron, & Keith, 1996; Stecher & Chun, 2001). Given the wide variation in alignment of standards and assessments, it would be helpful to know whether those alignment differences lead to differential impacts on instruction.

Whether student achievement tests are to be used for high-stakes decisions, such as hiring and firing of teachers, or lower-stakes decisions, such as curriculum revision, it is important that the tests accurately reflect student knowledge of the domain the tests are intended to assess. To have assessments that are not well aligned to the standards that are the foundation of the U.S. curriculum is unfair to teachers and students. Especially as there is movement toward common core national standards, developing assessments that can provide information on progress toward achievement of those standards is of paramount importance. The research described here suggests that there is some distance to go toward achieving the commonsense goal of fair, aligned assessments. But also, there are clear, proven strategies for improving alignment that should be followed. We hypothesize that the success of standards-based attempts to improve student outcomes depends on these improvements.

References

Barghaus, K. M., Porter, A. C., & Polikoff, M. S. (2010, April). *Constructing aligned assessments for middle school science students and teachers*. Paper presented at the annual meeting of the American Educational Research Association, Denver, CO.

Chatterji, M. (2002). Models and methods for examining standards-based reforms and accountability initiatives: Have the tools of inquiry answered pressing questions on improving schools? *Review of Educational Research*, *72*(3), 345–386.

Clune, W. H. (1993). Systemic educational policy: A conceptual framework. In S. H. Fuhrman (Ed.), *Designing coherent educational policy* (pp. 125–140). San Francisco, CA: Jossey-Bass.

Clune, W. H. (2001). Toward a theory of systemic reform: The case of nine NSF statewide systemic initiatives. In S. H. Fuhrman (Ed.), *From the capitol to the classroom: Standards-based reform in the state. One hundredth yearbook of the*

National Society for the Study of Education, Part II (pp. 13–38). Madison, WI: University of Chicago Press.

Cohen, D. K. (1989). Teaching practice: Plus ca change. In P. W. Jackson (Ed.), *Contributing to educational change: Perspectives on research and practice* (pp. 27–84). Berkeley, CA: McCutchan.

Cohen, D. K. (1995). What is the system in systemic reform? *Educational Researcher*, *24*(9), 11–17.

Cohen, D. K., & Spillane, J. P. (1992). Policy and practice: The relations between governance and instruction. *Review of Research in Education*, *18*, 3–49.

Cuban, L. (1984). *How teachers taught: Constancy and change in American classrooms 1890-1980*. New York, NY: Longman.

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

Goodlad, J. I. (1984). *A place called school*. New York, NY: McGraw-Hill.

Hamilton, L. S., & Berends, M. (2006). *Instructional practices related to standards and assessments* (RAND Working Paper WR-374-EDU). Santa Monica, CA: RAND.

Jennings, J. H. (1998). *Why national standards and tests? Politics and the quest for better schools*. Thousand Oaks, CA: Sage.

Kane, M. T. (2008). Terminology, emphasis, and utility in validation. *Educational Researcher*, *37*(2), 76–82.

Koretz, D. M., Barron, S. I., Mitchell, K. J., & Stecher, B. M. (1996). *Perceived effects of the Kentucky Instructional Results Information System*. Santa Monica, CA: RAND.

Koretz, D. M., Mitchell, K. J., Barron, S. I., & Keith, S. (1996). *Final report: Perceived effects of the Maryland School Performance Assessment Program*. Los Angeles, CA: CRESST.

Martone, A., & Sireci, S. G. (2009). Evaluating alignment between curriculum, assessment, and instruction. *Review of Educational Research*, *79*(4), 1332–1361.

McLaughlin, M. W. (1990). The RAND Change Agent Study revisited: Macro perspectives and micro realities. *Educational Researcher*, *19*(9), 11–16.

National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics*. Reston, VA: Author.

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

O'Day, J. A., & Smith, M. S. (1993). Systemic reform and educational opportunity. In S. H. Fuhrman (Ed.), *Designing coherent education policy* (pp. 250–312). San Francisco, CA: Jossey-Bass.

Polikoff, M. S. (2010). Instructional sensitivity as a psychometric property of assessments. *Educational Measurement: Issues and Practice*, *29*(4), 3–14.

Popham, J. W. (2007). Instructional insensitivity of tests: Accountability's dire drawback. *Phi Delta Kappan*, *89*(2), 146–155.

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

Porter, A. C. (2006). Curriculum assessment. In J. L. Green, G. Camilli, & P. B. Elmore (Eds.), *Handbook of complementary methods in education research* (pp. 141–160). Mahwah, NJ: Lawrence Erlbaum.

Porter, A. C., Floden, R. E., Freeman, D. J., Schmidt, W. H., & Schwille, J. R. (1988). Content determinants in elementary school mathematics In D. Grouws & T. Cooney (Eds.), *Perspectives on research on effective mathematics teaching* (pp. 96–142). Reston, VA: National Council of Teachers of Mathematics.

Porter, A. C., Kirst, M. W., Osthoff, E. J., Smithson, J., & Schneider, S. A. (1993). *Reform up close: An analysis of high school mathematics and science classrooms*. Madison, WI: Wisconsin Center for Education Research.

Porter, A. C., Polikoff, M. S., Zeidner, T., & Smithson, J. (2008). The quality of content analyses of state student achievement tests and state content standards. *Educational Measurement: Issues and Practice*, *27*(4), 2–14.

Ravitch, D. (1995). *National standards in American education: A citizen's guide*. Washington, DC: Brookings Institution.

Resnick, L. B., Rothman, R., Slattery, J. B., & Vranek, J. L. (2003). Benchmark and alignment of standards and testing. *Educational Assessment*, *9*(1&2), 1–27.

Smith, M. S., & O'Day, J. A. (1991). Systemic school reform. In S. H. Fuhrman & B. Malen (Eds.), *The politics of curriculum and testing: Politics of Education Association yearbook* (pp. 233–267). Bristol, PA: Falmer Press.

Stecher, B. M., & Chun, T. (2001). *School and classroom practices during two years of education reform in Washington state*. Los Angeles, CA: CRESST.

Tyack, D. B., & Cuban, L. (1995). *Tinkering toward utopia: A century of public school reform*. Cambridge, MA: Harvard University Press.

Webb, N. L. (1999). *Alignment of science and mathematics standards and assessments in four states* (Research Monograph No. 18). Madison, WI: National Institute for Science Education.

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

Webb, N. L. (2005). *Alignment analysis of mathematics standards and assessments, Michigan, high school*. Madison, WI: Author.

Webb, N. L. (2006). *Alignment analysis of mathematics standards and assessments, Tennessee Grades 3-9*. Madison, WI: Author.

Webb, N. L. (2007). Issues related to judging the alignment of curriculum standards and assessments. *Applied Measurement in Education*, *20*(1), 7–25.

Wixson, K. K., Fisk, M. C., Dutro, E., & McDaniel, J. (2002). *The alignment of state standards and assessments in elementary reading*. Ann Arbor, MI: Center for the Improvement of Early Reading Achievement.

Manuscript received August 16, 2010
Final revision received February 11, 2011
Accepted April 10, 2011