



University of Pennsylvania
ScholarlyCommons

Publicly Accessible Penn Dissertations

Spring 5-17-2010

The Content of Instruction Under Standards-Based Reform

Morgan S. Polikoff

University of Pennsylvania, polikoff@dolphin.upenn.edu

Follow this and additional works at: <http://repository.upenn.edu/edissertations>

 Part of the [Educational Assessment, Evaluation, and Research Commons](#)

Recommended Citation

Polikoff, Morgan S., "The Content of Instruction Under Standards-Based Reform" (2010). *Publicly Accessible Penn Dissertations*. 406.
<http://repository.upenn.edu/edissertations/406>

This paper is posted at ScholarlyCommons. <http://repository.upenn.edu/edissertations/406>
For more information, please contact libraryrepository@pobox.upenn.edu.

The Content of Instruction Under Standards-Based Reform

Abstract

Standards-based reform (SBR) has been an important policy in U.S. K-12 education for at least the last 25 years. The theory of change for SBR suggests that, with coherent content standards in core academic subjects, measurable goals for student learning, and appropriate supports and accountability measures, teachers will modify their instruction to align with standards and assessments, and student learning will improve. While many studies have evaluated the effects of SBR on achievement, few have addressed the content of instruction, the primary means by which SBR is intended to have its effects. The few studies of instruction have focused mainly on pedagogy, or they have suffered from serious methodological and substantive problems that bring into question the validity of their findings.

The purpose of this dissertation is to address this gap in the literature by using survey data on the content of teachers' instruction in mathematics, science, and English language arts and reading (ELAR) to investigate changes in a) alignment to standards and assessments, b) hours of instruction, and c) emphasis on various levels of cognitive demand (e.g., memorizing, procedures) in the years 2003-2009. In addition, attributes of the standards-based accountability system are studied for their relationship with instructional alignment. The methods used include fixed-effects models and three-level hierarchical linear modeling. The results suggest that alignment to standards and assessments in mathematics increased during the years 2003-2009 by approximately .3 to .5 standard deviations. Alignment also increased to science standards at grades 3-8, but there were no consistent changes in alignment in ELAR or at other grades. Shifts of instructional time were small and generally non-significant. Teachers were found to increase their focus on lower-level thinking (e.g., procedure, memorization) in most grades and all three subjects. Generally, these shifts brought teachers' instruction into greater agreement with standards and assessments. Finally, alignment was higher when standards and assessments were well aligned with each other (average standardized effect size = $.22\sigma$), when the standards were broad and focused on many ideas ($.50\sigma$), and when the state had stronger policies emphasizing rewards and sanctions ($.12\sigma$). Implications for policy and research are discussed.

Degree Type

Dissertation

Degree Name

Doctor of Philosophy (PhD)

Graduate Group

Education

First Advisor

Andrew C. Porter

Second Advisor

Laura M. Desimone

Third Advisor

Richard M. Ingersoll

Keywords

Standards, assessment, alignment, No Child Left Behind, instruction

Subject Categories

Educational Assessment, Evaluation, and Research

THE CONTENT OF INSTRUCTION UNDER STANDARDS-BASED REFORM

Morgan S. Polikoff

A DISSERTATION

in

Education

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2010

Supervisor of Dissertation

Signature _____

Andrew C. Porter, Dean, George and Diane Weiss Professor of Education

Graduate Group Chairperson

Signature _____

Stanton E.F. Wortham, Judy and Howard Berkowitz Professor of Education

Dissertation Committee

Laura M. Desimone, Associate Professor of Education

Richard M. Ingersoll, Professor of Education and Sociology

The Content of Instruction under Standards-Based Reform

COPYRIGHT

2010

Morgan Scott Polikoff

iii

DEDICATION

To Mom and Dad.

To Rich.

To Bubba.

To Joel.

ACKNOWLEDGEMENTS

This dissertation is the culmination of four years of work with an outstanding advisor and mentor. Andy Porter has not only given me many opportunities to succeed, he has inspired me to become a top-notch researcher. I take away many things from my experience working with him—not just great colloquialisms, but also an understanding of how to do high quality research, how to write in a compelling way, and how to navigate the professoriate. It goes without saying that his guidance was integral in helping me achieve success in graduate school and beyond.

Of course, my work with Andy was really my work with Andy and Laura. I was blessed to have, for all intents and purposes, two advisors. Laura Desimone has always been willing to read my work, offer outstanding feedback, give me direction, and make me think about education policy issues in a new and different way. I thank Laura for her strong guidance and her support, especially as a member of my dissertation committee.

I also thank Richard Ingersoll, both for teaching me about the organization that is “the school,” and also for providing a unique perspective on my dissertation work. His wisdom and expertise was a critical complement to that of my other committee members, and I know that my dissertation work was much better for having him on my committee.

Of course, none of this work would be possible without the data. Thanks very much to John Smithson for providing me with the data and answering my frequent questions.

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305C050041-05 to the University of

Pennsylvania. The opinions expressed are those of the authors and do not represent views of the U.S. Department of Education.

ABSTRACT

THE CONTENT OF INSTRUCTION UNDER STANDARDS-BASED REFORM

Morgan S. Polikoff

Andrew C. Porter

Standards-based reform (SBR) has been an important policy in U.S. K-12 education for at least the last 25 years. The theory of change for SBR suggests that, with coherent content standards in core academic subjects, measurable goals for student learning, and appropriate supports and accountability measures, teachers will modify their instruction to align with standards and assessments, and student learning will improve. While many studies have evaluated the effects of SBR on achievement, few have addressed the content of instruction, the primary means by which SBR is intended to have its effects. The few studies of instruction have focused mainly on pedagogy, or they have suffered from serious methodological and substantive problems that bring into question the validity of their findings.

The purpose of this dissertation is to address this gap in the literature by using survey data on the content of teachers' instruction in mathematics, science, and English language arts and reading (ELAR) to investigate changes in a) alignment to standards and assessments, b) hours of instruction, and c) emphasis on various levels of cognitive demand (e.g., memorizing, procedures) in the years 2003-2009. In addition, attributes of the standards-based accountability system are studied for their relationship with

instructional alignment. The methods used include fixed-effects models and three-level hierarchical linear modeling. The results suggest that alignment to standards and assessments in mathematics increased during the years 2003-2009 by approximately .3 to .5 standard deviations. Alignment also increased to science standards at grades 3-8, but there were no consistent changes in alignment in ELAR or at other grades. Shifts of instructional time were small and generally non-significant. Teachers were found to increase their focus on lower-level thinking (e.g., procedure, memorization) in most grades and all three subjects. Generally, these shifts brought teachers' instruction into greater agreement with standards and assessments. Finally, alignment was higher when standards and assessments were well aligned with each other (average standardized effect size = $.22\sigma$), when the standards were broad and focused on many ideas ($.50\sigma$), and when the state had stronger policies emphasizing rewards and sanctions ($.12\sigma$). Implications for policy and research are discussed.

TABLE OF CONTENTS

	Page
List of Tables	ix
List of Figures	xi
Chapter 1: Introduction	1
Chapter 2: Review of the Literature	15
Chapter 3: Data and Methods	37
Chapter 4: Results	96
Chapter 5: Summary and Discussion	147
Appendices	170
Bibliography	181

LIST OF TABLES

	Page
Table 1: Studies Investigating the Effects of Standards-Based Reform on the Content of Teachers' Instruction	20
Table 2: Summarized Teacher Sample Sizes	57
Table 3: Teacher Sample Sizes by Year	58
Table 4: Teacher Sample Sizes by Grade Band	59
Table 5: Teacher Sample Sizes by State	61
Table 6: OLS Regressions of Sample Descriptive Statistics on Year of Administration	63
Table 7: Comparison of Sample Descriptive Statistics with National Averages	65
Table 8: Cronbach's Alpha Scale Reliabilities	83
Table 9: Sample Descriptive Statistics	85
Table 10: Correlations among the Policy Attributes	94
Table 11: Fixed-Effects Regressions of Alignment to Mathematics Standards and Assessments	97
Table 12: Fixed-Effects Estimates of Changes in Alignment within Individual States, Mathematics	100
Table 13: Fixed-Effects Regressions of Alignment to ELAR standards and Assessments	102
Table 14: Fixed-Effects Estimates of Changes in Alignment within Individual States, ELAR	103
Table 15: Fixed-Effects Regressions of Alignment to Science Standards and Assessments	104
Table 16: Fixed-Effects Estimates of Changes in Alignment within Individual States, Science	105
Table 17: Comparison of Fixed Effects Regressions of Alignment to Standards and Assessments Using Main Data and Re-centered Data	107
Table 18: Fixed Effects Regressions of Hours of Instruction on Year of Administration	112
Table 19: Changes in Emphasis on Five Levels of Cognitive Demand in Mathematics Instruction	114
Table 20: Changes in Emphasis on Five Levels of Cognitive Demand in ELAR Instruction	115
Table 21: Changes in Emphasis on Five Levels of Cognitive Demand in Science Instruction	117
Table 22: Comparison of Change in Cognitive Demand for Teachers' Mathematics Instruction with Cognitive Demand Emphasis of Standards and Assessments, 2003-2009	119
Table 23: Comparison of Change in Cognitive Demand for Teachers' ELAR Instruction with Cognitive Demand Emphasis of Standards and Assessments, 2004-2009	120

Table 24: Comparison of Change in Cognitive Demand for Teachers' Science Instruction with Cognitive Demand Emphasis of Standards and Assessments, 2003-2009	121
Table 25: Changes in Emphasis on Coarse-Grained Topics in Mathematics Instruction	123
Table 26: Changes in Emphasis on Coarse-Grained Topics in ELAR Instruction	124
Table 27: Changes in Emphasis on Coarse-Grained Topics in Science Instruction	125
Table 28: Comparison of Change in Topic Coverage for Teachers' Mathematics Instruction with Topic Emphasis of Standards and Assessments, 2003-2009	127
Table 29: Comparison of Change in Topic Coverage for Teachers' ELAR Instruction with Topic Emphasis of Standards and Assessments, 2004-2009	128
Table 30: Comparison of Change in Topic Coverage for Teachers' Science Instruction with Topic Emphasis of Standards and Assessments, 2003-2009	129
Table 31: Summary of Fully Unconditional Models	132
Table 32: Full HLM of Alignment to Standards and Assessments in Mathematics	133
Table 33: Full HLM of Alignment to Standards and Assessments in ELAR	136
Table 34: Full HLM of Alignment to Standards and Assessments in Science	139
Table 35: Comparison of Models: 2003-2009 Compared With 2007-2009	141
Table 36: Comparison of Models: Main Data Compared with Re-centered Data	143
Table 37: Standardized Regression Coefficients for Predicting Alignment from Policy Attributes	157
Table C.1: Changes in Emphasis on Five Levels of Cognitive Demand Using Recentered Data	174
Table C.2: HLM of Alignment to Standards and Assessments in Mathematics Using 2007-2009 Data	175
Table C.3: HLM of Alignment to Standards and Assessments in ELAR Using 2007-2009 Data	176
Table C.4: HLM of Alignment to Standards and Assessments in Science Using 2007-2009 Data	177
Table C.5: HLM of Alignment to Standards and Assessments in Mathematics Using Re-Centered Data	178
Table C.6: HLM of Alignment to Standards and Assessments in ELAR Using Re-Centered Data	179
Table C.7: HLM of Alignment to Standards and Assessments in Science Using Re-Centered Data	180

LIST OF FIGURES

	Page
Figure 1: Theoretical framework for the study of standards-based reform	46
Figure 2: Illustrative content map	54
Figure 3: Histogram of FOCUS in science	86
Figure 4: Content maps for the most and least focused standards in mathematics	87
Figure 5: Content maps for the least consistent standards and assessments in science	89
Figure 6: Content maps for the most consistent standards and assessments in science	90
Figure 7: Histogram of power index for 50 U.S. states, 2003 & 2009	91

Chapter 1

Introduction

Throughout the last 25 years in U.S. K-12 education, standards-based reform (SBR) has been an important policy, guiding actions at the federal and state levels. Standards-based reform arose out of concerns over educational quality expressed in *A Nation at Risk* (National Commission on Excellence in Education, 1983), prominent theories about educational reform (M. S. Smith & O'Day, 1991), accompanying statewide initiatives (Clune, 2001), and the standardizing actions of teachers (National Council of Teachers of Mathematics, 1989) and state leaders (Guthrie et al., 1989). The theory of change for SBR starts with the building of coherent education policy, most notably through the construction of content standards in core academic subjects and the establishment of measurable goals (M. S. Smith & O'Day, 1991). These goals generally take the form of targets for student learning measured by aligned, standardized assessments. With appropriate supports and accountability measures, the theory proposes that teachers will modify their instruction to align with the standards and assessments, and student learning will improve (Clune, 1993, 2001; O'Day & Smith, 1993; M. S. Smith & O'Day, 1991). The modern incarnation of SBR is the No Child Left Behind Act ("No Child Left Behind Act of 2001," 2002), which has codified standards-based reform in federal law, using the carrot of federal dollars to effectively mandate the basic tenets of standards and aligned assessments with supports and sanctions to improve student learning.

Since the early years of SBR's implementation, researchers have conducted dozens of studies on the policy's effects on student achievement (e.g., Betts & Danenberg, 2002; Carnoy & Loeb, 2004; Center on Education Policy, 2006; Dee & Jacob, 2009; Grissmer & Flanagan, 2001; Hanushek & Raymond, 2005; Jacob, 2005) and teachers' instruction (e.g., Koretz, Barron, Mitchell, & Stecher, 1996; Luna & Turner, 2001; Pedulla et al., 2003; Shepard & Dougherty, 1991; M. L. Smith, 1991; Stecher & Barron, 2001). While these studies have generally found significant effects, they have often omitted or glossed over the first critical step in SBR's theory of change—the extent to which teachers are aligning the content of their instruction to standards and assessments. Even studies of instructional effects have focused primarily on pedagogical issues, such as test preparation, cheating, and the use of reform versus traditional methods of instruction. The neglect of content in studies of SBR's effects is surprising and troubling, both because content is central to the policy, and because the content of instruction is highly predictive of gains in student learning (Cueto, Ramirez, & Leon, 2006; Gamoran, Porter, Smithson, & White, 1997; Sebring, 1987).

Furthermore, the research on content effects that does exist is limited by both substantive and methodological problems. In terms of substance, while the studies are suggestive of increasing alignment and other changes in the content of instruction, they provide little evidence about the ways in which state SBR policies can encourage instructional alignment. Except for a few studies that suggest that teachers in states where the policies have more power (i.e., more types of or stronger rewards and sanctions) report greater efforts at instructional alignment (Clarke et al., 2003; Firestone,

Mayrowetz, & Fairman, 1998; Pedulla et al., 2003), there is little useful information to guide policymakers seeking to improve instructional alignment and, through alignment, student learning.

A second substantive problem with the extant research on content effects is that the dependent variables are so broad as to be uninformative about the nature of changes in teachers' instruction. One type of dependent variable is teachers' self-reported instructional alignment. While large majorities of teachers in survey studies report increasing instructional alignment with tests and/or standards, there is not good evidence about the magnitude or nature of these changes (Hamilton & Berends, 2006; Koretz, Barron et al., 1996; Pedulla et al., 2003; Taylor, Shepard, Kinner, & Rosenthal, 2002). What does it mean for teachers to report increased instructional alignment? Such questions also assume that teachers understand the content messages of standards and assessments, which may not be accurate (Hill, 2001; McDonnell, 2004). A second type of dependent variable is teachers' reported coverage of core content strands (e.g., algebra). Again, while teachers report increasing their coverage of content strands that are emphasized on the standards and assessments (Koretz, Mitchell, Barron, & Keith, 1996; Stecher & Barron, 2001; Stecher & Chun, 2001), it is unclear that an increase in coverage of algebra, for instance, necessarily indicates an instructional change that is in the spirit of the standards or assessments. There are many components of algebra instruction, and the teachers may or may not be increasing focus on those topics that are emphasized in the standards and assessments.

There are also potential methodological problems with the extant research on content effects of SBR. For one, the studies generally rely on teachers' self-report about the extent to which they are changing or have changed their instruction. Some evidence suggests that it is difficult for teachers to validly report on their own instructional changes (Desimone, Le Floch, Ansell, & Taylor, 2005). Other research indicates that asking survey questions about perceived behavioral change results in larger estimates of change than other methods such as asking about both pretest and posttest behaviors at posttest (Lam & Bengo, 2003). Furthermore, there is some evidence that individuals are likely to report changes in behavior when such changes are expected, as is the case with the introduction of standards-based reforms, even when such changes have not taken place (Ross, 1989; Ross & Conway, 1986). Indeed, Schwartz and Oyserman (2001) argued that the reliability and validity of survey reports about behavioral change are highly troublesome:

Asking program participants to report on how their behavior has changed over the course of the intervention, or what their behavior was prior to the intervention, is likely to result in theory-driven reconstructions. These reconstructions are useless as measures of objective change, although they may be of interest as measures of participants' subjective perceptions. To assess actual change, we need to rely on before-after, or treatment-control, comparisons, and if we have missed asking the right question before the intervention, little can be done after the fact to make up for the oversight. (p.144)

While their quote was in the context of program evaluation, it may well be applicable to wide-scale policy evaluation, particularly for a policy as salient to teachers as standards-based reform. And given that the theory for standards-based reform clearly predicts instructional changes in the form of increased alignment, the validity of teacher-reported change in alignment appears questionable, at best.

Another methodological concern is that, especially in those studies where teachers are asked whether and to what extent they have aligned their instruction with standards and/or assessments, one-item survey questions about instruction are quite common. The limitations of one-item scales in terms of reliability are obvious, and research questions the use of such indicators in reporting on instructional effects (Mayer, 1999).

A more useful approach than that taken in prior analyses is to use data on the self-reported content of teachers' instruction. The data are collected at a particular point in time and they are at a greater level of detail. Thus, the data can be used to investigate change over time. These data appear to be a reliable and valid way to describe the content of teachers' instruction (Porter, Kirst, Osthoff, Smithson, & Schneider, 1993). Also, the data have been used to investigate instructional alignment, which has been shown to be highly predictive of gains in student learning (Gamoran et al., 1997).

These data have several important advantages over the data used in previous investigations of SBR's effects on the content of instruction. For one, the data provide much greater detail about the nature of instruction, allowing investigation at a detailed level of the ways in which teachers have responded to standards and assessments. For another, the data can be directly compared with content analyses of standards and assessments to estimate instructional alignment in a way that requires no assumptions about teachers' understanding of the content messages of those documents. Additionally, the data can be connected to both teacher-level survey data and state-level policy data to investigate the ways in which local contexts and state policies relate to teachers' instructional alignment. No prior investigation has used large-scale, multi-state data to

compare the content of teachers' instruction with the content of state standards and assessment documents at the fine-grained level of detail that is most useful for predicting student learning (Gamoran et al., 1997). Thus, this dissertation contributes to an understanding of the effectiveness of standards-based reform in general and state standards-based policy attributes in particular in influencing teachers' instructional alignment.

Significance

The Content of Instruction is a Predictor of Student Learning. Standards-based educational policy is based on the idea that the content of teachers' instruction is a powerful predictor of what students actually learn. Research bears out the importance of content as a predictor of student learning in diverse settings (Cueto et al., 2006; Gamoran et al., 1997; Sebring, 1987). Indeed, after accounting for content coverage, between-class differences in student learning gains may be non-significant (Gamoran et al., 1997). Even at the cross-national level, research suggests that differences in curriculum content explain much of the large achievement differences found among nations (Schmidt et al., 2001). Across a multitude of studies, the content of instruction is predictive of student learning, whether it is measured precisely (e.g., Gamoran et al., 1997; Smithson & Collares, 2007) or using rough measures such as coursetaking (Carbonaro, 2005; Gamoran & Nystrand, 1994; Hoffer, 1994).

Furthermore, the content of teachers' instruction in U.S. schools has not historically been externally controlled (Kliebard, 2004; Lortie, 1975). Rather, content decisions have often been left to individual teachers in their classrooms. Thus, the

architects of standards-based reform viewed the content of instruction as an important policy lever for improving student learning through a focus on “depth of understanding, knowledge construction through analysis and synthesis of real life problems, hands-on experiences, and the integration of content and pedagogy” (M. S. Smith & O’Day, 1991, p. 248). Aligned instruction was the first goal of standards-based reform.

Standards-Based Reform Is Affecting Teachers and Students. Decades of research on the implementation of educational reforms predating standards-based reform found limited or superficial change in teaching practices across a wide variety of educational reforms (Chubb & Moe, 1990; D. K. Cohen, 1990; Elmore, 2000; Lortie, 1975; Meyer & Rowan, 1980). The lack of wide-scale change was attributed to such features of schools as the variability of school inputs (e.g., children, teacher quality) (Bidwell, 1965), the excessiveness of school bureaucracy (Chubb & Moe, 1990), the structural nature of teaching that emphasizes individualism (Elmore, 2000; Lortie, 1975), or the desire of school leaders to buffer teachers from external pressures (Meyer & Rowan, 1980). In the years before standards-based reform, it had become largely accepted that schools were “loosely-coupled” organizations where the elements (e.g., teachers, leaders, policymakers, students) were connected, but in a way that preserved each element’s identity and authority (Weick, 1976).

Standards-based reform is a policy designed to “tighten” the coupling of educational policy and practice. Research to date suggests the policy has been at least partly effective in that regard. While some research on the implementation of SBR indicates that its effects on teachers have been smaller than expected (Hess, 2002; Hill,

2001; McDonnell & Choisser, 1997; M. L. Smith, 1997; Supovitz, 2009), surveys generally reveal that teachers almost unanimously report effects of tests and standards on instruction (Hamilton & Berends, 2006; Koretz, Barron et al., 1996; Koretz, Mitchell et al., 1996; Pedulla et al., 2003). Some of these effects are described in more detail in the review of the literature.

Effects of SBR on student learning have been mainly positive (Betts & Danenberg, 2002; Carnoy & Loeb, 2004; Center on Education Policy, 2006; Grissmer & Flanagan, 2001; Hanushek & Raymond, 2005; Jacob, 2005; Sullivan, 2006). The most methodologically sophisticated research (Figlio & Ladd, 2008) suggests positive effects of pre-No Child Left Behind (NCLB) accountability systems of .2 to .35 standard deviations on student achievement as measured by the National Assessment of Educational Progress (NAEP) and the Iowa Test of Basic Skills (Hanushek & Raymond, 2005; Jacob, 2005), though little narrowing of achievement gaps (Hanushek & Raymond, 2005). A recent study (Dee & Jacob, 2009) comparing states that implemented NCLB-like accountability early to those implementing NCLB accountability after 2002 found similar effects of .15 to .26 standard deviations in 4th and 8th grade mathematics, but no significant effects in English language arts/reading (ELAR). In these studies, the effects of standards-based reform were of roughly equal magnitudes across racial/ethnic and other student groups. These effect sizes are equal to or slightly larger than effect sizes found for typical comprehensive school reforms in early years of implementation (.13 to .25 standard deviations, Borman, Hewes, Overman, & Brown, 2003).

While researchers and policymakers have been notably focused on the effects of standards-based policies on academic achievement, less work has been done on instruction in general and the content of instruction in particular. When research has focused on instruction, it has often been centered on pedagogical effects, such as test preparation. Less research has focused on student opportunity to learn the core content, defined in the curriculum frameworks, which is the centerpiece of the original vision of standards-based reform (M. S. Smith & O'Day, 1991). Thus, while improvements in student learning and teacher responses on surveys are suggestive of changes in the content and quality of teachers' instruction, research cannot yet specify the nature and magnitude of those changes.

States Have Implemented Standards-Based Reform Policies Differently, and That Variability Could Explain Differences in Teachers' Aligned Instruction. Recent research highlights that, while NCLB makes standards-based reform a part of national law, the leeway NCLB grants to states has resulted in widely varying content expectations (Porter & Polikoff, 2007; Porter, Polikoff, & Smithson, 2009; Reys, 2006; Reys et al., 2007). The variability in implementation of NCLB across states is manifested not only in the content of state standards (Porter et al., 2009; Reys, 2006; Reys et al., 2007), but also the rigor of state cut scores on teacher licensure examinations (Porter & Polikoff, 2007), and the annual targets chosen for adequate yearly progress (Porter, Linn, & Trimble, 2005). For both standards and assessments, the key documents designed to influence teachers' instructional choices, there are large differences in the nature of the content in, the degree of focus of, and the alignment between state standards and

assessments, suggesting a troubling lack of coherence in current standards-based policy systems (Polikoff, Porter, & Smithson, 2009; Porter et al., 2009). There have also been reviews of state content standards that have argued that they differ on various dimensions of quality (American Federation of Teachers, 2001; Finn, Petrilli, & Julian, 2006).

Beyond the content and alignment of standards and assessments, states have leeway over such standards-based reform policies as the types of rewards and sanctions for schools and students and the extent to which standards-based reform policies remain stable over time. Taken together, the attributes of state policies around standards-based reform are highly variable (Porter & Polikoff, 2007). As state policy plays an increasingly prominent role in affecting the classroom, the features of state policies—specificity, stability, consistency, power, and authority—have been shown to affect the implementation of standards-based policies and, ultimately, student learning (Clune, 2001; Desimone, 2002; Desimone, Smith, Hayes, & Frisvold, 2004). While a few qualitative studies (Clarke et al., 2003; Firestone et al., 1998) and one national survey (Pedulla et al., 2003) suggest that effects of SBR on the content of instruction are larger in states high on power (i.e., sanctions and/or rewards for schools and/or students), the magnitude and nature of such changes is unclear, and no other state policy attributes have been directly investigated for effects on instructional content. States wishing to leverage instructional change through standards and assessments need information on the ways policies affect teachers' decisions about what to teach. Thus, the focus here is on the content of instruction and how states can improve the impact of aligned, standards-based accountability systems on teachers.

The significance of this study is based on three points. First, this is the first study to use detailed data on the content of teachers' instruction, described at the intersection of topics and cognitive demands, to investigate the effects of standards-based reform on the content of teachers' instruction. This is in contrast to previous studies, which have primarily relied on problematic one-item indicators or teacher perceptions of instructional change. Second, this is the first study to directly compare the content of teachers' instruction with the content of state standards and assessments in describing instructional alignment. Third, this study is the first to connect multiple attributes of state policies to teachers' instructional alignment.

Research Questions

From the literature briefly described above and in more detail below, the large majority of teachers report changing the content of their instruction to align with the content messages in standards and assessments (Firestone et al., 1998; Hamilton & Berends, 2006; Koretz, Barron et al., 1996; Koretz, Mitchell et al., 1996; Lane, Stone, Parke, Hansen, & Cerrillo, 2000; Stecher, Barron, Chun, & Ross, 2000; Taylor et al., 2002). However, the quality of the evidence on the nature of teachers' changes in instructional content is weak. There is also little evidence about the ways that states can design standards-based policy systems to impact the content of teachers' instruction and encourage greater alignment to rigorous curriculum frameworks. There is a need for evaluation of the extent to which teachers are responding to the content messages of the standards and assessments.

To accomplish this evaluation, this dissertation uses data on teachers' self-reported content of instruction collected since 2003. These instructional data are compared with the content of standards and assessments as coded by trained content analysts using the Surveys of Enacted Curriculum, a nationally-recognized content analysis procedure (Porter, 2002). In addition, key state standards-based reform policies are measured in a state policy database. Together, the data are used to address the following set of questions:

- 1) To what extent have teachers aligned their instruction with standards and assessments in English language arts and reading, mathematics, and science, since 2003?
- 2) In what ways have teachers changed the distribution of instructional time and the topics and cognitive demand levels of their instruction in the years since 2003, and to what extent are the changes in topics and cognitive demand levels reflective of the content messages of standards and assessments?
- 3) To what extent are state policy contexts, features of standards and assessments, and features of local context related to alignment?

Questions 1 and 3 are focused on instructional alignment and the relationship of state policies with alignment. Question 2 provides useful descriptive information about the nature of changes in teachers' instruction that affect alignment. Results from the research provide evidence about the ways in which state policies, standards, and assessments can be better designed to improve the impact of SBR on the content of instruction.

Definition of Terms

Content—the material that is taught, defined at the intersection of topics and cognitive demands (Porter, 2002).

Pedagogy—the methods teachers use to convey content to students.

Instruction—the intersection of content and pedagogy: “interactions among teachers and students, around content, in environments” (D. K. Cohen, Raudenbush, & Ball, 2003, p. 122).

Curriculum—the intended curriculum, as represented through formal documents (Schmidt et al., 2001).

Standards—the lists or descriptions of the content that students are to know or be able to do (i.e., content standards).

Assessments—measures of student learning administered by states.

Delimitations of the Study

The study is limited by the sample represented (described in more detail in chapter three). The sample of states and teachers may not be nationally representative, and the results of this research should not necessarily be construed to apply to all teachers nationwide. Furthermore, since the teachers in this study were only K-12 teachers at U.S. public schools, the results may not be representative of teachers at other kinds of schools.

Organization of the Dissertation

Chapter 1 provides a brief description of the background for the study, the significance of the study, and the research questions. Chapter 2 provides a review of the literature on the effects of standards-based reforms on the content of teachers' instruction.

Chapter 3 describes the data and methods used to address the research questions. Chapter 4 presents the main results for the three research questions. Chapter 5 discusses the results and the implications for research and policy.

Chapter 2

Review of the Literature

The purpose of this chapter is to situate this study in the literature on the effects of standards-based reform on the content of teachers' instruction. The focus of this chapter is narrowly constrained to focus on effects on content, rather than effects on pedagogy or student achievement, because content is the focus of this dissertation. Studies were identified for inclusion in this literature review through searches of ERIC and Google Scholar, as well as manual searches of reference lists for key books and articles. Because few studies addressing content effects were published in peer-reviewed journals, reports and dissertations were included where applicable. The first part of the chapter contains a brief summary of the intended effects of standards-based reforms on teachers' content decisions. The next part identifies and summarizes the empirical research on the effects of SBR on the content of teachers' instruction.

Intended Effects of Standards-Based Reform

The theory of change of standards-based, systemic accountability has been well articulated since at least the early 1990s. The need for a 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 U.S. had created a fragmented system (O'Day & Smith, 1993; M. S. Smith & O'Day, 1991), with incoherent curricula (D. K. Cohen, 1989; Goodlad, 1984) and poor quality teacher preparation and professional development (M. S. 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 (D. K. Cohen & Spillane, 1992).

This fragmentation helped ensure that educational reforms rarely, if ever, achieved their target of change (Cuban, 1984; Tyack & Cuban, 1995). And when change did come, it was not often in the format envisioned by policymakers. Change was often additive rather than substitutive, and the 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 often 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 from on high, without consideration for the varied needs of schools (Tyack & Cuban, 1995). In short, by the late 1980s and early 1990s, 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 the teaching and learning going on in the classroom (M. S. Smith & O'Day, 1991). The focus was on the state as the primary level of policymaking. Smith and O'Day outlined three facets of systemic reform: a unifying vision, a coherent system of instructional guidance, and a restructured school governance system. In describing a unifying vision, they argued that (a) all students should have access to a high quality school, (b) schools should embrace a commitment to democratic values, and (c)

teachers and learners should embrace exploration and intellectual rigor. Additionally, they argued that goals should be made explicit and measurable, so that progress toward them could be tracked. Goals could include inputs, but they should be focused on teaching and learning and outcomes. And finally, goals should focus not just on average outcomes, but the full distribution (p. 246-247).

The most important step in reaching educational goals was establishing instructional coherence. Smith and O'Day recognized that teaching and learning were at the core of the school's work. They advocated the creation of deep, rigorous curriculum frameworks in the core content areas. The frameworks would not be prescriptive of teaching methods or even curriculum; rather, they would provide guidance by establishing a set of core content expected of all students. Along with frameworks, they recommended the adoption of curricular materials that were high quality and aligned to the frameworks. To support teachers, they advocated coherent pre-service and in-service professional development in content and effective 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 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, at least in the original vision, the decisions on whether to use test results for accountability purposes would be left to states.

The final component of the systemic approach to educational reform was restructured school governance. Smith and O'Day proposed that states should provide

increased resources to schools along with any accountability provisions. States also should not let partisan politics and varying agendas over time interfere with the progress of reform. Then, schools would be responsible for ensuring progress toward the goals, with the state only intervening if sufficient progress was not made. Three critical features of the schools' responsibility would be the selection and development of personnel, the sharing of authority and responsibility within the school, and the capital resources (e.g., rooms, technology) to reach the goals. Districts would be responsible for providing resources to schools and ensuring that no students fall through the cracks.

With the three components in place, the theory suggested that several important changes would happen to U.S. education. For one, the quality of instruction would improve, based on the introduction of the content frameworks and the emphasis on teacher control over the implementation of the standards. Teacher collaboration would also improve, as some of the responsibility formerly vested in the school and district leadership would be devolved to teachers. Additionally, teacher in-service and pre-service training would improve with the new focus on coherence and capacity-building. The assessment and accountability system would encourage schools and teachers to focus on all students. In short, the causal model was plain enough, even if the steps advocated required a refocusing of school, district, and state priorities.

To some extent, none of the ideas proposed by Smith and O'Day were new. Many states had been employing various components of the systemic approach for years. However, in synthesizing top-down and bottom-up approaches to education reform, they sought a new path. Their 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).

The Effects of Standards and Assessments on the Content of Instruction

Throughout the 1990s and 2000s, researchers have investigated the effects of standards and assessments on the content of teachers' instruction. These studies have included qualitative (e.g., Borko & Elliott, 1998; M. L. Smith, Edelsky, Draper, Rottenberg, & Cherland, 1990), quantitative (e.g., Pedulla et al., 2003; Stecher et al., 2000), and mixed methods studies (e.g., Hamilton & Berends, 2006; M. L. Smith, 1997), though most of the studies have been based on surveys of teachers. Primarily, they have been studies based in individual states—teachers in at least 22 individual states have been investigated—though a few national studies have also been conducted (e.g., Pedulla et al., 2003; Ysseldyke et al., 2004). The studies have targeted all levels of schooling and the four core content areas, though the primary subjects have been mathematics and ELAR. Table 1 provides a list of the studies identified that investigate effects on the content of teachers' instruction. Together, the studies provide important information about the effects of standards-based reform on teachers' content decisions, which is summarized here. However, the studies leave important questions unanswered, or answered unsatisfactorily. Furthermore, all but five of the studies pre-date the period studied here, and the large majority pre-date NCLB. Thus, the literature on recent effects is quite limited, and the earlier literature points in the direction for new research.

Table 1
Studies Investigating the Effects of Standards-Based Reform on the Content of Teachers' Instruction

Citation	Location	Method	N	Subject/grade
Borko & Elliott, 1998	KY	case study	2 teachers	Math
Borko et al., 2002	KY	case study	4 schools	All
Burroughs, Groce, & Webeck, 2005	NC, TX, & MS	interviews, surveys	34 teachers	Elementary
Charles, 2008	LA	interviews, surveys	8 schools, 112 teachers	Elementary
Clarke et al., 2003	KS, MI, MA	interviews	360 teachers	All
Falk & Drayton, 2004	MA	interviews, observations	6 schools, 40 teachers	Science
Firestone et al., 1998	ME & MD	interviews, observations	5 districts, 82 teachers	8th grade math
Grant, 2001	NY	case studies	2 teachers	Secondary social studies
Hagge & Waltman, 2007	IA	surveys	131 schools, 1338 teachers	All
Hamilton & Berends, 2006	CA, PA, GA	case study, surveys	2840 teachers	Elementary & middle math & science
Herman & Klein, 1996	CA	interviews, observations, surveys	13 schools, 27 teachers	Math
Jones et al., 1999	NC	surveys	16 schools, 236 teachers	Elementary
Koretz et al., 1996	KY	surveys	178 teachers	4th grade, 8th grade math
Koretz et al., 1996b	MD	surveys	224 teachers	5th grade, 8th grade math
Lomax et al., 1996	National	document analysis, surveys	2229 teachers	Math and science, 4-12
Luna & Turner, 2001	MA	interviews	2 schools, 10 teachers	9th & 10th English
McDonnell, 2004	KY	interviews, daily logs	12 schools, 46 teachers	All levels, multiple subjects
McMillan et al., 1999	VA	surveys	19 schools, 732 teachers	All
Pedulla et al., 2003	National	surveys	4195 teachers	All
Schorr et al., 2003	NJ	interviews, observations	63 teachers	4th math
Schulz, 2008	GA	interview	5 schools, 6 teachers	2nd - 4th grade
Shepard & Dougherty, 1991	not listed	surveys	100 schools, 360 teachers	All, grades 3, 5, 6
Smith & Rottenberg, 1991	AZ	case studies	2 schools, 19 teachers	All
Smith et al., 1990	AZ	interviews, observations	2 schools, 29 teachers	not listed
Smith, 1997	AZ	case studies, interviews, surveys	4 schools	Elementary
Spillane, 2004	MI	case studies, interviews, surveys	9 districts, 283 teachers	Math & science
Stecher & Barron, 2001	KY	surveys	152 schools, 479 teachers	Writing and math; 4-7
Stecher & Chun, 2001	WA	surveys	70 schools, 400 teachers	4th and 7th
Stecher et al., 1999	KY	case study, surveys	391 teachers	4th & 7th writing, 5th & 8th math
Stecher et al., 2000	WA	surveys	70 schools, 400 teachers	4th and 7th
Sullivan, 2006	VA	surveys	337 teachers	Computer technology
Taylor et al., 2002	CO	surveys	510 teachers	All
Wong et al., 2003	Chicago	document analysis, observations, surveys	4 schools, 71 teachers	10th and 11th English and math

In general, most of the studies of the effects of standards and assessments on instruction have focused more on pedagogical effects than on effects on content. Where content is considered, it is often in survey questions about whether the teachers have

aligned their instruction, or whether and how much teachers have covered certain broad topics. No studies have focused on the fine-grained content decisions that are most important for predicting student learning gains (Gamoran et al., 1997), and no studies were identified that compared the content of instruction with the content of standards and assessments at a fine-grained level of detail (i.e., specific topics and levels of student expectation). The privileging of pedagogy over content in research on instructional effects is an important oversight, as the content of instruction and its alignment with the content of assessment is an important predictor of gains in achievement (Gamoran et al., 1997; Smithson & Collares, 2007). The purpose of this section is to review the evidence about the effects of standards-based policies on the content of teachers' instruction. Two primary effects on the content of instruction emerge in this review: a curricular narrowing, where instructional time is shifted to tested subjects and, to a lesser extent, away from non-tested subjects; and an increased alignment, where teachers report increasing the consistency between the content of their instruction and the content emphasized in standards and assessments.

Curricular narrowing. The first effect of standards-based reform on the content of instruction is a shifting of instructional time to tested subjects and away from non-tested subjects. This finding is confirmed in qualitative (Clarke et al., 2003; Schulz, 2008; M. L. Smith et al., 1990; M. L. Smith & Rottenberg, 1991) and quantitative studies (e.g., Hamilton & Berends, 2006; Pedulla et al., 2003; Stecher et al., 2000), and in surveys of national samples (Pedulla et al., 2003) and individual states.

The largest national survey of 4,195 teachers at all grades and across the core content areas and special education found that teachers in schools of all types (elementary, middle, and high; low, medium, and high stakes for students and teachers) reported increasing the time spent teaching in tested areas and decreasing or not changing the time teaching in non-core areas. Teachers were asked how much the amount of time they spent on certain activities had changed in their school in order to prepare students for the state-mandated test. Because the survey was national, “tested areas” was defined as whatever content areas were tested or had stakes attached to them in each state. Non-core areas included foreign language, physical education, fine arts, and industrial or vocational education. Between 68% and 79% of teachers reported increasing time spent on instruction in tested areas, and between 42% and 63% of teachers reported decreasing the time spent on instruction in non-tested areas, depending on state accountability system type. Teachers reported increasing their instruction in tested areas more than they decreased their instruction in non-core areas. For non-core areas, the net proportion of teachers indicating a decrease in time (the percent indicating a decrease minus the percent indicating an increase) was between -11% and 29% for fine arts, 0% and 22% for physical education, 0% and 13% for foreign language, and 5% and 26% for industrial/vocational education. The results of this national survey suggest that teachers may be redistributing instructional time across subjects, though it is unclear where the newly-added time is coming from, since fewer teachers reported decreases in coverage in non-tested subjects than increases in tested ones.

Surveys of teachers in individual states support the conclusion that teachers report shifting instruction to tested areas and away from non-tested areas. In an early survey of teachers in Kentucky (Koretz, Barron et al., 1996), 93% reported that their schools' emphasis on material likely to be tested on KIRIS (the state assessment) had increased, and 88% of fourth grade teachers reported that KIRIS had caused some teachers to deemphasize or neglect untested subject areas. Even within content areas, teachers reported decreasing emphasis on untested topics, including 86% of eighth grade mathematics teachers. Five years later (Stecher & Barron, 2001), the effects of KIRIS remained—76% of fourth grade teachers reported increasing the time spent teaching science, a tested subject in that grade, as compared to 13% of fifth grade teachers, where science was not tested. A majority of teachers similarly indicated an increase in time spent teaching tested content areas in North Carolina (Jones et al., 1999), Washington (Stecher et al., 2000), and Colorado (Taylor et al., 2002). The survey studies that pre-date NCLB uniformly report majorities of teachers indicating content redistribution.

More recent studies report smaller year-on-year changes in instructional focus. A survey study of teachers in California, Georgia, and Pennsylvania found that just 22% to 38% of teachers in those states reported an increase in emphasis on mathematics from 2003-2004 to 2004-2005, with fewer reporting increases in emphasis on ELAR or decreases in emphasis on social studies or the arts (Hamilton & Berends, 2006). In Iowa, the proportion of teachers reporting changes in emphasis between 2004/5 and 2006/7 was also at 50% or lower (Hagge & Waltman, 2007). Overall, these survey studies support the conclusion that, across a wide range of settings, during the early years of standards-based

reform, large majorities of teachers perceived a shift in the content of their instruction from non-tested to tested subjects—a narrowing of the curriculum. Two studies suggest that the degree of this shift is diminishing as standards-based reform remains in place. Several studies also indicate that more teachers report increases in coverage of tested areas than report decreases in coverage of non-tested areas.

It is possible that the teachers surveyed perceive changes in the content of their instruction that are not really occurring. Qualitative and observational studies tend to support the conclusions of survey studies, however. One study including day-long observations of 29 classrooms in six Arizona elementary schools and 81 additional observations of two teachers (M. L. Smith et al., 1990) indicated that teachers were turning science activities into reading activities by decreasing hands-on work and having students read the textbook instead. Furthermore, as the test date approached, time normally spent on science was diverted to prepare for the reading test. While the date and small number of teachers observed are limitations of this study, the findings of curricular narrowing were also supported by teachers in interviews. In another interview study of Louisiana teachers, the teachers reported that non-tested subjects were neglected and English and mathematics instruction were emphasized instead (Charles, 2008). In Michigan, when science was added to the elementary school testing program, teachers reported covering science more frequently (Spillane, 2004). Overall, taken together with findings from national and state-specific surveys, there is evidence that a majority of teachers have redistributed some instructional time from non-tested subjects to tested ones.

The absolute magnitude of these redistributions is impossible to ascertain, due to the ways in which the data are reported. As mentioned previously, almost all teachers in early studies reported redistributing instructional time (e.g., Koretz, Barron et al., 1996), with later studies finding fewer teachers reporting changes (e.g., Hamilton & Berends, 2006). However, no available studies report the changes in terms of hours of instruction or other absolute measures of magnitude. Thus, while it seems clear that teachers believe they have changed the number of hours they teach tested and non-tested subjects since the implementation of standards-based reform, it is unclear whether these changes comprise minutes, hours, or days worth of instruction. The analysis presented here allows for the examination of trends in content redistribution from 2003 through 2009, which reveals the extent to which teachers have continued to shift the content of instruction across subjects as standards-based reform has matured. Furthermore, the data allow for the calculation of absolute magnitudes of redistribution in terms of hours of instruction per week, something which has not been done before.

Instructional alignment. The theory of change for standards-based reform indicates that the major effect on teachers' instruction will be to lead them to align their instruction with standards and assessments. A number of qualitative and quantitative studies, spread across a large variety of settings, have investigated the extent to which the expected instructional alignment is taking place. Again, the majority of these studies are survey studies of teachers in individual states, and these generally find that a high proportion of teachers report aligning their instruction with standards and/or assessments (e.g., 87% to 90% of KY teachers reporting alignment to the state test, Koretz, Barron et

al., 1996). Unlike with curricular narrowing, however, there are also some contrary findings, generally coming from qualitative studies involving interviews, observations, and/or logs (McDonnell, 2004; M. L. Smith, 1997). The contrary findings, taken together with the quality of the data from the survey studies, bring into question the extent to which alignment has actually increased.

Before teachers can align their instruction to standards and assessments, they must be willing to change their instruction by adding and removing topics. An early experimental policy-capturing study of 66 teachers asked them whether they would add or remove five topics from their instruction, given certain external pressures (Floden, Porter, Schmidt, Freeman, & Schwille, 1980). The six pressures were textbooks, district content standards, published test results, other teachers, principals, and parents. On a 1-to-7 scale, with 1 indicating virtual certainty of teaching the topic and 7 indicating the opposite, teachers indicated a great willingness to add new topics to their instruction (mean = 1.0 to 2.5), with tests and objectives being the most influential. In contrast, even when all six pressures were pushing against teachers, they were ambivalent about removing topics (mean = 3.8). These results support the claim that teachers are willing to change the content of their instruction, and that their content decisions are influenced by standards and assessments. The results also reinforce the finding from surveys that more teachers report adding time and topics to their instruction than removing them (Pedulla et al., 2003). However, the findings are merely suggestive of actual effects because only teachers' intentions were studied.

Many studies indicate that teachers report increasing the alignment of their instruction with standards and/or assessments. One source of evidence comes from survey questions where teachers are directly asked if they or their district are aligning their instruction. The proportion of teachers reporting efforts in alignment is universally large when such questions are asked. In Kentucky, a survey of 4th and 8th grade teachers found that 90% of 4th grade and 87% of 8th grade teachers focused “a moderate amount” or more on improving the content match between their instruction and the state test (Koretz, Barron et al., 1996). Similarly large proportions were found in California and Georgia in mathematics and science, where as many as 96% of teachers, depending on state and grade, reported aligning their instruction with state standards (Hamilton & Berends, 2006). In non-tested grades and subjects, the proportion indicating alignment was lower—45% to 57% in science in Pennsylvania. In Colorado, 86% of teachers reported adding something to their curriculum to improve alignment with the state test, and 82% indicated removing content to improve alignment (Taylor et al., 2002). A national survey of over 4000 teachers found that 76% or more of teachers, depending on state accountability type, reported alignment between their district curriculum and the state standards (Pedulla et al., 2003). Across these multiple survey studies, large proportions of teachers in tested subjects and grades report efforts to improve instructional alignment with standards and assessments.

A second way alignment with standards and assessments has been gauged is through teacher report of the degree to which they have changed their instruction to address focal content from the standards or assessments. An example of this approach is

the work of Stecher and Barron in Kentucky (2001). They asked teachers how often they covered core content from Kentucky's mathematics standards, finding that 50% to 56% of teachers in the accountability grade reported covering the content areas once a week or more. In three of the four core content areas, this proportion was significantly higher than the proportion of teachers in non-accountability grades covering those core content areas once a week or more. In Maryland, 80-81% of fifth grade teachers reported increasing instruction on mathematical communication and problem solving, two strands emphasized by the state's performance assessment (Koretz, Mitchell et al., 1996). In contrast, by a 26% margin, teachers reported decreasing instruction on number facts and computation, strands less emphasized on the state test. In Washington, teachers were surveyed about their amount of focus on each writing strand in the standards. Across all but one strand, 29% or more of teachers reported increasing their coverage in the past year (Stecher & Chun, 2001).

While these three studies may be suggestive that teachers have actually improved their instructional alignment to standards at the individual strand level, there are a number of important caveats about this data. For one, the studies do not report how well represented each strand is in the standards and assessments. Thus, the strands teachers are emphasizing most could be the strands least emphasized in the standards and assessments. Furthermore, these indicators do not get to the fine-grained topics identified in the standards and assessments, so it is impossible to evaluate alignment at a detailed level. Nevertheless, these findings largely support teachers' claims of alignment, insofar

as they indicate that the content in the standards is, broadly speaking, increasingly being covered.

The consistency of the findings from surveys studies is not always reflected in qualitative studies based on interviews or observations. In a mixed methods study of four elementary schools in Arizona, more than half of the teachers indicated little or no effect of the assessment on their instruction. While schools were responding to the reform's formal regulations, detailed case-study observations revealed that teachers were misinterpreting the underlying intent, such as by creating separate test-preparation classes rather than aligning their regular instruction with the test or standards (M. L. Smith, 1997). McDonnell's (2004) interview studies of teachers in North Carolina and Kentucky indicated that teachers did not understand that changing the content of instruction was a primary purpose of the state assessment system until after several years of implementation. Analysis of teaching logs and classroom assessments also indicated that major curriculum reform goals emphasized in the standards and assessments, such as the linking of content across subjects, were largely not taking place. When teachers turned in assignments they thought were aligned with the standards and assessments, content analyses revealed that fewer than half of the assignments were well aligned. Overall, these findings largely support previous work on teachers' implementation of educational reforms, which suggests that uneven implementation has multiple causes, including the sometimes conflicting role of district policies (Wong, Anagnostopoulos, Rutledge, & Edwards, 2003), the misinterpretation of content specifications (Hill, 2001), and the tendency to graft new approaches onto existing ones (D. K. Cohen, 1990).

Across the extant research, there is some evidence that teachers have modified their instruction to align with state standards and assessments. For one, teachers overwhelmingly report having done so when directly asked. For another, they report increasing their focus on specific content strands emphasized in the standards and assessments and, sometimes, decreasing their focus on content not tested. However, there is evidence of uneven implementation, or that, when actual observations of classes or analyses of teacher lessons are done, the alignment is not as strong as was indicated by survey responses. Furthermore, there is evidence that the interpretation of the content messages of standards and assessment documents is a challenging barrier to implementation. Another barrier to implementing aligned instruction may be presented by states or districts when they report to teachers that certain curricula, textbooks, or assessments are aligned with standards even when the alignment has not been rigorously studied. Since teachers respond most to more proximal documents than standards, such as programs of study and pacing guides (Wong et al., 2003), messages from districts may not lead to improved alignment with standards and assessments if these documents are not as well aligned as is claimed. That is, teachers might believe they are practicing aligned instruction because they are using a certain textbook, but in reality the textbook might not be well aligned with standards. In short, while there are numerous reports of increased instructional alignment resulting from standards-based reform, there are several counterexamples or reasons why teachers might be overstating the extent of their instructional alignment.

Additionally, there are certain measurement issues in the extant literature on changes in alignment that limit the utility of prior research. For instance, as was the case with shifts in content coverage from one subject to another, the literature does not support estimation of the magnitude of changes in alignment. None of the studies described here define alignment in a measurable way (i.e., with units). Rather, the only measure of alignment is the proportion of teachers indicating their instruction is aligned. Other measurement issues include uncertainty about the extent to which individuals can validly report on behavioral change (Desimone et al., 2005) or attribute those changes to a particular policy (Ross, 1989; Ross & Conway, 1986), especially on single-item indicators such as the questions about instructional alignment used in the previously-mentioned survey items (Mayer, 1999). The research presented here uses data that require no assumptions about teachers' ability to estimate changes in instruction and allow for the estimation of the magnitude of alignment and changes in alignment over time.

State Policies and Content Effects. An important feature of the current incarnation of standards-based reform is that, while NCLB is a national policy, states have considerable freedom to implement the policy. As a result, states have created sets of standards-based reform policies that differ substantially from one another. For instance, as of 2009, 24 states have mandatory high school exits exams, making state tests high stakes for students, while 26 states do not (Center on Education Policy, 2009). States vary considerably in the level of rigor associated with their performance standards, such that schools are held to much higher standards in some states than others (de Mello, Blankenship, & McLaughlin, 2009). Furthermore, states differ substantially in the content

of their content standards in the core academic subjects, on average having just 25% of content in common across any two state standards documents in a particular grade/subject (Porter et al., 2009). State standards documents differ not only in their content, but also in their form—certain state standards documents are considerably less focused than other documents, covering as much as eight times as many topics in English, four times as many topics in science, and three times as many topics in mathematics (Porter et al., 2009).

It seems likely that these large differences in state policies and the content and features of state standards and assessments should impact teachers' decisions about what to teach, and that evidence about such effects could inform an understanding of how to better design state policies to influence teachers' decisions. State policy attributes have been shown to affect teachers' decisions about professional development (Desimone, Smith, & Phillips, 2007), state-level implementation of standards-based curricula (Clune, 2001), and student achievement on various cognitive domains (Desimone et al., 2004). State policies also appear to affect the degree of district policy innovation (Porter, Floden, Freeman, Schmidt, & Schwille, 1988; Spillane, 2004) and teachers' content decisions (Porter et al., 1988), though the research on content decisions pre-dates much of the implementation of standards-based reform. In the recent literature on SBR's effects on the content of instruction, there is little evidence on the ways in which the variation in state policy is related to teachers' decisions about content.

The only study that directly compared teachers across multiple states on the basis of state policy effects on content was a national survey of over 4,000 teachers (Pedulla et

al., 2003). The study first characterized each state in terms of the stakes for teachers/schools and the stakes for students associated with state tests. High stakes for teachers/schools meant that the state assigned rewards or sanctions on the basis of student test scores. Medium stakes meant that the state published data on student performance, for instance in the newspaper, but did not assign rewards or sanctions. Low stakes meant that test results were neither published nor attached to rewards or sanctions. For students, high stakes meant that student promotion or graduation was attached to test results; medium stakes meant that student test results were included on the students' transcripts; and low stakes meant that test results were neither used for graduation/promotion or student transcripts. On the basis of these classifications, states were clustered primarily in five of the nine possible cells—high stakes for schools and high, moderate, or low stakes for students; or moderate stakes for schools and high or low stakes for students.

Given these classifications, teachers were randomly sampled, stratified by classification, and surveyed about their opinions of standards-based reform and its effects on their instruction. A total of 4,195 of 12,000 teachers responded, for a response rate of 35%. Teachers were asked about their coverage of tested subjects and non-tested subjects. The degree of teachers' reported increase in coverage of tested subjects was significantly larger for teachers in states where there were high stakes for students (mean = 3.80 to 3.85 on a scale of 1 = decreased a great deal to 5 = increased a great deal) than for teachers in states with low or moderate stakes for students (mean = 3.47 to 3.65). The degree of teachers' reported decrease in coverage of non-tested subjects was also greater in states with high stakes for students (mean = 2.66 to 2.74) than states with low or

moderate stakes for students (mean = 2.85 to 3.01). Within a particular level of stakes for students, the decrease in coverage of non-tested content was also greater in states with high stakes for schools/teachers than in states with moderate or low stakes for schools/teachers. However, there were no overall differences in coverage of tested content across levels of stakes for schools/teachers.

In terms of alignment, responses on a seven-item scale (1 = strongly disagree to 4 = strongly agree) indicated that teachers in states with high stakes for schools/teachers reported greater alignment (mean = 2.56 to 2.71) than teachers in states with moderate stakes for schools/teachers (mean = 2.54). However, there were no differences in teacher-reported alignment across levels of stakes for students. Overall, the results of this study were suggestive that the degree of stakes for students and schools/teachers was related to the degree of content effects, in all cases with a significant difference finding greater effects in high-stakes states.

Few other studies provide an evaluation of the relationship between state policy attributes and the degree of content effects. In McDonnell's (2004) study of teachers in Kentucky and North Carolina, teachers were asked to submit assignments that were "most aligned" with state assessments. When those assignments were content analyzed by trained raters, only 32% of North Carolina teachers' assignments were rated as being similar to the assessment, as compared to 53% for Kentucky teachers. While there are many possible explanations for this difference, one is that the stakes for schools attached to assessments in North Carolina were much lower than those in Kentucky. A survey study of teachers in California, Georgia, and Pennsylvania (Hamilton & Berends, 2006)

found that, while most teachers in all states reported aligning their instruction to standards and assessments in English, math, and science, the proportions were higher in California and Georgia (90 to 95% in math, 82 to 96% in science) than in Pennsylvania (70 to 82% in math, 45 to 57% in science). The authors proposed that alignment to standards and assessments in science in Pennsylvania was so low because that state had not yet implemented a testing system in science. The lower proportion of teachers reporting alignment to standards and assessments in mathematics was attributed to either the newness of the standards-based reform policies in Pennsylvania, the historic degree of local control in that state, or the fact that mathematics was tested in only four grades as opposed to all grades 3-8 (Stecher & Naftel, 2006). While these studies are suggestive of potential relationships between the policy attributes of power and stability and aligned instruction, the low numbers of states and large numbers of potential alternative explanations limit the strength of these conclusions.

The remaining studies identified in the literature focus on surveys or qualitative methodologies in individual states. While there are differing policy contexts identified across these studies, the lack of common methodologies, survey questions, and reporting metrics do not allow for comparison of effects across policy contexts. Overall, there is some evidence that certain state policy attributes, particularly power, may be related to differences in the content of teachers' instruction. However, the evidence is far from overwhelming, and the magnitude of any such effects is unclear. Finally, few policy attributes have been investigated. Thus, an important contribution of this study is the

examination of multiple state policy attributes and their relationships with the content of teachers' instruction.

Summary of Literature Review

The purpose of this section was to situate this dissertation in the literature on standards-based reform and its effects on the content of teachers' instruction. First, the theory of change underlying standards-based reform was described. It was argued that the first causal step in the theory of change is that teachers should modify the content of their instruction to align with state standards and assessments. Next, the literature on SBR's effects on the content of teachers' instruction was described. Two primary effects were identified—a shifting of content from non-tested to tested subject areas and an increase in teachers' self-reported alignment of instruction to standards and assessments. The quality of the research on content alignment was challenged on methodological grounds, and it was argued that the research did not provide useful information to policymakers about the relationship between state policy and content effects. Finally, the limited literature on state policy effects on the content of teachers' instruction was described in some detail. There was some evidence that teachers in states with policies emphasizing power reported greater content effects. There was also one study that suggested that teachers' self-reported alignment might increase as the SBR policy remains stable. The literature did not support conclusions about other policy attributes.

Chapter 3

Methods and Data

The literature review on the effects of standards-based educational policies on the content of teachers' instruction suggested large effects. In many survey studies, 80% or more of teachers reported aligning the content of their instruction with standards and/or assessments (e.g., Hamilton & Berends, 2006; Koretz, Barron et al., 1996), and large majorities of teachers reported shifting the content of their instruction from non-tested to tested subjects (e.g., Koretz, Barron et al., 1996; Koretz, Mitchell et al., 1996; Stecher et al., 2000). Clearly, whether or not teachers are accurately interpreting the content messages of standards and assessments, they perceive great changes in the content of their instruction as a result of these policy instruments. However, the literature was weak in several regards. For one, surveys often relied on teachers' self-reported change in instruction, a potentially problematic measure (Desimone et al., 2005; Lam & Bengo, 2003; Ross, 1989; Ross & Conway, 1986; Schwartz & Oyserman, 2001), and teachers' causal attribution of that change to the standards-based reform policy. For another, teachers' reports about increased content alignment were often based on one-item questions which are likely to have low reliability (Mayer, 1999). Questions about content alignment also assumed that teachers understood the content messages of standards and assessments, which is likely not the case (e.g., Hill, 2001; Spillane, 2004). The studies were also limited in that they did not report magnitudes of the effects except in percentages of teachers. Finally, with few exceptions (e.g., Hamilton & Berends, 2006; Pedulla et al., 2003), the studies did not allow for comparison of effects across state

policy contexts, limiting their utility in terms of understanding the reasons why teachers changed the content of their instruction.

The research presented here was designed to investigate the issue of SBR's effects on the content of teachers' instruction in a way that addresses some of the shortcomings of previous research and allows for the identification of specific attributes of state policy that are related to effects on teachers' content decisions. The purpose of this chapter is to report the conceptual framework that guided this research, delineate the data and variables used to address the research questions, and describe the specific analytical methods used.

Research Questions

Three research questions guided this research:

- 1) To what extent have teachers aligned their instruction with standards and assessments in English language arts and reading, mathematics, and science, since 2003?
- 2) In what ways have teachers changed the distribution of instructional time and the topics and cognitive demand levels of their instruction in the years since 2003, and to what extent are the changes in topics and cognitive demand levels reflective of the content messages of standards and assessments?
- 3) To what extent are state policy contexts, features of standards and assessments, and features of local context related to alignment?

Theoretical Frameworks for the Study of Standards and Assessments

Theories about the potential effects of standards and assessments come from multiple disciplinary perspectives, including management (Rowan, 1996), political science (McDonnell, 2004), and institutional/organizational theory (Elmore, 1996). For the proposed analysis, two frameworks (Porter et al., 1988; Supovitz, 2009) were combined and organized to create a broad model for investigating the effects of state standards-based reform policies on teachers. The conceptual framework applies not only at the level of the formal policy as written, but also at the level of the policy's perceived influence on teachers. That is, features of the policies themselves and teachers' perceptions of the policies are seen as important in impacting teachers' instructional decisions.

Supovitz's (2009) framework describes the four categories of theories underpinning the use of standards-based reform. The first is alignment theory, or the belief that test-based accountability can help tighten the various parts of the education enterprise. The second are motivational theories, which propose that the incentives created by standards-based reform will motivate improvement among the actors in the educational system. The third is informational theory, or the belief that the provision of information from standards-based assessments will encourage improvement. The fourth is symbolic theory, or the idea that the accountability system signals important values to educators, parents, and communities. A framework proposed by researchers at Michigan State's Institute for Research on Teaching (Porter et al., 1988) provides important details to fill in the key variables in Supovitz's organizational scheme by proposing five policy

attributes: specificity, consistency, authority, power, and stability. Specificity and consistency fit in the category of alignment, while authority and power fit in the category of motivation. Stability underlies all of Supovitz's categories. Together, the two frameworks can organize thinking about the effects of state standards and assessment policies on teaching and, ultimately, student learning.

Alignment. Perhaps the most prominent theories about the effects of educational reform are based on ideas of coupling or alignment. Even the title given the reform—"systemic"—invokes a coherent effort operating throughout the system. The first feature of alignment is specificity, or the degree to which the policy is clear and promotes common understanding of the policy's intent (Clarke et al., 2003; D. K. Cohen & Ball, 1990; D. K. Cohen & Spillane, 1992; Finn & Kanstoroom, 2001; Porter et al., 1988; Ravitch, 1995; Schwille et al., 1988; Schwille et al., 1982; Schwille, Porter, & Gant, 1980). More specific standards, or ones that are less vague and less open to misunderstanding, are more likely to be implemented as intended. More specific standards are also more able to be measured with a high degree of reliability and validity, and thus better able to be part of an aligned standards-based system. In this study, specificity is operationalized by the degree of *focus* of the state standards—do they cover a few topics in great depth or many topics shallowly?

The second feature of alignment is consistency, or the extent to which the various parts of the policy are coherent with one another and provide the same message to the implementer (Clarke et al., 2003; D. K. Cohen & Ball, 1990; D. K. Cohen & Spillane, 1992; Cuban, 1990; Darling-Hammond, 1990; Finn & Kanstoroom, 2001; Porter et al.,

1988; Raizen, 1988; Schwille et al., 1988; Spillane, 2004; Sykes, 1990). Of all the attributes, consistency is the most fundamental to standards-based reform. The consistency of the policy message is seen as being critical in K-12 education's fragmented governance system (M. S. Smith & O'Day, 1991). In this study, the consistency of state policy is operationalized as the *consistency* (alignment) of state standards with state assessments of student learning.

Consistency applies not only at the level of the formal policy, but also at the school and classroom level, impacting teachers' perceptions of the coherence of the system. Decades of educational research have identified features of schools and teaching that counteract rationalizing trends such as those embodied in standards-based reform. Foremost among these are the conflicting messages teachers receive about the content and practices they are to be using in their classrooms (M. S. Smith & O'Day, 1991). Organizational and structural features of schools, such as the cellular nature of teaching, the poor socialization of teachers into the profession, and the transient nature of many teachers' commitments, have contributed to an inability to achieve large-scale educational reform (Little, 1990; Lortie, 1975). Similarly, some have argued that teaching rarely changes in any meaningful way, because the governance of education is too complex and cumbersome to provide consistent messages to teachers (D. K. Cohen & Spillane, 1992).

Standards-based reform was intended to overcome these obstacles (M. S. Smith & O'Day, 1991), but that intention may or may not have been enacted. In short, teachers are subject to a wide variety of external pressures that could affect their decisions about what

and how to teach, not all of which will necessarily be aligned with the spirit of the standards and assessments. In the absence of some external forces removing or mitigating these conflicting messages, there is good reason to believe that little meaningful change will take place. To capture consistency at the level of teacher perceptions, this study uses survey data from teachers on the extent to which their instruction is *influenced by SBR-type sources* such as standards and assessments, as opposed to being *influenced by other sources* such as parent desires.

Motivation. Along with alignment, theories of motivation are widely cited as explaining effects of standards-based policies. Motivation can be intrinsic to the teacher or created through extrinsic sources. Teachers' intrinsic motivation can be affected through authority (Spady & Mitchell, 1979). Authority is based on the personal belief in the legitimacy of the policy and is classified in four types: traditional, based in history, customs, and values; charismatic, based on the affective relationship between subordinate and superior; legal, based on rules and laws; and expert, based on possession of specialized knowledge or skills. Authority is widely seen as being important in the implementation of standards-based reform policies (Clarke et al., 2003; D. K. Cohen & Spillane, 1992; Porter et al., 1988; Schwille et al., 1988; Schwille et al., 1980; Spillane, 2004). The *authority* of standards-based reform policies is estimated by examining state standards documents and determining the extent to which they have attributes that increase their authority, such as mentioning teacher involvement in the development process or citing laws or statutes.

An alternative to intrinsic motivation of teachers to enact the standards is external motivation. External motivation is thought to be achieved through the application of power (Clarke et al., 2003; D. K. Cohen & Spillane, 1992; Porter et al., 1988; Schwille et al., 1988; Schwille et al., 1982; Spillane, 2004). Power, often described as ‘rewards and sanctions,’ is the use of resources, controlled by the superior, to motivate the subordinate (Spady & Mitchell, 1979). Power is an important part of NCLB—schools face reconstitution and closure threats if they do not meet acceptable targets for student achievement. In this case, the threat of sanctions is supposed to incentivize schools to improve (Supovitz, 2009). Power can also be exerted on teachers indirectly, through the establishment of requirements for student grade-to-grade promotion or graduation. There can be a trade-off between authority-based and power-based approaches (McDonnell, 2005). Here, *power* is measured through the extent to which the state has policies that provide rewards or sanctions to schools, teachers, students, or teacher education programs on the basis of certain measures of performance.

Information. A third theory about the effects of standards-based policy systems is based on information (Supovitz, 2009). According to this theory, standards-based systems will work if they provide teachers the information they need to improve their practice. Curriculum policy is based on the notion that knowledge about how best to instruct children is incomplete. That is, standards-based policies will provide teachers with the missing information they need to make better instructional decisions. Data are seen as important for schools’ and teachers’ instructional improvement (Boudett, City, & Murnane, 2005; Holcomb, 1999). For instance, the results of standards-based assessments

of student learning could give teachers a more complete set of information about their students' performance, including strengths and targets for improvement.

Information is also important for the sense-making that Spillane (2004) describes as critical in the district- and teacher-level interpretation and implementation of standards. Furthermore, confirming the systemic notion presented by Smith and O'Day (1991), Spillane argues that additional supports, such as professional community and high-quality, coherent professional development, are necessary for successful implementation. Other research confirms that incomplete information, often in the form of unclear language in standards-based reform policies and policy documents (D. K. Cohen & Ball, 1990; Hill, 2001) or unclear or inaccessible results from student assessments (Karr, Marsh, Ikemoto, Darilek, & Barney, 2006; Wayman & Stringfield, 2006), can deter successful implementation of standards-based reform policies. For this study, information is captured by two variables. The first is the extent to which the teacher has extensive *professional development on standards-based reform* or related instructional improvement. The second is the extent to which the teacher reports the existence of a *professional community* around instructional improvement.

Stability. The attribute of stability, identified in Porter and colleagues' framework, does not fit under just one of the theories in Supovitz's framework. Stability is the extent to which the policy has been in place for a long time (Hagge & Waltman, 2007; McMillan, Myran, & Workman, 1999; Porter et al., 1988; Raizen, 1988; Schwille et al., 1988; Spillane, 2004). Stability is an important underlying attribute that affects alignment, motivation, and information. For one, standards that are more stable give

teachers and other implementers a better opportunity to learn about, adapt to, and implement the policy, increasing the likelihood of alignment. Stable policies are also important for motivation, as forcing teachers to respond to rapidly changing instructional targets would likely be seen as demoralizing. Unstable policies also diminish the utility of the information provided by standards-based reform—if teachers receive feedback on student performance against some set of standards and those standards are rapidly changing, that information will be of limited value. In short, stability is an important feature of standards-based reform policies in that it supports the means by which those policies bring about change. For this study, *stability* is measured by the number of years the state standards have been in place.

Symbolism. Supovitz (2009) describes a fourth set of theories, grouped under the title of symbolism. The notion of symbolism suggests that standards-based systems are in place to prove to the public that educational institutions are accountable. Alternatively, standards and assessments might represent a symbolic validation of educational structures as they are, without requiring serious change (Airasian, 1988). Ravitch (1995) argues that standards and assessments are an important signaling device, indicating to various stakeholders what the meaning is of, for instance, a high school diploma. This theory fits well with Cuban's arguments about the vulnerability of educational institutions (1990). He claims that schools are the targets of reforms by dominant social groups because they offer a means of feigning reform that does not threaten the core of society. Schools have no means with which to "defend" themselves against these reforms—they are simply subject to wave after wave of reform, imposed upon them from above. Thus, schools and

teachers make superficial changes in their practices and structures but shield the instructional core from change. Systemic reform was explicitly designed to break this cycle by focusing on the instructional core of teaching and learning (M. S. Smith & O'Day, 1991). Symbolism is not included in the final framework described here, because if the symbolic interpretation of standards-based reform policies were true, we would not expect meaningful effects of standards-based reform on instruction. Instead, the standards and assessment system would be operating as a symbolic cover, whereby the public schools could appear focused on improvement without actually working toward it.

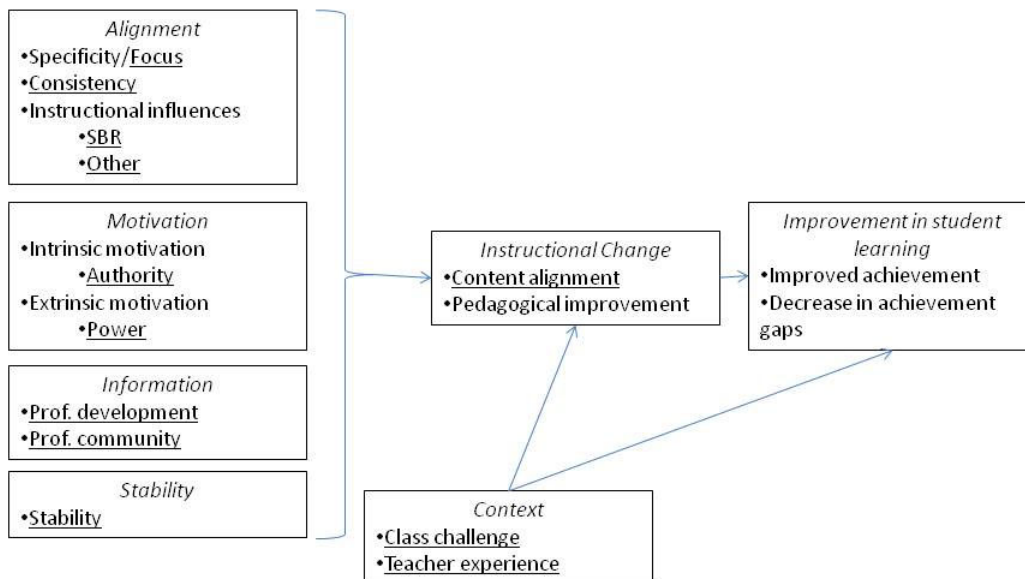


Figure 1. Theoretical framework for the study of standards-based reform.

Figure 1 shows a model of the theory of change for the impact of standards-based reform on instruction and student learning. The underlined variables are the ones measured and included in the models described below. These are described in more detail later. On the left are the main theories as to why standards-based reform should have effects. Based on the presence or absence of these features, we would expect variation in the degree of implementation of standards-based practices (e.g., adoption of content standards). These would impact the quality of student learning, which is not measured here. The impact on student learning would of course be expected to be greatest on closely aligned assessments, but the original design of standards-based reform proposed that effects on student learning would be substantial and generalizable (M. S. Smith & O'Day, 1991). Contextual variables are teacher experience and the challenge of the target class.

Data

The data for the proposed secondary analysis come from four sources. Data on teachers' instruction come from more than 30,000 collected surveys of teachers' instruction using the Surveys of Enacted Curriculum (SEC) content taxonomies (Porter, 2002). The teachers are from grades K-12 in mathematics, ELAR, and science. The data have been collected over the years 2003-2009 for use in research studies or by states, districts, or schools for evaluating instructional alignment. While some SEC data were collected prior to 2003, these data were not in a format that was compatible with the current framework, so they could not be used. Second, data on the content of standards and assessments also come from the SEC. Since 2002, trained content analysts have

analyzed state standards and assessment documents in the three content areas, using the same content taxonomies. Third, data on the power of state standards-based reform policies come from a database of state policies assembled from various published sources. Fourth, data on the authority of state standards-based reform policies come from examinations of state content standards documents.

The SEC grew out of the work of researchers at Michigan State University's Institute for Research on Teaching (Porter et al., 1988). Originally, a three-dimensional language was created for analyzing the content of instruction in mathematics. The dimensions were topics, levels of cognitive demand, and modes of presentation. Eventually, the third dimension was dropped and the others were modified. Meanwhile, languages were developed for science and ELAR. All of the content languages were developed by analyzing curriculum materials, state and district assessments, and standards documents, as well as through interviews with education professionals. The framework, now managed by the Council of Chief State School Officers, can be used by teachers to report the content of instruction. It can also be used by content analysts to analyze documents such as state content standards and assessments.

The data for both instruction and content analyses are in the form of two-dimensional matrices of topics by levels of student expectation, sometimes called levels of cognitive demand. For each subject area there is a list of broad topics (see Appendix A for the list of broad topics in the three content areas). Underneath each broad topic, there is a list of fine-grained topics. For instance, in science, one broad topic is Animal Biology, with fine-grained topics including Nutrition, Circulation, and Excretion, among

others. In total, there are 217 fine-grained topics in mathematics, 163 in ELAR, and 211 in ELAR. Perpendicular to the topics is a set of five levels of cognitive demand. These are slightly different across content areas, as seen in Appendix B.

The task for teachers completing the SEC is recommended to be done in three steps. First, teachers should identify all the fine-grained topics from the list that they did not teach to a “target class” in a particular time period. Often, the time period is a semester, but it can be any amount of time—results are always aggregated up to represent one full academic year. Once teachers have identified all the topics they have not taught to the target class, they identify, for each topic taught, how many lessons were spent on that topic, using a scale of 1 = slight coverage (less than one class/lesson), 2 = moderate coverage (one to five classes/lessons), and 3 = sustained coverage (more than five classes/lessons). Finally, for each topic taught, teachers identify the relative emphasis on each level of cognitive demand, using a scale of 0 = no emphasis (not an expectation for this topic), 1 = slight emphasis (accounts for less than 25% of the time spent on this topic), 2 = moderate emphasis (accounts for 25% to 33% of the time spent on this topic), and 3 = sustained emphasis (accounts for more than 33% of the time spent on this topic). Once the teacher has filled out enough content matrices to represent the full academic year, the data are turned into proportions of total content, where the value in each cell indicates the proportion of the total school year’s content that focused on the particular topic and cognitive demand level described by that cell. In addition to completing the content matrix, teachers complete a “Part A” survey with information on their classes, professional development, and school climate. These variables are also used here.

Data on teachers' self-reported instruction can be compared with content analyses of state standards and assessment documents to estimate alignment. The content analysis procedure for standards and assessments begins with the assembly of a content analysis team of three to five individuals, usually consisting of content experts in the particular subject area. The coders are trained; for training, the content analysts are first instructed on the basics of the content analysis procedures. They then complete sample codes and discuss confusing terms in the content framework, in order to establish a common understanding of the procedures and language. Finally, when coders indicate their readiness to proceed, the content analysis procedures begin.

The actual coding process is completed independently. For assessments, each analyst is responsible for examining every assessment item and assigning it to between one and three cells. Each item's weight on the assessment is then evenly divided among the assigned cells (e.g., if a three-point item is assigned to two cells, each cell receives 1.5 points). Respondents are allowed to flag items for discussion with the other coders, but the final decision on placing the item into a particular cell or cells is left up to the individual coder. For coding standards, each analyst examines the finest-grained level represented in the standards document (usually objectives). Because objectives are often "broader" than assessment items, the convention is that coders may place each objective in up to six cells. All objectives are weighted equally, and the weight for each objective is then divided evenly among the cells chosen. For both standards and assessments, the result for each coder is a matrix of proportions indicating the proportion of total content

in the document falling in each of the cells in the framework. The proportions are averaged, cell-by-cell, across raters.

One complication with the SEC data is that the content languages, both for the teacher surveys and for the content analyses of standards and assessments, changed between 2006 and 2007. Before 2007, there were five SEC content languages: math K-8, math 9-12, English K-12, science K-8, and science 9-12. In 2007, combined math K-12 and science K-12 frameworks were created, changing the list of available topics in each grade. In addition, small changes were made to the topic list for ELAR. The methods used to address these discrepancies are discussed in more detail below, but the evidence from most analyses suggests that the changes in the content languages had little effect on the substantive nature of the findings presented in chapter 4.

The third data source is a database of state policies on power, assembled from data provided in *Education Week's* annual "Quality Counts" series (Education Week, 2002, 2003, 2004, 2005, 2006, 2007, 2008, 2009), and other policy reports and databases (e.g., American Federation of Teachers, 2001; Center on Education Policy, 2009; Finn et al., 2006). The database is a modification of a database used in previous analyses of state policy (Desimone et al., 2004; Desimone et al., 2007). The variables in this database contribute to a measure of state policy attributes for power, which has been shown to be related to such outcomes as student learning and teacher participation in professional development (Desimone et al., 2004; Desimone et al., 2007). Other state-level variables for consistency, focus, and stability come from the content analyses of the standards and assessments themselves.

The final data source is the full set of state standards documents themselves. For each set of content standards in the database, the actual standards document was collected in PDF, Microsoft Word, or HTML form, and indices of authority were calculated by examining the standards documents against an authority rubric created by the author.

Alignment. The primary use of the SEC data and the content analyses of standards and assessments is for the calculation of alignment. Because all three provide data in the form of proportions summing to one across cells in the framework, the content coverage of teachers and the content of the documents can be directly compared. A simple alignment index (Porter, 2002) is:

$$\text{Alignment} = 1 - \frac{\sum_i |x_i - y_i|}{2}$$

where x_i is the proportion of content in cell i of matrix x (e.g., the matrix representing teachers' self-reported instruction) and y_i is the proportion of content in cell i of matrix y (e.g., the matrix representing state content standards). As the content languages can be used for teachers' instruction, content standards, assessments, curriculum (e.g., textbooks), or professional development, alignment can be calculated between any two.

The alignment index ranges from 0 to 1, with 0 indicating perfect misalignment (no content in common) and 1 indicating complete alignment (identical content in identical proportions). Because the alignment index is based on proportions, not absolute time, it is possible for alignment to be high, and yet, the amount of instruction might be vastly different (Porter, Smithson, Blank, & Zeidner, 2007). There is also no absolute standard for alignment—the indices are best used for normative comparisons. For instance, an analysis of the alignment of state standards across states indicated that, in

fourth grade ELAR, the alignment of standards documents ranged from .09 to .48, with a mean value of .24 (Porter et al., 2009). Thus, the two states aligned .48 had roughly 50% of their respective content in common, while the states aligned .09 had less than 10% of their content in common. Neither value is “good” in the sense of meeting some absolute threshold, but the states aligned .48 clearly have much more similar expectations for their students in fourth grade ELAR than the states aligned .09.

The alignment index has been used for descriptive analyses of state standards and assessments (Porter, 2002; Porter et al., 2009), as an independent variable in estimating student achievement and gains in achievement (Gamoran et al., 1997; Smithson & Collares, 2007), and as a dependent variable in a randomized experiment (Porter, Blank, Smithson, & Osthoff, 2005; Porter et al., 2007). In the third case, the treatment was a professional development sequence emphasizing teacher empowerment through receipt of data on their instruction and the alignment with state standards and assessments. Effects on curricular alignment were estimated, with one positive effect (effect size = .36) on alignment found for the alignment of mathematics instruction with state standards. Furthermore, the distribution of alignment of instruction with standards or assessments was seen to be symmetric and roughly normal, with means below .50. In the study reported here, alignment indices are used as descriptive variables and as both independent and dependent variables.

The interpretation of alignment indices can also be aided by the use of content maps, like the one shown in Figure 2. The maps, which are created in Microsoft Excel software, display the content in the standards, assessment, or instruction in graphical

form. The vertical axis on the maps contains topics, and the horizontal axis contains the levels of cognitive demand. The maps appear like topographical maps, with peaks representing focal topics and valleys representing topics not covered. Maps can be created for coarse-grained topics, such as the map in Figure 2. Or, maps can be created for fine-grained topics within a particular coarse-grained topic. Because both cognitive demand and topics are categorical variables, while the maps are completely accurate at the intersection of a topic and a cognitive demand, they cannot be interpreted in between categories.

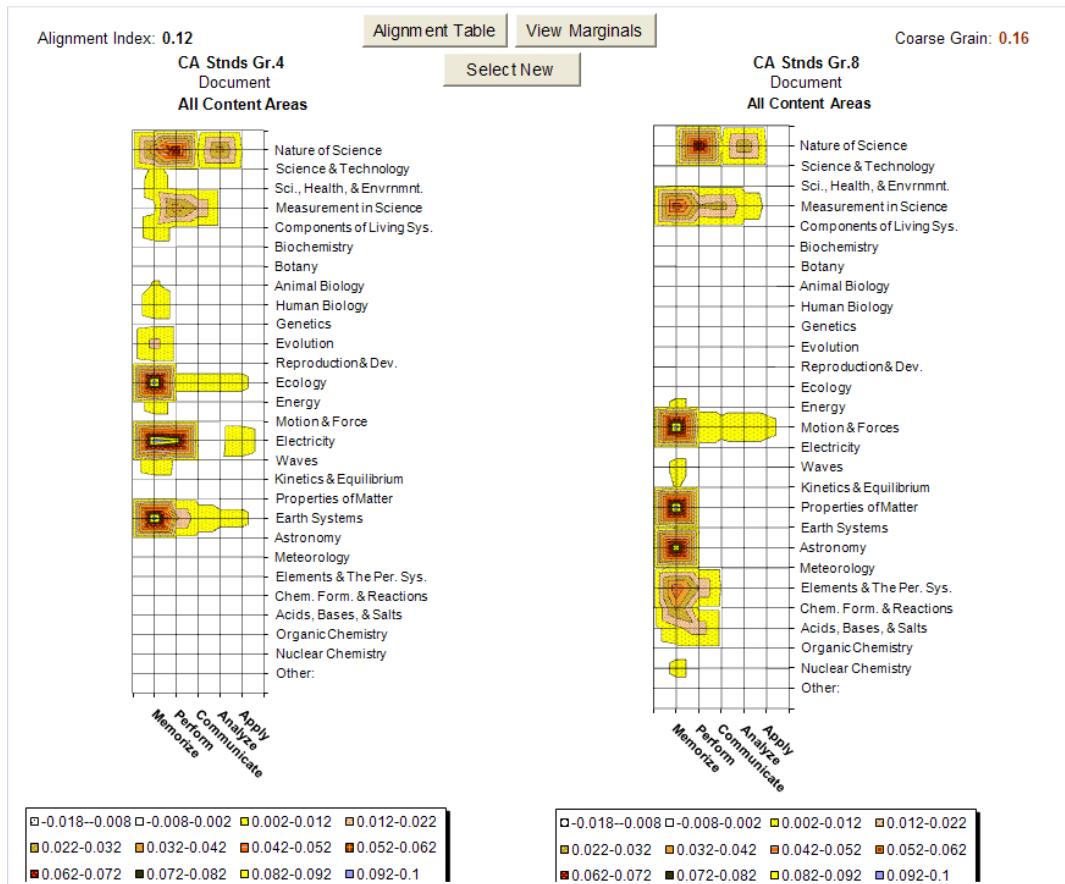


Figure 2. Illustrative content map.

Data Quality. The quality of both the teacher self-report data and the content analyses of standards and assessments have been investigated. The quality of the content analyses of ELAR and mathematics standards and assessments in two states at three grades each was studied using generalizability theory (Porter, Polikoff, Zeidner, & Smithson, 2008). The two-by-two-by-two-by-three design allowed for investigation of the extent to which reliability of the content analyses was dependent on state, subject, grade level, or document type. Across the 24 documents, the estimated reliability for three coders was .70 or higher in 20 of 24 cases. For four coders, the reliability was .70 or higher on 22 of 24 and .80 or higher on 18 of 24 documents. Thus, the content analysis procedures were deemed reliable, especially in cases where four or more raters were used. Unfortunately, there have been no analyses of the quality of content analyses of science standards and assessments. For the documents in the database used here, recent analyses (2008-2009) are based on four or more raters, while earlier analyses are generally based on three to five raters.

The quality of teacher self-report data was investigated in an earlier study of high school mathematics and science teaching (Porter et al., 1993). In that study, teachers filled out daily logs on the content of their instruction using a three-dimensional framework based on the SEC. Two of the dimensions were topics and cognitive demand, the two that remain in the SEC, though the particular topics and levels of cognitive demand were somewhat different in that taxonomy than in the current taxonomy. For 62 classes in the study, external observers also coded the lessons using the same framework. The agreement was calculated between the teachers and the external observers on several

dimensions. There was, on average, 68% percent agreement on the fine-grained topic taught and 59% average agreement on the cognitive demand level taught, though there were nine levels of cognitive demand in that taxonomy, as opposed to five here. Thus, at the level of the individual lesson, teacher report of content coverage exhibited substantial agreement with observer-rated content coverage. These results were viewed as lower estimates of reliability for content reported across a whole year, just as the reliability of one item on a student achievement test is lower than the reliability across many items.

The data were also studied by correlating teacher reports of the content of instruction as aggregated across the daily logs with reports based on end-of-semester surveys (Porter et al., 1993). The proportion of total instructional time on each coarse-grained content area was the variable studied. In mathematics, the correlations of the proportions between the aggregated logs and the end-of-semester surveys were .29 or higher for nine of ten coarse-grained topics and .50 or higher for five topics. In science, the correlations were .60 or higher for all but one of the eight coarse-grained topics. Cognitive demand levels were also investigated, but the set of cognitive demand levels on the end-of-semester surveys differed from those on the logs, so the results are not useful. Fine-grained topics were not investigated. Still, the results were seen as being supportive of the validity of the end-of-semester surveys in representing the content on the daily logs. No similar analyses have been conducted in ELAR.

Another piece of evidence for the validity of teachers' SEC responses is a study of the effect of content coverage on student learning gains (Gamoran et al., 1997). Student achievement gains on a constructed mathematics achievement test were regressed on the

alignment between the test and teachers' self-reported instruction. Results indicated correlations of .45 between content coverage and achievement gains. This evidence suggests that the SEC not only accurately measures the content covered in a classroom, but also that alignment of SEC-reported instruction with the content of assessments is predictive of differential change in student achievement across teachers, at least in mathematics.

Sample

The teacher data for this study are held at the Wisconsin Center for Education Research. The data have been collected throughout the years 2003-2009 for various research studies or for use by districts, schools, or individual teachers to examine the content of their instruction. The data have not been collected systematically, nor are they intended to be nationally representative. None of the data were specifically collected for this dissertation.

Table 2
Summarized Teacher Sample Sizes

Subject	Full sample		States with coded standards		States with coded assessments		States with coded standards and assessments	
	n	states	n	states	n	states	n	states
Math	16,395	32	10,814	23	5,071	12	4,748	11
ELAR	14,992	27	12,522	19	4,431	10	4,344	10
Science	9,553	30	5,346	22	1,837	12	1,666	10

Table 2 summarizes the samples by academic subject and type of available data. In all cases, the largest possible samples were used for each research question, in order to maximize power. Thus, the composition of the samples used for each research question

differ somewhat. The total teacher samples range from 16,395 for mathematics to 9,553 teachers for science. These are the samples used for the descriptive analyses (RQ2). When only teachers with alignment indices to standards or assessments are counted, there are between 5,346 and 12,522 for standards and between 1,837 and 5,071 for assessments. These are the samples used for calculating changes in alignment across years (RQ1). When only teachers with complete data (i.e., in states and grades with content analyzed standards and assessments) are counted, there are roughly 4,750 mathematics teachers in 11 states, 4,350 ELAR teachers in 10 states, and 1,650 science teachers in 10 states. These are the samples used for the investigations of state policy (RQ3). Clearly, the ELAR and mathematics samples are stronger than science. Nevertheless, across subjects there are large samples of teachers from multiple states.

Table 3
Teacher Sample Sizes by Year

Year	Mathematics				ELAR				Science			
	n	n with standards	n with assessments	n with both	n	n with standards	n with assessments	n with both	n	n with standards	n with assessments	n with both
2002/3	122	79	0	0	0	0	0	0	194	103	23	23
2003/4	1287	725	590	463	1108	881	329	324	439	239	132	110
2004/5	2696	1759	822	769	2187	1699	533	505	1738	1017	353	340
2005/6	3107	2182	1271	1227	3399	3027	1066	1050	2089	1105	452	402
2006/7	3868	2420	1094	1002	4186	3386	1184	1170	1897	1037	357	301
2007/8	3745	2664	734	728	2965	2599	1022	1007	2127	1241	329	301
2008/9	1570	985	560	559	1147	930	297	288	1069	604	191	189
Total	16395	10814	5071	4748	14992	12522	4431	4344	9553	5346	1837	1666

Tables 3, 4, and 5 describe the sample by year, grade band, and state respectively. As seen in Table 3, the teachers are distributed somewhat unevenly across years, with most teachers in the years 2004-05 to 2007-08. In all three subjects, in the full samples used for the descriptive analyses, there are more than one thousand teachers in each

subject and year 2003-04 to 2008-09 except for 2003-04 in science. For the restricted analyses used for the HLMs, the samples are much smaller across years—still 288 or more in all years 2004 to 2009 in mathematics and ELAR, but as low as 110 in science in 2003-04. There are no ELAR teachers in the year 2002-03, but a few mathematics and science teachers in that year. Clearly, whether for the descriptive or HLM analyses, the samples are strongest in the middle years of the data (2005 to 2008)

Table 4
Teacher Sample Sizes by Grade Band

Grade band	Mathematics				ELAR				Science			
	n	n with standards	n with assessments	n with both	n	n with standards	n with assessments	n with both	n	n with standards	n with assessments	n with both
K-2	1808	1333	105	105	6571	5876	0	0	826	356	0	0
3-8	10536	8179	4302	4053	6807	6134	4184	4124	5952	4281	1433	1408
9-12	4051	1302	664	590	1614	512	247	220	2775	709	404	258
Total	16395	10814	5071	4748	14992	12522	4431	4344	9553	5346	1837	1666

The distribution of teachers by grade band, shown in Table 4, is also variable across subjects. Grade bands are presented, rather than grades, because grade bands are used for the descriptive analyses in research questions 1 and 2. The grade bands are K-2, the grades prior to NCLB testing; 3-8, the main grades for NCLB testing; and 9-12, the high school grades. These grand bands were chosen for three reasons. First, it is believed that the grades within each band are subject to roughly similar influences from standards-based accountability due to the grade-level timing of NCLB testing requirements. Second, for the descriptive analyses, the statistical power to detect instructional changes would have been too low if individual grades were used. Third, the presentation and interpretation of results for each of the analyses would have been overly complicated with 13 sets of regressions for each subject for each analysis, as opposed to three sets.

For the descriptive analyses in question 2, the samples are strongest for mathematics grades 3-8 and 9-12, ELAR grades K-2 and 3-8, and science grades 3-8, with more than 4,000 teachers in each. For these descriptive analyses, there are more than 1,000 teachers in all grade bands except K-2 science. For alignment to standards in research question 1, the samples are strong in the same grade bands, with the exception of K-2 mathematics. In contrast, for alignment to assessments in question 1, the samples are approximately 4,000 each in mathematics and ELAR grades 3-8, approximately 1,400 in science grades 3-8, and less than 1,000 in other subjects and grades. The HLMs use data on all grades, as mentioned previously. Clearly, the sample is strongest for all analyses in grades 3-8.

As for the distribution of teachers across states, shown in Table 5, several conclusions are clear. For one, a large number of states are represented for the descriptive analyses and the analyses of changes in alignment to standards over time. There are 24 states with 100 or more teachers in mathematics, 15 states with 100 or more teachers in ELAR, and 18 states with 100 or more teachers in science in the sample for research question 2. The sample of teachers in the descriptive analysis of changes in alignment to standards also includes moderate numbers of states—23, including 14 with 100 or more teachers in mathematics; 19, including 11 with 100 or more teachers in ELAR; and 22, including 10 with 100 or more teachers in science. The samples for the HLM analyses include teachers from fewer states. For mathematics, the teachers in the HLM sample are concentrated in Indiana, Montana, Ohio, Oklahoma, and Oregon. For ELAR, the teachers

in the HLM sample are concentrated in Indiana, New York, and Ohio. For science, the teachers in the HLM sample are concentrated in Illinois and Oklahoma.

Table 5
Teacher Sample Sizes by State

State	Mathematics				ELAR				Science			
	n	n with standards	n with assessments	n with both	n	n with standards	n with assessments	n with both	n	n with standards	n with assessments	n with both
AL	519	352	0	0	2	0	0	0	1	0	0	0
AR	63	0	0	0	0	0	0	0	0	0	0	0
CA	116	74	0	0	0	0	0	0	316	282	0	0
CO	1	0	0	0	5	0	0	0	0	0	0	0
DE	339	324	0	0	255	239	0	0	121	115	0	0
FL	0	0	0	0	0	0	0	0	89	9	9	9
ID	735	396	274	124	628	620	0	0	263	194	0	0
IL	2738	1947	0	0	1299	945	0	0	3084	1987	695	695
IN	867	822	646	621	1058	1038	646	626	131	124	33	33
IA	256	0	0	0	243	0	0	0	255	0	66	0
KS	1	1	1	1	0	0	0	0	0	0	0	0
ME	239	84	58	4	352	197	212	197	363	92	92	92
MA	186	25	0	0	96	44	0	0	124	63	0	0
MI	953	433	0	0	475	222	0	0	173	146	0	0
MN	912	738	0	0	192	175	7	7	38	16	0	0
MS	316	60	74	60	22	0	0	0	497	210	0	0
MO	76	0	0	0	1	0	0	0	212	81	0	0
MT	708	563	563	563	157	39	47	39	238	50	44	40
NH	55	25	5	2	18	1	0	0	6	3	3	3
NJ	240	76	0	0	0	0	0	0	44	5	0	0
NY	69	0	0	0	2166	1717	1155	1155	51	0	0	0
NC	656	396	0	0	136	122	128	122	62	12	0	0
OH	1497	1497	1174	1174	6780	6780	2115	2115	344	305	109	109
OK	1771	1622	1475	1475	126	55	89	55	1293	1188	442	442
OR	1108	1010	617	617	419	281	0	0	233	85	84	84
PA	397	53	0	0	56	0	0	0	552	140	0	0
TN	252	0	0	0	370	0	0	0	44	0	0	0
TX	206	33	0	0	0	0	0	0	100	65	8	0
UT	0	0	0	0	6	6	6	6	0	0	0	0
VT	248	169	0	0	37	15	0	0	59	0	0	0
VA	2	0	0	0	2	2	0	0	7	0	0	0
WA	3	0	0	0	0	0	0	0	5	0	0	0
WV	438	0	59	0	1	0	0	0	43	0	0	0
WI	428	114	125	107	90	24	26	22	805	174	252	159
Total	16395	10814	5071	4748	14992	12522	4431	4344	9553	5346	1837	1666

Because of the differences in sample composition across years and the fact that the samples are not probability samples, there are two important questions that must be addressed. The first question is an internal validity question, which affects the

interpretation of the findings from research questions 1 and 2. This is, to what extent do the samples change in composition across years? The second question is an external validity question, which affects the extent to which the conclusions drawn in this research are applicable to U.S. teachers generally. This is, to what extent do the sample characteristics indicate that the samples are similar in composition to the population of U.S. K-12 teachers?

To address the first question, a set of descriptive statistics was created for the teacher samples using data from the SEC part A. These descriptive statistics were then individually regressed on the year of SEC administration. A significant coefficient on year of administration would indicate that the sample of teachers changed on that characteristic over the study period. A total of 25 sets of OLS regressions were run—3 subjects, 3 grade bands, and 3 samples (excluding ELAR K-2 and science K-2 for alignment to assessments, because there are no K-2 assessments in the database in those subjects). The three samples were the full sample with all available teachers, the sample with alignment to standards, and the sample with alignment to assessments.

Because the four variables used for comparison are in terms of scales that do not allow for exact point estimates, estimates were created by taking the midpoint of each scale point. Class size was on a scale of 10 or fewer, 11-15, 16-20, 21-25, 26-30, and 30 or more, which was rescaled to 8, 13, 18, 23, 28, and 35. Proportion of nonwhite students was on a scale of less than 10%, 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, and 90+%, which was rescaled to 5%, 15%, 25%, 35%, 45%, 55%, 65%, 75%, 85%, and 95%. Proportion of LEP students was on a scale of zero, less than 10%, 10%-25%, 26%-

50%, and more than 50%, which was rescaled to 0%, 5%, 17.5%, 38%, and 75%. And teacher experience was on a scale of less than 1 year, 1-2 years, 3-5 years, 6-8 years, 9-11 years, 12-15 years, and 15+ years, which was rescaled to 0, 2, 4, 7, 10, 14, and 18 years.

Table 6
OLS Regressions of Sample Descriptive Statistics on Year of Administration

Variable	Math			ELAR			ELAR		
	Sample 1	Sample 2	Sample 3	Sample 1	Sample 2	Sample 3	Sample 1	Sample 2	Sample 3
Grades K-2									
Class size	-0.11	-0.43*	0.42	0.13*	0.09	-	0.13	-0.10	-
Percent nonwhite students	-5.11*	-2.77*	-0.16	-1.71*	-1.79*	-	-2.43*	-2.05	-
Percent ELL students	-1.00*	-1.88*	-0.54	1.00*	0.90*	-	-0.40	-0.52	-
Teacher experience	-0.13	-0.36*	-0.54	-0.29*	-0.30*	-	-0.42*	0.00	-
n	1808	1333	105	6571	5876	0	826	356	0
Grades 3-8									
Class size	-0.35*	-0.01	-0.23*	0.08	0.32*	0.22*	-0.33*	-0.39*	-0.16
Percent nonwhite students	-1.33*	-1.73*	-0.26	2.42*	3.44*	2.16*	-2.49*	-2.00*	0.59
Percent ELL students	0.08	0.23*	0.23	1.07*	1.09*	0.89*	-0.47*	-0.65*	0.73*
Teacher experience	-0.08	-0.14*	-0.21*	-0.03	-0.17*	-0.24*	0.14*	0.10	-0.11
n	10536	8179	4302	6807	6134	4184	5952	4281	1433
Grades 9-12									
Class size	0.28*	0.33*	-0.19	0.01	-0.45	-0.64*	-0.05	-0.43*	-0.66*
Percent nonwhite students	-1.86*	-2.39*	0.33	3.93*	2.92*	2.67	0.08	-3.99*	1.11
Percent ELL students	-0.55*	-0.36*	-0.16	0.67*	-0.28	-0.29	0.02	-0.88*	-0.40
Teacher experience	-0.09	-0.51*	-0.43*	-0.27	-0.43	-0.63*	-0.04	-0.02	-0.14
n	4051	1302	664	1614	512	247	2775	709	404

Note. * regression coefficient different from 0, $p < .05$. Sample 1 is full sample, Sample 2 includes only teachers with alignment to standards, Sample 3 includes only teachers with alignment to assessments.

The results of the OLS regressions of descriptive statistics on year of administration are shown in Table 6. Of the 100 regression coefficients, 57 show a significant effect of year of survey administration, indicating that the samples change over time in significant ways. Part of the finding of large numbers of significant

coefficients is certainly due to the large samples of teachers in most of these models. By variable, the number of significant coefficients is 13 for class size, 17 for percent nonwhite students, 15 for percent ELL students, and 12 for teacher experience. The most troubling grade band and subject combinations in terms of having a large proportion of significant changes are ELAR K-2 and ELAR 3-8. The strongest are science K-2 and science 9-12. There is at least one significant change in all but three of the 25 samples.

Because of the relatively large sample sizes, the magnitudes of the coefficients are probably more useful than the significance levels. There is some consistency across most of the models, except for ELAR grades 3-8 and 9-12. In most models, there is a modest decrease in class size of .2 to .3 students per year. There is also a small decrease of perhaps .5% per year in the proportion of ELL students, a decrease of approximately 2% per year in the proportion of nonwhite students, and a decrease of roughly .2 years per year in teacher experience. In ELAR grades 3-8 and 9-12, in contrast, the trends seem to be the opposite. Here, there are increases of 2-3% per year in the proportion of non-white students, increases of .5% to 1% in the proportion of ELL students, and mixed changes for class size.

Unfortunately, these descriptive statistics suggest that the samples are changing in meaningful ways across years. The method used in this dissertation to attempt to control for changes in sample composition is fixed effects for state and grade, but results of parallel models to those in Table 6 using fixed effects find similar patterns of change with no consistent differences in magnitudes or significance levels (not presented). Whether these sample composition changes invalidate the findings of the descriptive analyses is

unclear. But it is at least clear that changes in sample composition should be taken into account as a potential alternative explanation of the descriptive findings presented below.

Table 7
Comparison of Sample Descriptive Statistics with National Averages

Variable	Sample 1	Sample 2	Sample 3	Sample 4	Population
Mathematics					
Percent nonwhite students	34.2%	33.6%	32.4%	33.3%	43.5%
Elementary (K-5) class size	20.5	20.6	20.3	20.2	20.4
Secondary (9-12) class size	21.5	21.7	20.2	20.3	24.7
Percent of female secondary (9-12) teachers	61.4%	63.2%	63.9%	64.5%	54.9%
Percent of teachers with highest degree Bachelor's	50.1%	50.1%	51.7%	50.4%	50.8%
n	16,395	10,814	5,071	4,748	
ELAR					
Percent nonwhite students	43.2%	43.8%	45.2%	44.5%	43.5%
Elementary (K-5) class size	18.7	18.7	18.6	18.5	20.4
Secondary (9-12) class size	21.9	22.0	22.0	22.1	24.7
Percent of female secondary (9-12) teachers	76.1%	78.1%	78.3%	79.3%	73.9%
Percent of teachers with highest degree Bachelor's	39.1%	38.3%	35.7%	34.3%	50.8%
n	14,992	12,522	4,431	4,344	
Science					
Percent nonwhite students	36.2%	37.0%	33.9%	34.9%	43.5%
Elementary (K-5) class size	21.1	22.0	22.0	21.9	20.4
Secondary (9-12) class size	22.6	22.1	22.1	21.8	24.7
Percent of female secondary (9-12) teachers	56.1%	61.3%	58.3%	64.0%	48.7%
Percent of teachers with highest degree Bachelor's	52.3%	55.8%	53.1%	53.8%	50.8%
n	9,553	5,346	1,837	1,666	

Note. Population data are from 2008 Digest of Education Statistics. Sample 1 is full sample, Sample 2 includes only teachers with alignment to standards, Sample 3 includes only teachers with alignment to assessments. Sample 4 includes only teachers with alignment to standards and assessments.

The second question to be answered about sample characteristics is an external validity question. This is, to what extent do the observed characteristics of teachers in the various samples used here match those of U.S. teachers? A set of descriptive statistics on the teacher sample drawn from the SEC Part A is shown in Table 7. These are compared

with national figures taken from the Digest of Education Statistics (Snyder, Dillow, & Hoffman, 2008). Because the Digest contains information on only certain variables, the descriptive variables used here are somewhat different from those used for the longitudinal comparisons of descriptive statistics. Statistical tests of differences are not presented—because of the large sample sizes, nearly all sample averages are statistically different from national averages.

The descriptive statistics highlight first that the four different kinds of samples are roughly similar in terms of their composition as measured on these five variables. In all three subjects, there is a slight increase in the proportion of teachers who are female as the samples become more restrictive (i.e., going from sample 1 to sample 4). On the other variables, however, there are no consistent patterns in the descriptive statistics across the different samples.

The second conclusion from the descriptive statistics in Table 7 is that the samples are close to national averages on most indicators. However, the figures do not match perfectly. For instance, the sample of mathematics classes and science classes both have 5% to 10% more white students than is typical of classes nationwide. In contrast, the ELAR classes have the correct amount of nonwhite students. In all three subjects, the class sizes in secondary classrooms are lower by two to four students than the national average. In contrast, the class sizes in elementary classrooms are very close to national averages in mathematics, slightly lower than average in ELAR, and slightly higher than average in science. In all three subjects, the samples of teachers are slightly more female than is typical. The difference is especially large in science. In mathematics and science,

the samples of teachers are approximately as educated as the national average. In ELAR, the sample is much more educated than is typical (i.e., fewer teachers with a top degree of Bachelor's, more with a top degree of Master's).

Overall, the comparisons of sample descriptive statistics with national averages suggest that the samples are not radically different from national figures on most indicators. However, there are some indicators, such as student race in mathematics and science and teacher education in ELAR that are moderately different than expected. Partly, these disparities may be due to imbalances in the states and grades of the sample teachers. Certainly, it would be better to have more descriptive statistics with which to compare these teachers with national averages. However, the results suggest that the classes represented in the dataset are not an extraordinary group in terms of teacher education level, teacher gender, class size, or student race, which is a promising finding that supports the external validity of the conclusions presented here.

Analysis

The multiple research questions lend themselves to multiple statistical analyses.

R.Q. 1: To what extent have teachers aligned their instruction with standards and assessments in ELAR, mathematics, and science, since 2003? Using data on the content of instruction and the content of standards and assessments, alignment indices are calculated for each teacher in the database, and these alignment indices are compared across years. Teachers are grouped by grade band, as in the presentation of descriptive statistics (K-2, 3-8, 9-12). Two approaches are taken for this research question.

This first approach is to separately regress the indices for alignment to standards and assessments in each subject on year, controlling only for the composition of the sample (i.e., state and grade, or the intersection of state and grade). The method chosen is fixed-effects regression, which accounts for differences in individual grades and individual states in terms of average alignment. The coefficient on the year variable then represents the average change in alignment from one year to the next within states and grades, averaged across states and grades. Only states and grades where the standards or assessments remained constant throughout the full study period are used in this approach. The model for the fixed effects regression is

$$Align_{ijk} = \beta_0 + \beta_1 * Year + u_j + f_k + \varepsilon_{ijk}$$

where YEAR is the year of SEC administration, u_j is a set of fixed effects for state, and f_k is a set of fixed effects for grade. Alternatively, one set of fixed effects could be used for the state/grade intersection. The model would otherwise be the same:

$$Align_{ijk} = \beta_0 + \beta_1 * Year + u_{jk} + \varepsilon_{ijk}$$

The coefficient β_1 indicates the relationship between alignment and year of administration, free from the effect of any state- and grade-level effects (u_{jk}). Both approaches are taken and the results compared.

A second approach is to look within particular states that are well-represented in the sample across years and where the standards or assessments remained constant across those years. Then a series of within-state models are run, where alignment is regressed on YEAR using fixed effects for grade:

$$Align_{ij} = \beta_0 + \beta_1 * Year + u_j + \varepsilon_{ij}$$

Here, u_j are fixed effects for grade. This analysis illustrates the changes in alignment at the state level, but says little about national changes in instructional alignment over the years. For the states where such data are available, these analyses are done.

Both of these approaches are used with the full dataset and again with a reduced dataset, where the list of eligible SEC topics is limited to only those topics that appeared in both the pre-2007 and post-2007 SEC surveys. The proportions in each of the cells in the SEC taxonomy are re-centered and alignment indices are calculated on the re-centered proportions. The results suggest whether the findings of the main analysis are affected in some way by the change in SEC form.

R.Q. 2: In what ways have teachers changed the distribution of instructional time and the topics and cognitive demand levels of their instruction in the years since 2003, and to what extent are the changes in topics and cognitive demand reflective of the content messages of standards and assessments? To illuminate the findings of the first research question, the second question provides descriptive information about the changes in teachers' instruction taking place during the study period. Furthermore, by comparing teachers' changes in instruction with the content of standards and assessments, it is possible to see whether the changes they have made are in line with the content messages of the documents.

Again, teachers are first grouped in the same grade spans for these analyses. The marginal proportions for each cognitive demand level are calculated for each teacher. Both the main SEC data and the re-centered data are used and presented separately, though the results are nearly identical. Fixed-effects regressions are run, with fixed

effects for grade and state; the coefficient on year indicates the change in cognitive demand coverage across time. Finally, the mean marginal proportions at the start of the study period are compared with the average marginal proportions in the standards and assessments across grades in each grade band. The results indicate the extent to which teachers made significant changes in the topics and cognitive demands they emphasize in their instruction during the study period, and whether or not those changes moved the content of their instruction into greater agreement with the standards and assessments to which they teach.

A parallel approach is used for analyzing coverage of coarse-grained topics in the SEC framework: 16 in mathematics, 14 in ELAR, and 27 in science. However, the analysis of topics is complicated somewhat by the change in SEC forms between 2006 and 2007. Thus, the primary approach is to use the re-centered data and conduct fixed-effects regressions by grade band, parallel to those described above for cognitive demand levels. For all those topics identified as having a significant change in coverage across time, the main data are then analyzed. The analysis is again fixed effects regression, but the main data are split into two halves—one for 2003 to 2006, and another for 2007 to 2009—and analyzed twice. In order for a coarse-grained topic to be reported as having a significant change, the requirement is a finding of significant change in the analysis with the re-centered data and a finding of significant change in the same direction using either the 2003-2006 or 2007-2009 main data, or both. This conservative approach ensures that any topic reported as having changed over the years 2003 to 2009 changed whether the data was the main data or the re-centered data.

Another form of content shift identified in the literature is a redistribution of instructional time to tested subjects and, to a lesser extent, away from non-tested subjects. While the SEC does not ask questions about coverage of other subjects, it does ask about instructional time in mathematics, ELAR, and science on the respective forms. To address this research question, another set of fixed-effects regressions is run with average hours taught per week in each academic subject as the dependent variable. The results indicate the extent to which teachers have shifted instructional time to, from, or among mathematics, ELAR, and science during the study period.

R.Q. 3: To what extent are state policy contexts, features of standards and assessments, and features of local context related to alignment and changes in alignment? The previous analyses provide descriptive results about the content of teachers' instruction under standards-based reform and the alignment of teachers' instruction with state standards and assessments. The final research question asks to what extent curricular alignment is related to teacher, school, and state policy variables.

To address this research question, three-level hierarchical linear models (HLMs) are used. A multi-level framework is necessary for this data in order to properly account for the clustering of teachers within grades, years, and states (Raudenbush & Bryk, 2002). Separate models are run for each subject. For each subject, the level-1 model is for teachers and the level-3 model is for states. The level-2 model is for individual grades and years within states. Teachers are nested within level 2 (e.g., a teacher's data is for a particular year in a particular grade in a particular state). In order to ensure proper standard errors, however, the level-2 data must be clustered within states. The three level

model is presented and described below. As recommended by Raudenbush and Bryk (2002), the first and simplest model is the fully unconditional model, shown below in Model 1.

$$\text{Level 1: } \text{ALIGN}_{ijk} = \pi_{0jk} + \varepsilon_{ijk}, \varepsilon_{ijk} \sim N(0, \sigma^2)$$

Model 1 $\text{Level 2: } \pi_{0jk} = \beta_{00k} + r_{0jk}, r_{0jk} \sim N(0, \tau_\pi)$

$$\text{Level 3: } \beta_{00k} = \gamma_{000} + u_{00k}, u_{00k} \sim N(0, \tau_\beta)$$

$$\text{Combined: } \text{ALIGN}_{ijk} = \gamma_{000} + \varepsilon_{ijk} + r_{0jk} + u_{00k}$$

where

ALIGN_{ijk} is the alignment of teacher i 's instruction with standards or assessments in state, grade, and year j and state k ;

π_{0jk} is the mean alignment in state, grade, and year j and state k ;

β_{00k} is the mean alignment in state k ;

ε_{ijk} is a random effect, assumed normally distributed with mean 0 and variance σ^2 ;

r_{0jk} is a random effect, assumed normally distributed with mean 0 and variance τ_π ;
and

u_{00k} is a random effect, assumed normally distributed with mean 0 and variance τ_β .

This model partitions the variance among the three levels, indicating how much of the variance in alignment is found at each of the three levels, according to the following equations.

$$\% \text{ variance at level 1} = \frac{\sigma^2}{\sigma^2 + \tau_\pi + \tau_\beta}$$

$$\% \text{ variance at level 2} = \frac{\tau_\pi}{\sigma^2 + \tau_\pi + \tau_\beta}$$

$$\% \text{ variance at level 3} = \frac{\tau_\beta}{\sigma^2 + \tau_\pi + \tau_\beta}$$

This model is evaluated six times—twice for each of the three subject areas, once each with alignment to standards and assessments as the dependent variables. Where there is significant variance to be accounted for at levels 2 and 3, the conditional models described below are also evaluated.

The conditional model includes information on state policies, features of standards and assessments, and local contexts as level-1 and level-2 variables. All level-1 variables are group-mean centered, and all non-dummy level-2 variables are grand-mean centered. The mathematics model is presented here, along with the relevant independent variables. Because the SEC Part A survey varies from one subject to another, some of the independent variables differ across models. These are explained at the end of this section.

$$\begin{aligned} \text{Level 1: } \text{ALIGN}_{ijk} = & \pi_{0jk} + \pi_{1jk} * \text{EXPERIENCE}_{ijk} + \pi_{2jk} * \text{CHALLENGE}_{ijk} \\ & + \pi_{3jk} * \text{SBRINFLUENCE}_{ijk} + \pi_{4jk} * \text{OTHERINFLUENCE}_{ijk} + \pi_{5jk} * \\ & \text{PROFCOMM}_{ijk} + \pi_{6jk} * \text{SBRPD}_{ijk} + \varepsilon_{ijk} \end{aligned}$$

$$\begin{aligned} \text{Level 2: } \pi_{0jk} = & \beta_{00k} + \beta_{01k} * \text{FOCUS}_{jk} + \beta_{02k} * \text{STABILITY}_{jk} + \beta_{03k} * \\ & \text{CONSISTENCY}_{jk} + \beta_{04k} * \text{AUTHORITY}_{jk} + \beta_{05k} * \text{POWER}_{jk} + \sum_{j=6}^n \beta_{0jk} * \text{GRADE}_{jk} \\ & + \sum_{j=n+1}^m \beta_{0jk} * \text{YEAR}_{jk} + r_{0jk} \end{aligned}$$

$$\pi_{1jk} = \beta_{10k}$$

$$\pi_{2jk} = \beta_{20k}$$

Model 2 $\pi_{3jk} = \beta_{30k}$

$$\pi_{4jk} = \beta_{40k}$$

$$\pi_{5jk} = \beta_{50k}$$

$$\pi_{6jk} = \beta_{60k}$$

$$\text{Level 3: } \beta_{00k} = \gamma_{000} + u_{00k}$$

$$\beta_{01k} = \gamma_{010}$$

$$\beta_{02k} = \gamma_{020}$$

$$\beta_{03k} = \gamma_{030}$$

$$\beta_{04k} = \gamma_{040}$$

$$\beta_{05k} = \gamma_{050}$$

$$\beta_{10k} = \gamma_{100}$$

$$\beta_{20k} = \gamma_{200}$$

$$\beta_{30k} = \gamma_{300}$$

$$\beta_{40k} = \gamma_{400}$$

$$\beta_{50k} = \gamma_{500}$$

$$\beta_{60k} = \gamma_{600}$$

$$\begin{aligned} \text{Combined: } \text{ALIGN}_{ijk} = & \gamma_{000} + \gamma_{010} * \text{FOCUS}_{jk} + \gamma_{020} * \text{STABILITY}_{jk} + \gamma_{030} * \\ & \text{CONSISTENCY}_{jk} + \gamma_{040} * \text{AUTHORITY}_{jk} + \gamma_{050} * \text{POWER}_{jk} + \sum_{j=6}^n \gamma_{0j0} * \\ & \text{GRADE}_{jk} + \sum_{j=n+1}^m \gamma_{0j0} * \text{YEAR}_{jk} + \gamma_{100} * \text{EXPERIENCE}_{ijk} + \gamma_{200} * \\ & \text{CHALLENGE}_{ijk} + \gamma_{300} * \text{SBRINFLUENCE}_{ijk} + \gamma_{400} * \text{OTHERINFLUENCE}_{ijk} + \\ & \gamma_{500} * \text{PROFCOMM}_{ijk} + \gamma_{600} * \text{SBRPD}_{ijk} + \varepsilon_{ijk} + r_{0jk} + u_{00k} \end{aligned}$$

where for mathematics teacher i in state, grade, and year j in state k ,

FOCUS is an index of the extent to which the state's standards in a particular

grade/subject are focused on a few big ideas. It is a composite of (a) the

number of cells in SEC framework that contain .556% or more (1 day in a 180-day year) of the total content in the standards, (b) the number of cells in the SEC framework required to account for 50% of the content in the standards, (c) the total number of cells covered, and (d) the proportion of total content included in the 10 most prominent cells (reverse coded). Each value is standardized within subject and the four values are averaged;

STABILITY is the number of years the state's content standards have been in place for the teacher's particular grade/subject;

CONSISTENCY is the alignment of the state's standards with its assessments;

AUTHORITY is an index of the authority associated with the state's standards-based reform policies, derived from an examination of state content standards documents. Electronic copies of the documents in Word or PDF form were obtained, and each document was analyzed as to whether or not it had each of nine characteristics: (a) the development process for the standards is described¹; (b) teachers participated in the writing of the standards; (c) content experts participated in the writing of the standards; (d) laws or statutes are mentioned in the document; (e) a professional organization (e.g., NCTM) is mentioned²; (f) the standards are claimed to be aligned with the standards of a professional organization; (g) the

¹ Standards documents that describe the development process should enhance the legitimacy of these documents with teachers because they increase the traditional authority of these documents through a description of the history of the standards.

² Many states mention national professional standards in their documents to say they were consulted in the development process, without claiming that the state standards are explicitly aligned with these national professional standards. Expert authority should be greater in these cases.

standards are claimed to be aligned with some textbook or other external resource (e.g., a curriculum guide); (h) a rationale is offered as to the importance of following the standards; and (i) certain objectives are explicitly identified as being focal standards more likely to be tested. The final characteristic is reverse coded, because identifying focal standards undermines the authority of the complete set of standards. For each characteristic, each standards document is rated with a 1 or a 0, and the indicators are added to form a total score for authority ranging from 0 to 9;

POWER is an index of the power of the state's accountability policies, derived from the database of state policies. Policies are (a) the state has a statewide student identification system³, (b) the state assigns ratings to schools based on additional state-developed criteria (i.e., in addition to the Adequate Yearly Progress ratings required by NCLB), (c) the state provides rewards to high-performing or improved schools, (d) graduation is contingent upon statewide exit or end-of-course exams, (e) end-of-course exams are based on state 10th grade standards or higher, (f) the state holds teacher education programs accountable for performance of graduates in the classroom setting, and (g) the state holds teacher education programs accountable by publishing pass rates/rankings of institutions. States are given a 1 if they

³ A statewide student identification system is an important technical element that enables the use of accountability measures.

have a certain policy in a certain year and a 0 if they do not. The indicators are added to form a total power score;⁴

GRADE is a set of dummy variables for grade;

YEAR is a set of dummy variables for year;

EXPERIENCE is a measure of the teacher's years of experience teaching the target class;

CHALLENGE is an index of the level of challenge associated with the teacher's target class. This is calculated using 4 questions. The first is the number of students in the target class, on a six-point scale. The second is the achievement level of most of the students in the target class, on a four point categorical scale. The third is the percentage of students in the target class who are Limited English proficient, on a five-point scale. The fourth is the number of different preps the teacher has. Each of the questions is collapsed to a three-point scale by combining categories, and the responses are averaged;

SBRINFLUENCE is an index of the extent to which the teacher reports that his/her instructional practices are influenced by state standards and assessments. This is calculated as the average of teacher responses to the degree to which their instruction is influenced by (a) [Their] state's

⁴ Some of the indicators in the POWER scale have one missing year of data, where *Education Week* did not report on those policies. Thus, two sets of POWER scales were created. The difference in the scales was as follows. If the missing year was 2007, one version of the scale would set the 2007 value for each missing policy to be the same as the 2006 value. The other version of the scale would set the 2007 value for each missing policy to be the same as the 2008 value. The two scales correlated .85 or greater with one another and produced nearly identical results in terms of magnitude and significance of effects, so the latter version of the scale was chosen and is the one used in all analyses presented here.

curriculum framework or content standards, (b) state tests or result from tests, (c) district tests or results from tests, and (d) national mathematics education standards. The questions are on a scale of 1 = strong negative influence to 5 = strong positive influence;

OTHERINFLUENCE is an index of the extent to which the teacher reports that his/her instructional practices are influenced by things other than state standards and assessments. This is calculated as the average of teacher responses to the degree to which their instruction is influenced by (a) their pre-service preparation, (b) students' special needs, (c) parental or community preferences, and (d) preparation of students for the next grade or level. The questions are on a scale of 1 = strong negative influence to 5 = strong positive influence;

PROFCOMM is an index of the extent to which the teacher reports his/her school has a professional community focused on instructional improvement. This is calculated as the average of teacher responses to questions on their opinion about several statements: (a) I am supported by colleagues to try out new ideas in teaching mathematics, (b) Mathematics teachers in the school regularly observe each other teaching classes, (c) Mathematics teachers in this school trust each other, (d) It's OK in this school to discuss feelings, worries, and frustrations with other mathematics teachers, (e) Mathematics teachers respect other teachers who take the lead in school

improvement efforts. The responses to all questions are on a scale of 0 = strongly disagree to 4 = strongly agree;

SBRPD is an index of the extent to which the teacher has had PD focused on standards-based reform principles. This is calculated as the average of teacher responses to questions on the amount of emphasis their PD placed on (a) State mathematics content standards, (b) Alignment of instruction to curriculum, (c) State or district mathematics assessment, (d) Interpretation of assessment data for use in mathematics instruction, and (e) classroom mathematics assessment. The scale for these questions is 0 = none to 3 = major. The mean response on the 0-to-3 scale was then multiplied by the total number of hours of PD the teacher reported taking;

The following scales were different for ELAR:

SBRINFLUENCE is an index of the extent to which the teacher reports that his/her instructional practices are influenced by state standards and assessments. This is calculated as the average of teacher responses to the degree to which their instruction is influenced by (a) [their] state's curriculum framework or content standards, (b) [their] district's curriculum framework, standards, or guides , (c) state tests or result from tests, and (d) district tests or results from tests. The questions are on a scale of 1 = strong negative influence to 5 = strong positive influence;

OTHERINFLUENCE is an index of the extent to which the teacher reports that his/her instructional practices are influenced by things other than state

standards and assessments. This is calculated as the average of teacher responses to the degree to which their instruction is influenced by (a) their pre-service preparation, (b) students' special needs, (c) parental or community preferences, (d) preparation of students for the next grade or level, (e) local priorities, directives, or policies, (f) their professional development experiences, and (g) screening, diagnostic, or classroom assessment results. The questions are on a scale of 1 = strong negative influence to 5 = strong positive influence;

PROFCOMM is an index of the extent to which the teacher reports his/her school has a professional community focused on instructional improvement. This is calculated as the average of teacher responses to questions on their opinion about several statements: (a) I am supported by colleagues to try out new ideas in teaching ELAR, (b) I receive support from the administration for teaching ELAR, (c) ELAR teachers in the school regularly share ideas and materials (d) ELAR teachers in the school regularly observe each other teaching classes, (e) Most teachers in the school contribute actively to making decisions about the curriculum, (f) I have adequate time during the regular school week to work with peers on ELAR curriculum or instruction, and (g) I have many opportunities to learn new things about teaching ELAR in my present job. The responses to all questions are on a scale of 0 = strongly disagree to 4 = strongly agree;

SBRPD is an index of the extent to which the teacher has had PD focused on standards-based reform principles. This is calculated as the average of teacher responses to questions on the amount of emphasis their PD placed on (a) State content standards, (b) Alignment of instruction to curriculum, and (c) State or district assessment (e.g., preparing, understanding, interpreting assessment data). The scale for these questions is 0 = none to 3 = major. The mean response on the 0-to-3 scale was then multiplied by the total number of hours of PD the teacher reported taking;

The following scales were different for science:

SBRINFLUENCE is an index of the extent to which the teacher reports that his/her instructional practices are influenced by state standards and assessments. This is calculated as the average of teacher responses to the degree to which their instruction is influenced by (a) [their] state's curriculum framework or content standards, (b) [their] district's curriculum framework, standards, or guides , (c) state tests or result from tests, and (d) district tests or results from tests. The questions are on a scale of 1 = strong negative influence to 5 = strong positive influence;

OTHERINFLUENCE is an index of the extent to which the teacher reports that his/her instructional practices are influenced by things other than state standards and assessments. This is calculated as the average of teacher responses to the degree to which their instruction is influenced by (a) their pre-service preparation, (b) students' special needs, (c) parental or

community preferences, and (d) preparation of students for the next grade or level. The questions are on a scale of 1 = strong negative influence to 5 = strong positive influence;

PROFCOMM is an index of the extent to which the teacher reports his/her school has a professional community focused on instructional improvement. This is calculated as the average of teacher responses to questions on their opinion about several statements: (a) I am supported by colleagues to try out new ideas in teaching science, (b) Science teachers in this school trust each other, (c) It's OK in this school to discuss feelings, worries, and frustrations with other science teachers, and (d) Science teachers respect other teachers who take the lead in school improvement efforts. The responses to all questions are on a scale of 0 = strongly disagree to 4 = strongly agree;

SBRPD is an index of the extent to which the teacher has had PD focused on standards-based reform principles. This is calculated as the average of teacher responses to questions on the amount of emphasis their PD placed on (a) State science content standards, (b) Alignment of science instruction to curriculum, (c) Instructional approaches (e.g., use of manipulatives), (d) In-depth study of science or specific concepts within science, (e) State or district science assessment, and (f) Interpretation of assessment data for use in science instruction. The scale for these questions is 0 = none to 3 = major. The mean response on the 0-to-3 scale

was then multiplied by the total number of hours of PD the teacher reported taking;

Table 8
Cronbach's Alpha Scale Reliabilities

Scale	Math		ELAR		Science	
	Alpha	# items	Alpha	# items	Alpha	# items
Focus	0.93	4	0.82	4	0.82	4
Authority	0.83	9	0.79	9	0.80	9
Power	0.94	7	0.94	7	0.94	7
Sbrinfluence	0.81	4	0.83	4	0.79	4
Otherinfluence	0.64	4	0.78	7	0.65	4
Profcomm	0.76	5	0.78	7	0.83	4
Sbrpd	0.89	5	0.81	3	0.89	6

Scale reliabilities were estimated using Cronbach's Alpha and are displayed in Table 8. Scales for SBRINFLUENCE, OTHERINFLUENCE, PROFCOMM, and SBRPD were created by analyzing SEC Part A data through exploratory factor analysis. Specific subsections of the survey were analyzed one at a time, and a PROMAX approach to oblique factor analysis was used. As is the convention, the number of factors for each factor analysis was set equal to the number of eigenvalues greater than or equal to one. Items were assigned to scales if they had factor loadings for those scales of .50 or greater, with all other factor loadings below .40. Scales emerging from the factor analysis largely conformed with a priori scales; however, the factor analytic scales were used due to their greater reliabilities.

All of the estimated coefficients are tested for statistical significance and reported on four levels: .10, .05, .01, and <.001. The null hypothesis is that none of the variables are related to the teacher's instructional alignment. While there are a large number of independent variables relative to the number of states, there are only five independent variables and several dummy variables at level 2. Furthermore, two of these variables (STABILITY and POWER) change across years for most states, and the other three change for those states whose standards and/or assessments change during the study period. Thus, concerns about oversaturation should be mitigated. Because previous studies (Desimone et al., 2004) indicate moderate to high correlations among some policy attributes (especially power and authority), the addition of level-2 variables to the model are examined using tests of improvement in model fit. Particular interactions of the policy attributes are also investigated.

Six such models are run, including one each for alignment to standards and assessments in each of the three content areas using only teachers with alignment data of both types. In all cases, teachers are excluded from the sample if the standards or assessments in the grade/state changed and the new standards/assessments are not included in the SEC database. In contrast, the values for the variables at level 2 are adjusted if the standards or assessments have changed and there are SEC data on the new version. That is, two teachers from different years in the same state and grade will have different values for level-2 variables if the content of standards changed in the interim and there are content analyses of both documents.

Variable	Mathematics		ELAR		Science	
	Mean	SD	Mean	SD	Mean	SD
Alignment						
Alignment to Standards	0.195	0.080	0.272	0.090	0.165	0.051
Alignment to Assessments	0.199	0.056	0.194	0.061	0.170	0.059
Level 2						
Focus	3.050	0.710	2.557	0.799	2.922	0.459
Stability	4.120	2.119	2.890	1.973	4.495	2.193
Consistency	0.298	0.113	0.220	0.054	0.261	0.065
Power	4.378	1.868	5.415	1.497	2.945	1.792
Authority	3.181	2.649	4.913	2.371	3.899	1.861
Level 1						
Experience	9.723	6.701	10.385	6.553	8.597	6.538
Challenge	0.798	0.330	0.804	0.341	0.907	0.365
SBRinfluence	3.350	0.694	3.294	0.757	3.034	0.739
Otherinfluence	3.063	0.522	3.130	0.528	2.886	0.559
Profcomm	2.608	0.654	2.402	0.707	2.880	0.740
SBRpd	60.707	88.857	85.717	98.053	86.264	108.650
Reform	1.131	0.794	1.613	0.865	1.753	0.937
Traditional	1.983	0.861	1.561	0.947	1.358	0.732
n	4197		3741		1551	

Descriptive Statistics

Descriptive statistics for the sample of teachers with alignment indices to standards and assessments (the HLM samples) are provided in Table 9. Average alignment is low for both standards and assessments—it is lowest in science, then mathematics, and finally ELAR. Alignment is highest and most variable between instruction and ELAR standards, and it is lowest and least variable between instruction and science standards. Teachers' average experience is roughly 8-10 years. In all subjects

teachers report that the content of their instruction is somewhat more influenced by standards-based reform-type influences (e.g., standards and tests) than other influences (e.g., parents).

Of note are the descriptive statistics for the policy attributes. While some of these have no meaningful scale (e.g., focus), others have more meaning. These attributes are described in more detail here, with examples given to illustrate high, low, and typical states.

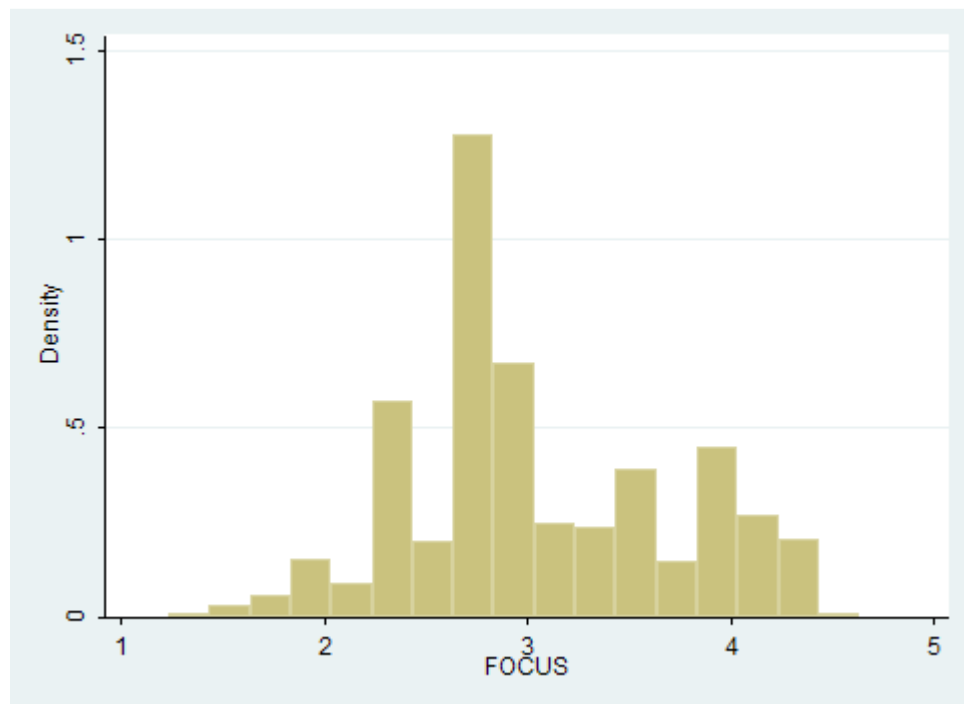


Figure 3. Histogram of FOCUS in science

Focus. The degree of focus of a state's standards was estimated by calculating four indices based on the content analyses of the standards documents. These four indices were then standardized and averaged. The average values were then re-centered to a mean of 3, so that there would be no negative values. A histogram of the focus variable in

science is shown in Figure 3—it reveals a slightly flattened distribution with a peak just below the mean. The focus variable ranges from .14 to 4.99 in mathematics, 1.34 to 5.38 in ELAR, and 1.24 to 4.57 in science.

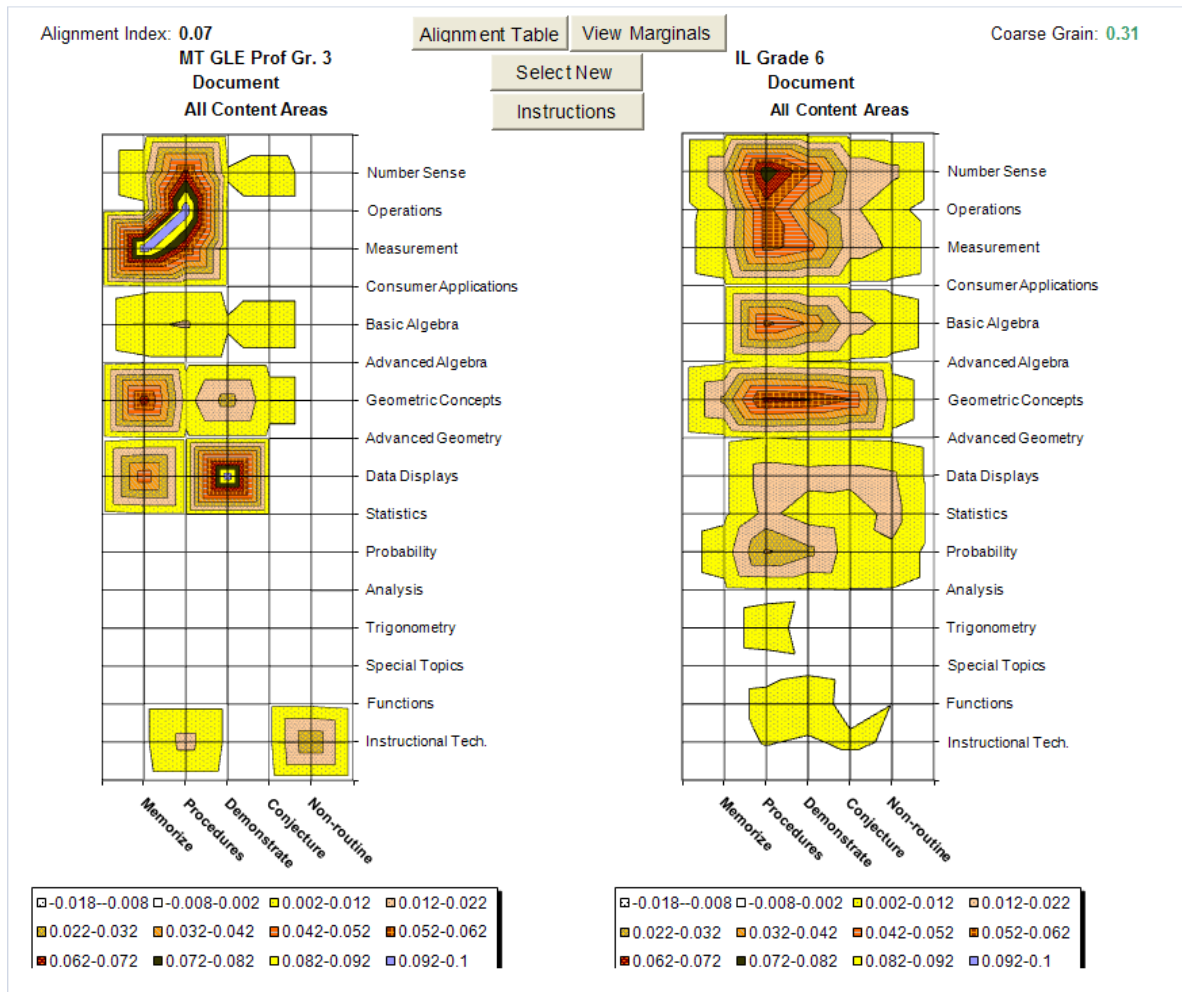


Figure 4. Content maps for the most and least focused standards in mathematics.

Again, because the FOCUS variable has no meaningful scale, it is useful to examine what high- and low-focus standards look like. The least focused state standards in mathematics are Illinois's sixth grade standards, and the most focused are Montana's third grade standards. A content map illustrating these two states' standards is shown in Figure 4. While Montana's standards have major focuses on procedures involving

operations and memorization of measurement, Illinois's standards have no such major focus. Instead, Illinois's standards are spread out across all levels of cognitive demand and all but five of the coarse-grained content areas. While Montana's grade three standards capture half of their content in just three cells and cover 83% of the curriculum in the ten most prominent cells, Illinois's grade six standards capture half their content in 62 cells and contain just 13% of the curriculum in the ten most prominent cells. These extremes indicate the wide variation in the degree of focus of state standards that is measured by the focus variable.

Stability. The stability of the standards is measured as the length the standards document has been in place. For the teachers in this sample, stability ranges from 0 to 9 years in ELAR and from 0 to 10 years in mathematics and science. Average stability was 3 to 5 years across subjects, with science standards the most stable during the study period.

Consistency. The alignment of a state's standards with its assessments defines the consistency of the standards-based reform policy. As seen in Table 9, the consistency of standards and assessments is generally low, though slightly higher and much more variable in mathematics than in ELAR and science. The range for consistency is .01 to .51 in mathematics, .07 to .36 in ELAR, and .09 to .37 in science. The lowest-aligned state standards and assessments in science are State U eighth grade, with an alignment of .09. Content maps of these standards and assessments are shown in Figure 5. For confidentiality purposes, the state cannot be identified.

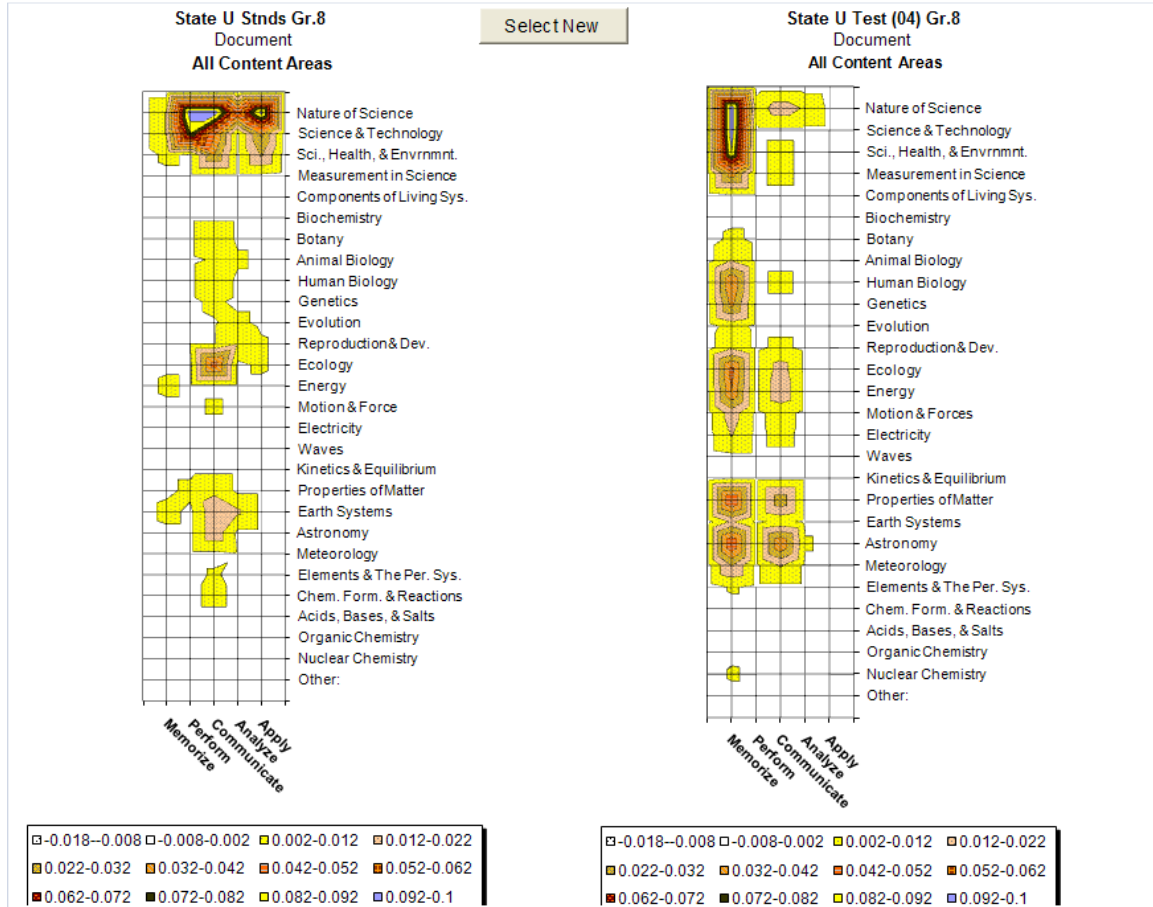


Figure 5. Content maps for the least consistent standards and assessments in science.

These content maps illustrate a highly inconsistent state standards-based reform policy. While there appear to be some overlaps in terms of coarse-grained topic areas (i.e., both the assessment and standards cover Nature of Science and Science & Technology), the assessment is concentrated almost exclusively on memorization. In contrast, the standards contain some content on procedures, communicating, and applying. Furthermore, the assessment tests several topics, such as Properties of Matter, Astronomy, and Energy that are hardly mentioned in the standards. In short, eighth grade teachers in this state would be receiving inconsistent messages from the standards and

assessments about the important content to teach, and it might be expected that the alignment of their instruction with standards and assessments might be lower as a result.

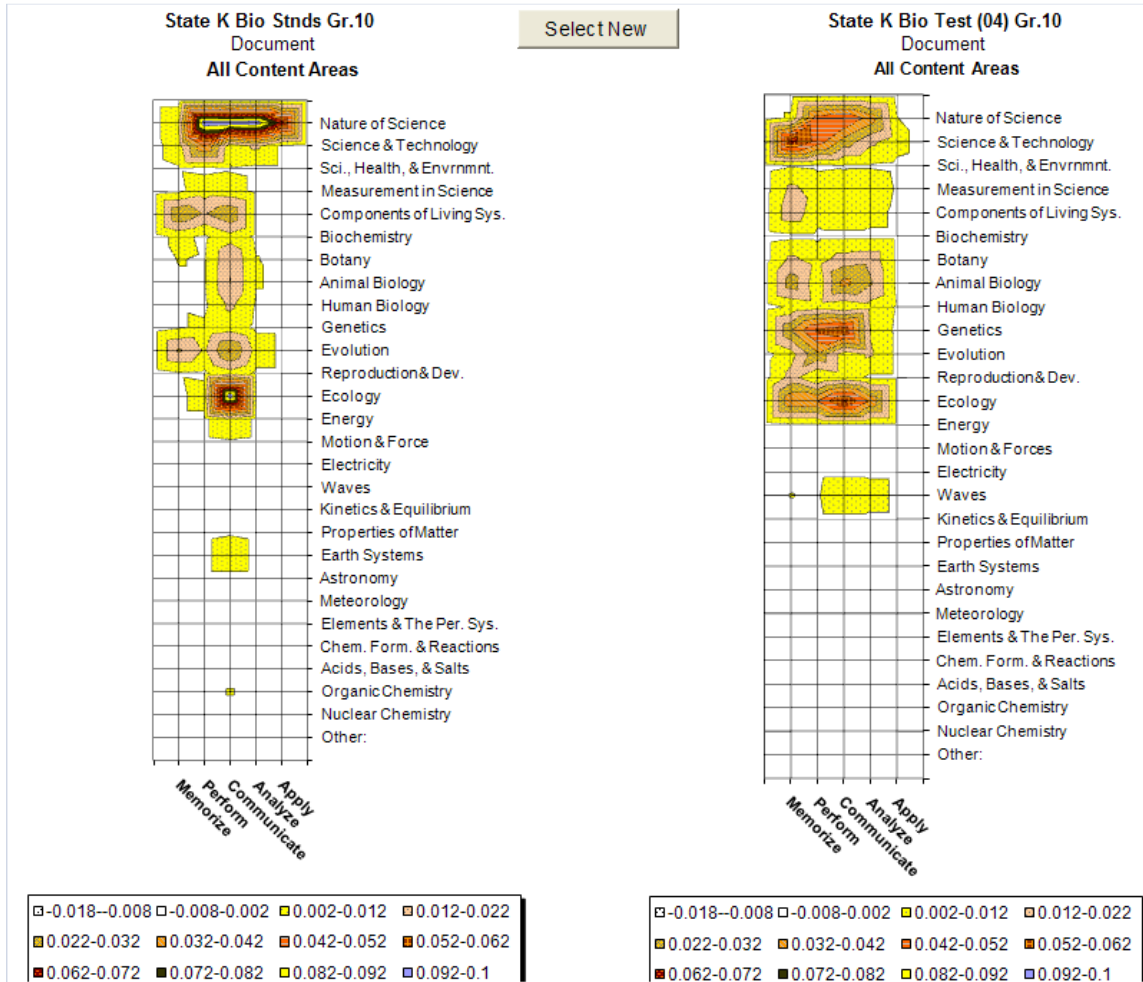


Figure 6. Content maps for the most consistent standards and assessments in science.

In contrast, Figure 6 shows a pair of content maps for the most consistent standards and assessments in science (alignment = .37). In this case, they are for high school biology in State K. The maps show that both the standards (on the left) and assessments have focuses on Ecology at the level of communicating, as well as Nature of Science at the levels of performing and communicating. While these are the most aligned standards and assessments in the database in science, there are still substantial areas of

disagreement. For instance, the test has a concentration in genetics that is hardly at all mirrored in the standards. Thus, while biology teachers in this state would be getting more consistent messages than eighth grade teachers in State U about what to teach, there would still be some conflicting ideas.

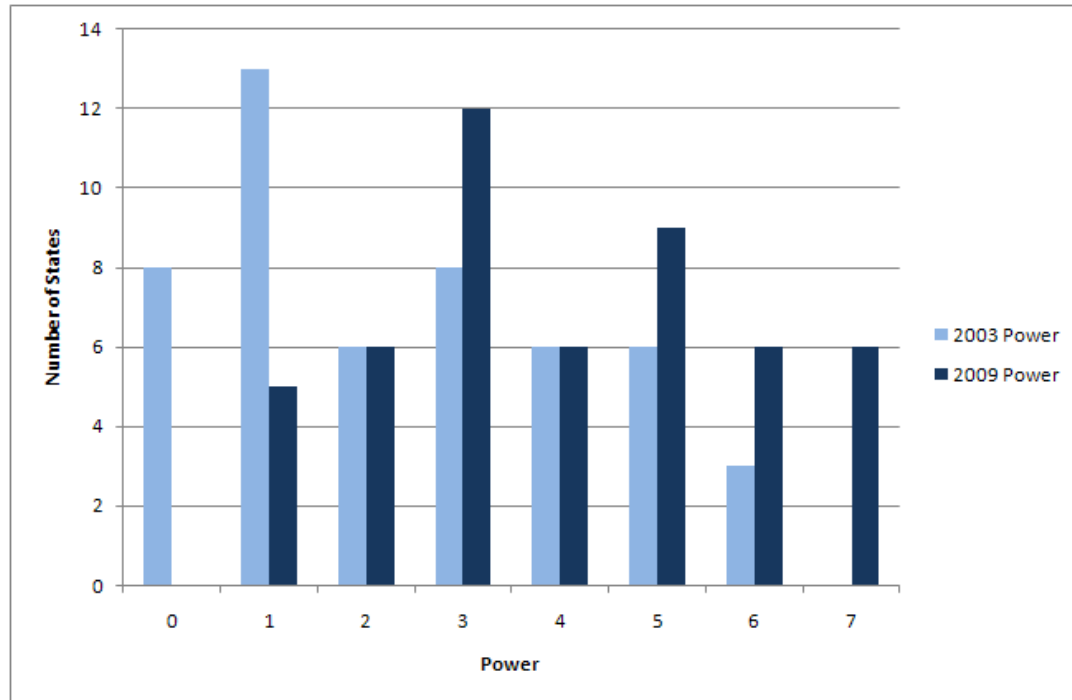


Figure 7. Histogram of power index for 50 U.S. states, 2003 & 2009.

Power. The power of a state's standards-based reform policies is operationalized as the extent to which the state has a number of policies relating to rewards and sanctions. There are seven policies, and the index is calculated for each year. Of the 50 U.S. states, four saw a decrease in power over the years 2003-2009, eight stayed the same, and 38 saw an increase. Figure 7, which shows a histogram of the power index in 2003 and 2009 for the 50 U.S. states, highlights the clear upward progression of power during the study

period. While eight states had a power index of 0 in 2003 and none had an index of 7, by 2009 six states had an index of 7 and none had an index of 0.

Of the states appearing in this study's sample, Ohio, New York, and Virginia had perfect scores of 7 by 2009. These states not only had graduation or promotion exams for students, but they offered rewards for high-performing or improved schools and they held teacher education programs and schools accountable in multiple ways. In contrast, Montana had a score of just 1 by 2009. The only power-related policy in place there by 2009 was that the state had a statewide student-identification system in place. Clearly, while there was a marked trend toward increased power across most states during the study period, the degree of power of state policies still varied 2009.

Authority. The final policy attribute, authority, was measured by examining state standards documents against a 9-point authority rubric. The authority index ranges from 0 to 7 in ELAR, 0 to 8 in mathematics, and 0 to 9 in science. Minnesota's ELAR standards (Minnesota Academic Standards Committee, 2003) are an example of a set of standards that earn a 0 for authority. This 58-page PDF document, which is available for download from the state Department of Education website, contains no information about how the standards were developed, who participated in the development, the importance of standards-based education, the relation of the standards with those created by national organizations of English teachers, or the relevant legal statute that requires the creation of the standards. Instead, the document contains a long list of objectives to be taught at each grade, with little supporting material.

In contrast to these are Ohio's ELAR standards (Ohio Department of Education, 2001), an even longer (320 page) document that contains not only objectives, but also detailed supporting information. For instance, the document says "the standards fulfill the requirement and timeline of Amended Substitute Senate Bill I for the State Board of Education to develop and adopt clear academic content standards ..." (page i), invoking legal authority. The document highlights the importance of standards, saying they "provide a set of clear and rigorous expectations for all students ... provide teachers with clearly defined statements of what students should know and be able to do ... [and] represent a research-based approach to literacy development" (p. 2). There is also a short description of the standards development process that mentions the role of educators on the development team (p. 4). Finally, the document provides an "Instructional Commentary" section that gives concrete advice for teachers on how to link the standards together and implement them in the classroom. In multiple ways, these standards represent a set of strong, authority-based expectations.

Correlations among the attributes. Previous research (e.g., Desimone et al., 2004) has noted moderate to high correlations among certain policy attributes. The correlations among the five policy attributes in each of the three subjects are presented in Table 10. The largest in magnitude is the correlation between FOCUS and STABILITY in English ($r = .64$), which suggests that standards that have been around longer tend to be more focused than newer standards. Other moderately large correlations are the correlation between FOCUS and POWER in English ($r = -.63$), which suggests that more focused English standards are found in states that do not have many policies emphasizing

power; and the correlation between power and authority in mathematics ($r = .62$), which suggests that states high on power tend to have mathematics standards high on authority.

Table 10
Correlations Among the Policy Attributes

	Focus	Stability	Consistency	Power	Authority
Mathematics					
Focus	1.00				
Stability	0.29	1.00			
Consistency	-0.25	-0.41	1.00		
Power	-0.49	-0.25	0.38	1.00	
Authority	-0.43	-0.42	0.25	0.62	1.00
ELAR					
Focus	1.00				
Stability	0.64	1.00			
Consistency	0.13	0.01	1.00		
Power	-0.63	-0.36	0.20	1.00	
Authority	0.03	0.17	0.46	0.22	1.00
Science					
Focus	1.00				
Stability	0.01	1.00			
Consistency	-0.50	-0.33	1.00		
Power	0.31	-0.21	-0.03	1.00	
Authority	-0.04	0.04	-0.34	-0.18	1.00

The correlations among the policy attributes are surprisingly variable across the three subjects. It is not clear what the explanation for these differences across subjects is. However, only the index for POWER is associated with a particular state across academic subjects—the other four indices are calculated for each standards document or standards and assessment pair. Thus, it appears to be the case that the relationships among the attributes differ by subject. Alternatively, the differences in correlations across subjects

could simply be due to differences in the states included in each sample. None of the correlations are sufficiently high as to merit concern about collinearity.

Chapter 4

Results

The first section of this chapter describes changes in alignment to standards and assessments in the three content areas between 2003 and 2009. The second section identifies changes made in the content of instruction, in terms of hours of instruction, topics taught, and cognitive demand levels emphasized. The third section relates alignment to state policy attributes and features of local content.

Changes in Alignment, 2003-2009

Because of the large changes in sample composition in terms of states and grades over time, it was not possible to merely compare means in the alignment index across time. Those means would be misleading, depending on which particular states and grades were represented in the sample in each year. Rather, the question is whether alignment increased within individual states and grades over time. Thus, two primary approaches were used to examine changes in alignment over time. The first was to use the full set of data and regress alignment to standards or assessments on year, accounting for sample composition through state and grade fixed effects. The second was to narrow the focus to individual states that were well-represented across years and conduct additional fixed-effects regressions with grade fixed effects. These results shed more light on changes in instructional alignment at the individual state level and illustrate whether the national trends appear to be operating similarly across states. The results for both sets of analyses are presented in the following section, separated by content area.

Mathematics. The first analysis is a fixed-effects regression of alignment on year of administration. First, two sets of models were run for alignment to standards, and two sets of models were run for alignment to assessments. One of each set had fixed effects for states and grades (e.g., 4th grade, 5th grade, Ohio, Pennsylvania), while the other had fixed effects for the state-grade interaction (e.g., fixed effects for 4th grade in Ohio, 5th grade in Ohio, 7th grade in Pennsylvania). The sets of models produced nearly identical estimates for the coefficient on year, so the more conceptually simple model with fixed effects for state and fixed effects for grade was chosen. Next, three separate models were run for standards and three for assessments, with one model for each grade band of K-2, 3-8, and 9-12. The results of these grade-band models are presented here.

Table 11
Fixed Effects Regressions of Alignment to Mathematics Standards and Assessments

Variable	Grades	B	SE(B)	t	Sig. (<i>p</i>)	Model R ²	n
Mathematics Standards							
Year of Administration	K-2	0.0043**	0.0015	2.90	0.004	0.47	1339
Year of Administration	3-8	0.0037***	0.0005	7.71	<.001	0.63	7935
Year of Administration	9-12	-0.0028*	0.0014	-2.01	0.044	0.44	1339
Mathematics Assessments							
Year of Administration	3-8	0.0027***	0.0005	4.99	<.001	0.20	4239
Year of Administration	9-12	0.0032*	0.0016	2.04	0.041	0.15	664

Note. * *p* < .05, ** *p* < .01, *** *p* < .001

Table 11 shows the results of the fixed effects regressions of alignment to standards and assessments on year of administration in mathematics. The model would not run for grades K-2 assessments because of small sample sizes. For standards in grades K-2 and 3-8 and assessments in grades 3-8 and 9-12, the coefficient is positive and

significantly different from zero. The estimated six-year change in the alignment index for these four increases ranges from .016 for grades 3-8 assessments to .026 for grades K-2 standards—increases in content overlap of a few percent. In contrast, for alignment to standards in grades 9-12, the coefficient is negative and significant, with a six-year estimated decrease of 1.7% in content overlap. The coefficients indicate the expected annual change in alignment within a state and grade, averaged across states and grades.

To get a better idea of the magnitude of the coefficients, they can be standardized. The estimated one-year increase in alignment to standards is approximately .06 standard deviations for grades K-2 and .04 standard deviations for grades 3-8, corresponding to an expected increase in alignment across the six-year period of .37 standard deviations for grades K-2 and .25 standard deviations for grades 3-8. The estimated one-year decrease in alignment to standards for grades 9-12 is .03 standard deviations, corresponding to a six-year decrease in alignment of .21 standard deviations. For assessments, the estimated one-year increase in alignment is .05 standard deviations for grades 3-8 and .06 standard deviations for grades 9-12, for six-year changes of .30 and .34 standard deviations, respectively. The results clearly suggest that, within individual states and grades, the alignment to standards and assessments increased over the study period by roughly one-third of a standard deviation in mathematics. The only exception was for high school standards, where there was a significant decrease in alignment.

A second way to analyze the data is to look within individual states and compare average alignment indices across years. These comparisons were again made by conducting fixed-effects regression within individual states, this time with fixed effects

for grades only. States were included in these analyses if there were 10 or more teachers in each of two or more years. The results of these analyses are presented in Table 12. States were analyzed by grade band for alignment to standards and assessments. Because of the large number of comparisons, only the states with significant changes in alignment are presented. There were a total of 6 states analyzed for grades K-2 standards, 16 for grades 3-8 standards, 10 for grades 3-8 assessments, 8 for grades 9-12 standards, and 4 for grades 9-12 assessments.

Of the 44 total regressions, 15 showed significant changes in alignment over time. In general, the state-level results in Table 12 support the aggregate results in Table 11. That is, for grades where there was a significant increase in alignment at the national level, the majority of state-level changes in alignment were also increases. This is true for K-2 and 3-8 standards and 3-8 and 9-12 assessments. In the one case where the national analysis found a decrease in alignment—grades 9-12 standards—two of the three significant state-level changes were negative. However, the significant changes in alignment found in the national analysis did not occur uniformly across states. For instance, there were significant decreases in alignment to standards for grades 3-8 teachers in California, Idaho, and Oklahoma, while the national trend was for increased alignment. Certain states had significant changes in alignment in multiple grades. For instance, Indiana had significant increases in alignment to grades 3-8 standards and grades 9-12 standards and assessments. Ohio had significant increases in alignment to standards and assessments in grades 3-8. Oregon had significant increases in alignment to

grades K-2 standards and 3-8 standards and assessments, but a significant decrease in alignment to grades 9-12 standards.

Table 12
Fixed Effects Estimates of Changes in Alignment Within Individual States, Mathematics

State	B	SE(B)	t	Sig. (<i>p</i>)
K-2 Standards				
Oregon	0.010	0.003	3.28	0.001
3-8 Standards				
California	-0.032	0.012	-2.64	0.011
Idaho	-0.015	0.007	-2.21	0.028
Illinois	0.007	0.001	5.83	<.001
Indiana	0.008	0.002	3.07	0.002
Maine	0.015	0.007	2.10	0.039
Ohio	0.008	0.001	11.10	<.001
Oklahoma	-0.002	0.001	-2.38	0.017
Oregon	0.005	0.002	2.59	0.010
3-8 Assessments				
Ohio	0.008	0.001	8.63	<.001
Oregon	0.003	0.002	1.86	0.063
9-12 Standards				
Delaware	-0.012	0.006	-2.10	0.037
Indiana	0.025	0.011	2.20	0.031
Oregon	-0.012	0.005	-2.36	0.022
9-12 Assessments				
Indiana	0.016	0.007	2.10	0.035

Note. Only states with significant changes presented.

In terms of magnitude, the significant increases in alignment were between .003 and .025 points per year on the alignment index. Because some of these states only have

data for two years, it is not appropriate to make extrapolations to the six-year study period for all of them. However, for those states with significant increases in alignment, the typical per-year change in alignment is roughly .08 to .10 on the alignment scale. The magnitudes are similar, if not slightly larger, for those states with significant declines in alignment. For Ohio, the only state with samples greater than 10 in all years, the expected six year change in alignment to grades 3-8 standards and assessments is approximately .064 on the alignment scale, nearly a full standard deviation. Overall, the results of the state-level analyses are supportive of the findings of the national analyses. The state-level analyses also point to Indiana, Ohio, and, to a lesser extent, Oregon as having somewhat consistent increases in alignment.

ELAR. Results of the fixed-effects models regressing alignment to ELAR standards and assessments on year of administration are presented in Table 13. In contrast to mathematics, where the results generally indicated significant increases in alignment, these models find no change in instructional alignment to standards or assessments in grades 3-8 and no change in alignment to standards in grades 9-12. There is a significant decrease in alignment to standards for grades K-2. The standardized coefficient is .04, indicating a six-year estimated decrease of approximately .27 standard deviations in alignment to standards. There are no assessments coded for grades K-2, and there are not enough teachers to estimate changes in alignment to assessments for grades 9-12. Looking across all the available data, these results suggest that there have largely been no change in teachers' ELAR instructional alignment to standards and assessments across

the study period, with some evidence of a decrease in alignment to standards at the lower grades.

Table 13
Fixed Effects Regressions of Alignment to ELAR Standards and Assessments

Variable	Grades	B	SE(B)	t	Sig. (<i>p</i>)	Model R ²	n
ELAR Standards							
Year of Administration	K-2	-0.0028***	0.0005	-5.14	<.001	0.25	5750
Year of Administration	3-8	-0.0004	0.0010	-0.40	0.686	0.32	5618
Year of Administration	9-12	0.0001	0.0033	0.02	0.982	0.29	439
ELAR Assessments							
Year of Administration	3-8	-0.0012	0.0008	-1.56	0.118	0.22	4185

Note. * *p* < .05, ** *p* < .01, *** *p* < .001

Looking at individual states, there are few significant changes over time. A total of 5 states were analyzed for grades K-2 standards, 11 for grades 3-8 standards, 6 for grades 3-8 assessments, and 2 for grades 9-12 standards. Of these 24 regressions, only 6 showed significant changes in alignment over time. Table 14 summarizes these significant changes. In general, the changes are mixed, with four states showing significant decreases in alignment for a particular grade band and document and two states showing significant increases. Indiana at grades K-2 shows a significant decrease in alignment to standards. Ohio, by far the largest state in the sample, saw small decreases of .002 to .003 units per year on the alignment index. The results suggest no particular trend in either direction in terms of alignment to either standards or assessments at the individual state level, supporting the finding of no major changes in alignment in ELAR nationally.

Table 14
Fixed Effects Estimates of Changes in Alignment Within Individual States, ELAR

State	B	SE(B)	t	Sig. (<i>p</i>)
K-2 Standards				
Indiana	-0.033	0.009	-3.74	<.001
Michigan	0.024	0.007	3.44	0.001
Ohio	-0.003	0.000	-7.72	<.001
3-8 Standards				
Illinois	0.008	0.004	2.00	0.046
Ohio	-0.002	0.001	-2.55	0.011
3-8 Assessments				
New York	-0.011	0.003	-3.98	<.001

Note. Only states with significant changes presented.

Science. Table 15 shows the results of the fixed effects models regressing alignment to standards or assessments on year of administration. There were no observations for K-2 assessments. For alignment to standards, the results suggest an increasing alignment over time at grades 3-8 and 9-12, and no significant change at grades K-2. The magnitude of the year-on-year changes is approximately .07 standard deviations for grades 3-8 and .05 standard deviations for grades 9-12, for a six-year estimated change of .43 standard deviations and .31 standard deviations, respectively. These changes are somewhat larger than the changes observed in mathematics. For alignment to assessments, there is a significant decrease in alignment at grades 3-8 and no significant change at grades 9-12. The magnitude of the decrease in alignment at grades 3-8 is .11 standard deviations per year, or .64 standard deviations for the study

period. Across all the models, the results suggest increasing alignment to standards at assessed grades and decreasing or stable alignment to assessments.

Table 15
Fixed Effects Regressions of Alignment to Science Standards and Assessments

Variable	Grades	B	SE(B)	t	Sig. (p)	Model R ²	n
Science Standards							
Year of Administration	K-2	-0.0041	0.0026	-1.61	0.108	0.20	353
Year of Administration	3-8	0.0042***	0.0006	6.87	<.001	0.18	4277
Year of Administration	9-12	0.0040*	0.0016	2.51	0.012	0.48	693
Science Assessments							
Year of Administration	3-8	-0.0052***	0.0008	-6.41	<.001	0.29	1480
Year of Administration	9-12	0.0008	0.0019	0.42	0.673	0.73	406

Note. * $p < .05$, ** $p < .01$, *** $p < .001$

Again, it is also useful to examine changes in alignment in individual states, to consider the extent to which they support the findings of the fixed effects analysis. The results are presented in Table 16. A total of 2 states were analyzed for grades K-2 standards, 16 for grades 3-8 standards, 8 for grades 3-8 assessments, and 4 each for grades 9-12 standards and assessments. Of these 34 regressions, 15 found significant changes in alignment over time. Largely, the results support the findings of the national analysis. The national analysis found significant increases in alignment to grades 3-8 and 9-12 standards and significant decreases in alignment to grades 3-8 assessments. Here, of the eight states with significant changes in alignment to grades 3-8 standards, seven had increases and one had a decrease. Three of the four states with significant changes in alignment to grades 3-8 assessments saw decreases, which was also consistent with the national findings. For grades 9-12 standards, just one state saw a significant change, an

increase in alignment in Ohio. There were mixed results for grades 9-12 assessments and one significant decrease in alignment to grades K-2 standards (Oklahoma).

State	B	SE(B)	t	Sig. (p)
K-2 Standards				
Oklahoma	-0.007	0.003	-2.15	0.033
3-8 Standards				
California	0.016	0.007	2.12	0.035
Idaho	0.039	0.008	5.09	<.001
Illinois	0.005	0.001	5.19	<.001
Indiana	0.013	0.005	2.56	0.012
Maine	-0.025	0.011	-2.45	0.017
Missouri	0.014	0.005	2.51	0.013
Ohio	0.011	0.001	7.94	<.001
3-8 Assessments				
Illinois	-0.006	0.001	-5.61	<.001
Maine	-0.019	0.009	-2.03	0.047
Oklahoma	0.004	0.002	2.11	0.036
Wisconsin	-0.018	0.002	-9.35	<.001
9-12 Standards				
Ohio	0.009	0.003	2.77	0.006
9-12 Assessments				
Ohio	0.018	0.005	3.69	0.001
Wisconsin	-0.011	0.004	-3.13	0.002

Note. Only states with significant changes presented.

The changes in alignment to science standards and assessments were somewhat larger than the changes in the other subjects. This was particularly true for grades 3-8 standards, where six of the states with significant increases in alignment had estimated

changes of .01 units per year in the alignment index. For Ohio, which has teacher samples greater than 10 in all years, the estimated change across these six years is .066 units, more than a full standard deviation. Overall, the results of the science state-by-state analyses are supportive of the national analyses, indicating increases in alignment to grades 3-8 standards and decreases in alignment to grades 3-8 assessments.

Comparison with Reduced SEC Taxonomy. In order to examine the sensitivity of these results to the changes in the SEC taxonomy that occurred between 2006 and 2007, the same analyses were re-run using the dataset where alignment was calculated on the re-centered data. First, all topics that did not appear in both the pre-2007 and current versions of the SEC were deleted from the data for teachers, standards, and assessments. Then, the remaining cell proportions were normalized so they would sum to one, and the alignment indices were calculated based on the re-centered data.

Of teachers' total instruction, on average, 8.1% of K-8 mathematics teachers' instruction was removed in the re-centering. That is, 91.9% of the content K-8 mathematics teachers reported covering appeared on both the pre-2007 and post-2007 surveys. The values for the other subjects and grades were 19.2% removed for mathematics grades 9-12, 8.7% removed for ELAR, 10.8% removed for science grades K-8, and 16.9% removed for science grades 9-12. For standards and assessments, the proportion of the content removed was 9.6% for mathematics standards, 8.7% for mathematics assessments, 4.5% for ELAR standards, 3.5% for ELAR assessments, 18.2% for science standards, and 18.9% for science assessments. Clearly, the amount of

content removed in this procedure was non-trivial on average, for both teachers and standards/assessments.

Table 17
Comparison of Fixed Effects Regressions of Alignment to Standards and Assessments Using Main Data and Recentered Data

Grades	Recentered Data			Main Data		
	B	SE(B)	t	B	SE(B)	t
Mathematics Standards						
K-2	0.0063***	0.0015	4.11	0.0043**	0.0015	2.90
3-8	0.0051***	0.0005	9.47	0.0037***	0.0005	7.71
9-12	0.0018	0.0015	1.19	-0.0028*	0.0014	-2.01
Mathematics Assessments						
3-8	0.0045***	0.0006	7.43	0.0027***	0.0005	4.99
9-12	0.0035*	0.0017	2.04	0.0032*	0.0016	2.04
ELAR Standards						
K-2	0.0030***	0.0006	5.05	-0.0028***	0.0005	-5.14
3-8	0.0038***	0.0010	3.66	-0.0004	0.0010	-0.40
9-12	0.0055	0.0034	1.61	0.0001	0.0033	0.02
ELAR Assessments						
3-8	0.0018*	0.0008	2.29	-0.0012	0.0008	-1.56
Science Standards						
K-2	-0.0077**	0.0027	-2.89	-0.0041	0.0026	-1.61
3-8	0.0021**	0.0007	3.15	0.0042***	0.0006	6.87
9-12	-0.0014	0.0017	-0.80	0.0040*	0.0016	2.51
Science Assessments						
3-8	-0.0037***	0.0008	-4.46	-0.0052***	0.0008	-6.41
9-12	-0.0071**	0.0021	-3.29	0.0008	0.0019	0.42

Note. * $p < .05$, ** $p < .01$, *** $p < .001$

The results of the fixed effects models were similar based on the two analyses for mathematics, as seen in Table 17. As in the main analysis, in the re-centered analysis, there were significant increases in alignment to mathematics standards in grades K-2 and 3-8 and assessments in grades 3-8 and 9-12. The magnitudes of these coefficients are .09 standard deviations per year for K-2 standards, $.05\sigma$ for 3-8 standards, $.08\sigma$ for 3-8 assessments, and $.06\sigma$ for 9-12 assessments. Over the study period, these represent total increases in alignment of .31 to .52 standard deviations, somewhat larger than the effects that were found in the main analysis. The significant decrease in alignment to standards at the higher grades from the main analysis was not duplicated—instead, there was no significant change in alignment to standards for mathematics grades 9-12.

The results were not as consistent in science and ELAR. In ELAR, when the re-centered data were used, there were significant increases in alignment to standards at grades K-2 and grades 3-8 and no change at grades 9-12. For assessments, there was a significant increase in alignment at grades 3-8 and no change at grades 9-12. In science, there was a significant decrease in alignment to standards at grades K-2, a significant increase at grades 3-8, and no change at grades 9-12. For assessments, there were significant decreases in alignment at grades 3-8 and 9-12. The difference in the results is suggestive that the changes to the framework, if not properly accounted for, may bias the alignment indices. However, the direction of that potential bias is unclear, as the results using the re-centered data varied both above and below the results from the main analyses.

Comparison across Content Areas. Overall, the results of the two sets of fixed effects regressions are fairly consistent in mathematics, suggesting mainly positive changes in alignment to standards and assessments. In ELAR, the main analysis suggested essentially no significant change in alignment across the grades and models. However, the re-centered analysis suggested increases in alignment at most grades. As for science, the two models suggested significant increases in alignment to standards at grades 3-8 and possibly 9-12, but no change or decreases for assessments and at earlier grades.

Certain states stand out as demonstrating significant changes in a number of subjects and grade levels. For instance, Ohio teachers showed increased alignment of instruction with grades 3-8 mathematics standards and assessments and grades 3-8 standards and 9-12 standards and assessments in science, though they exhibited slightly decreased alignment with grades 3-8 ELAR standards and assessments. Illinois teachers in grades 3-8 showed increasing alignment of instruction with standards in all three subjects, though there was a small decrease in alignment with grades 3-8 assessments in science. Indiana teachers showed increasing alignment to grades 3-8 standards in mathematics and science, as well as grades 9-12 standards and assessments in mathematics, though there was a significant decrease in alignment to grades 3-8 ELAR standards. It is important to mention again that the samples of teachers in these states changed across years, and that the analyses of state-level changes controlled only for the composition of the samples in terms of grade.

Given that these three states are among the states most highly represented in the database, it seems likely that the significance level of changes in alignment is related to the sample sizes of teachers in the particular grades/subjects. That is, there would likely be more states showing up as having significant changes in alignment if the samples were bigger across years. However, the state-level and national-level results clearly suggest that teachers are changing their instruction to align with standards, at least in mathematics and grades 3-8 in science, and the degree of change seems to vary across states.

Changes in the Content of Instruction

The previous section identified changes in teachers' instructional alignment to standards and assessments. While the results suggested increased alignment in certain subjects and grades, the results were not useful in describing the nature of changes in teachers' instruction. The literature offers a number of suggestions about the changes teachers have made in their instruction under standards-based reform. These include shifting content from non-tested to tested content areas (Clarke et al., 2003; Hamilton & Berends, 2006; Pedulla et al., 2003) and increasing coverage of focal topic strands on standards and assessments (Koretz, Mitchell et al., 1996; Stecher & Barron, 2001; Stecher & Chun, 2001). The detailed content data used here allow for analysis of the nature of changes in teachers' instruction and the extent to which those changes match with the broad content messages of the standards and assessments.

Instructional Time. One consistent finding from the literature is that teachers report a shifting of instructional time from non-tested to tested subjects. While the proportion of teachers reporting such a shift was near 100% in early studies of standards-

based reform, the proportion was smaller in more recent studies, perhaps suggesting a stabilization of content shifts over time. The data in the SEC Part A allow analysis of the extent to which hours of instruction have changed across time in each content area. Specifically, teachers were asked the number of hours per week they spent teaching each content area. As with the analysis of alignment, these data are analyzed using fixed-effects regression, with fixed-effects for state and grade. Again, teachers are first grouped by grade band (K-2, 3-8, and 9-12).

Results. The regression results, displayed in Table 18, suggest that previously-reported changes in instructional time have diminished or stopped by the years 2003 to 2009. There is a small increase in instructional hours reported in grades 3-8 mathematics. The coefficient of .028 indicates that the amount of instruction devoted to mathematics within a particular state and grade increased roughly .17 hours per week over the six-year study period, a change of just ten minutes in the amount of time devoted to mathematics (from a starting point of approximately 5 hours per week). There were no significant changes in earlier or later grades in mathematics or in any grades in ELAR. In science, there was a moderate decrease in instructional time in grades 9-12 of approximately .06 hours per study year, or just over 20 minutes per week for the six-year study period. There was also a moderately large negative coefficient for grades K-2 in science, but this was nonsignificant, perhaps due to the smaller sample at that grade. Overall, these results suggest that any lingering changes in the allocation of instructional time over the years 2003-2009 were small or nonexistent. Perhaps if other content areas, such as social studies or the arts, had been studied, different results would have been found.

Table 18
Fixed Effects Regressions of Hours of Instruction on Year of Administration

Grades	B	SE(B)	t	Sig. (p)	Model R ²	n
Mathematics						
K-2	0.014	0.040	0.36	0.721	0.12	1807
3-8	0.028**	0.011	2.62	0.009	0.11	10535
9-12	-0.011	0.020	-0.56	0.577	0.15	4050
ELAR						
K-2	-0.004	0.021	-0.18	0.855	0.02	6569
3-8	0.006	0.024	0.25	0.806	0.13	6806
9-12	-0.007	0.047	-0.16	0.872	0.13	1614
Science						
K-2	-0.041	0.041	-1.01	0.312	0.09	826
3-8	0.011	0.012	0.97	0.332	0.19	5819
9-12	-0.061**	0.021	-2.93	0.003	0.10	2775

Note. * $p < .05$, ** $p < .01$, *** $p < .001$

Levels of Cognitive Demand. A question emerging from the literature is whether teachers are “dumbing down” their instruction by teaching rote or procedural thinking instead of conceptual thinking or problem solving, and to what extent such changes are reflective of the incentives provided by the standards and accountability system. Because teachers report their instruction on the SEC at the intersection of topics and levels of cognitive demand, these data can be used to analyze changes in the allocation of instructional time to the various levels of cognitive demand. The definitions of the five levels of cognitive demand for each content area are presented in Appendix B. The levels go from B to F, with level B corresponding to memorization and level C corresponding to performing procedures. The three higher levels differ across subjects, but generally include explanation/communication, analysis, and application/connections.

For each teacher, a marginal proportion was created by adding across all cells representing each level of cognitive demand. These five marginal proportions were the dependent variables in fixed effects regressions, again with fixed effects for state and grade, and with an independent variable of year of administration. The results indicate the extent to which the average proportion of instructional time devoted to each of the five levels of cognitive demand has changed over the study period. The analysis was done on both the main data and the re-centered data. The results based on the main data are presented here in the text. The results were nearly identical for the re-centered data, as seen in Appendix Table C.1, with similar coefficients in terms of magnitude and only one change in significance level.

Results. The results of the analyses for mathematics, ELAR, and science using the main data are presented in Tables 19, 20, and 21, respectively. The dependent variable was a proportion that was re-scaled to 0-to-100. Thus, the coefficients in the tables indicate the estimated percent change in each marginal proportion per year. For instance, a coefficient of 1.0 would indicate that the proportion of instruction at that level of cognitive demand increased 1% per year during the study period.

The results for mathematics in Table 19 support the claim that instruction is being shifted from higher levels of cognitive demand toward lower levels. In particular, there are significant increases in the proportion of instruction at level C (perform procedures) at all three grade levels. There are also significant decreases in coverage of the top two levels of cognitive demand (conjecture, generalize, prove; and solve non-routine problems, make connections) for grades K-2 and 3-8. The one exception to this pattern is

a significant decrease in coverage of cognitive demand level B (memorize) in grades 9-12.

Table 19
Changes in Emphasis on Five Levels of Cognitive Demand in Mathematics Instruction

Grade	n	Level B		Level C		Level D		Level E		Level F	
		B	SE(B)	B	SE(B)	B	SE(B)	B	SE(B)	B	SE(B)
K-2	1808	0.22	0.26	0.97***	0.23	0.37†	0.21	-0.89***	0.17	-0.66***	0.19
3-8	10536	-0.05	0.06	0.57***	0.06	-0.05	0.05	-0.33***	0.05	-0.14**	0.05
9-12	2411	-0.47**	0.17	0.49*	0.20	0.16	0.18	-0.21	0.13	0.04	0.17

Note. Values are regression coefficients for fixed-effects regression of marginal proportion on year of administration. † $p < .10$, * $p < .05$, ** $p < .01$, *** $p < .001$

The magnitude of the shifts in cognitive demand coverage is largest in grades K-2. The coefficient of .97 for level C indicates a nearly 1% increase in coverage of procedures per year. This corresponds to a six-year total estimated change of 5.8%, quite close to the actual change based on the mean marginals of 5.9%, from 20.2% to 26.1%. For a 180-day school year, this amounts to a shift of more than 10 days of instruction toward procedural thinking. The large negative coefficients on levels E and F for grade K-2 similarly indicate significant shifts of 9.6 and 7.1 days' instruction, respectively, over the six-year study period, while the actual changes were the equivalent of 11.9 and 7.7 days' instruction, respectively.

The coefficients are somewhat smaller for the other grades, but they still represent meaningful shifts, especially when aggregated across six years. Another way to describe the magnitude of the shifts is in terms of standard deviations. For instance, the standardized coefficient for grades K-2 level C is .10. This is a moderately large standardized coefficient for a one-year change, and it indicates that the cumulative

increase in focus on performing procedures represents across the six years is three-fifths of a standard deviation.

Table 20
Changes in Emphasis on Five Levels of Cognitive Demand in ELAR Instruction

Grade	n	Level B		Level C		Level D		Level E		Level F	
		B	SE(B)	B	SE(B)	B	SE(B)	B	SE(B)	B	SE(B)
K-2	6571	0.48***	0.11	0.31***	0.09	0.78***	0.07	-0.92***	0.07	-0.66***	0.07
3-8	6807	-0.39**	0.11	0.10	0.10	1.05***	0.09	-0.37***	0.08	-0.40***	0.08
9-12	1614	-1.44***	0.28	-0.51*	0.21	1.47***	0.26	0.41*	0.19	0.08	0.22

Note. Values are regression coefficients for fixed-effects regression of marginal proportion on year of administration. † $p < .10$, * $p < .05$, ** $p < .01$, *** $p < .001$

The shifts in cognitive demand levels for ELAR are presented in Table 20. Again, the results largely support the claim that teachers have redistributed content away from higher levels of cognitive demand and toward lower levels, at least at lower grades. However, the pattern is not as consistent as in mathematics. There are significant increases in coverage of cognitive demand levels B, C, and D for grades K-2, as well as level D for grades 3-8 and 9-12. There are also significant decreases in higher-level cognitive demands E and F for grades K-2 and 3-8. However, there are a few changes that contradict the notion of a strict move from higher-level to lower-level teaching, both in grades 3-8 and 9-12. In both sets of grades there is a significant decrease in coverage of memorization. In grades 9-12, there is also a significant decrease in coverage of procedures and a significant increase in coverage of cognitive demand level E.

The magnitudes of the changes are larger for ELAR than for mathematics. Most notable are the increases in focus on cognitive demand level D, which represents generate/create/demonstrate, at all three grade levels. The largest is the coefficient for

cognitive demand level D for grades 9-12, which indicates a shift of roughly 2.5 days of instruction per year toward that level. Standardized, the magnitude of the year-to-year increases in coverage of cognitive demand level D are .10 standard deviations for grades K-2, .12 standard deviations for grades 3-8, and .15 standard deviations for grades 9-12. Over the five-year study period (there are no data from 2003 for ELAR), these effects are .49, .61, and .73 standard deviations, respectively. The actual changes from 2004 to 2009 based on mean marginal proportions are .39 standard deviations for grades K-2, .31 standard deviations for grades 3-8, and .39 standard deviations for grades 9-12.

Other particularly large changes are the decreases in coverage of cognitive demand levels E and F for grades K-2. These changes are .13 and .08 standard deviations per year, or .63 and .42 standard deviations for the study period, respectively. In terms of the actual means, the decreases were .53 and .63 standard deviations, respectively. Clearly, English teachers are making nontrivial changes in the allocation of their instructional time across levels of cognitive demand, and in elementary and middle grades these changes are primarily favoring lower-level cognitive demand at the expense of higher levels.

The results for science are displayed in Table 21, and they again largely support the contention that instruction is shifting to lower levels of cognitive demand. While there are no significant changes in cognitive demand levels at grades K-2, there are significant or marginally significant decreases in focus on the top two levels of cognitive demand at grades 3-8 and 9-12. Additionally, there are significant increases in coverage of levels C

and D for grades 3-8 and level B for grades 9-12. Overall, there are consistent shifts in cognitive demand from higher to lower levels at grades 3 and higher.

Table 21
Changes in Emphasis on Five Levels of Cognitive Demand in Science Instruction

Grade	n	Level B		Level C		Level D		Level E		Level F	
		B	SE(B)	B	SE(B)	B	SE(B)	B	SE(B)	B	SE(B)
K-2	826	0.08	0.44	-0.29	0.34	-0.12	0.31	0.09	0.25	0.24	0.28
3-8	5952	-0.03	0.11	0.15*	0.06	0.19**	0.06	-0.09†	0.05	-0.22**	0.06
9-12	2775	0.35*	0.16	0.03	0.09	0.05	0.09	-0.16*	0.08	-0.27*	0.12

Note. Values are regression coefficients for fixed-effects regression of marginal proportion on year of administration. † $p < .10$, * $p < .05$, ** $p < .01$, *** $p < .001$

One difference between the results for science and the results for the other two content areas is that the magnitude of the shifts is smaller in science. The largest shift is an increase of .35% per year in memorization at grades 9-12, a change over the six-year period of 2.1%, or approximately 3.8 days' instruction in a 180 day school year. This amounts to a standardized change of .03 standard deviations per year, or .19 standard deviations over the six-year period. The other significant changes have smaller effect sizes. These results suggest that, while there are changes taking place in the allocation of science teachers' instruction across levels of cognitive demand, such that more rote and procedural teaching is favored, these changes are relatively small in magnitude, especially when compared with the changes in other content areas. One possible explanation for the smaller magnitudes is that science is not tested in all grades.

Comparison with Standards and Assessments. One important question about teachers' changes in the cognitive demand of their instruction is whether those changes

reflect the messages they receive through standards and assessments about the content that is to be taught. To address this question, within each grade band, the marginals for all available standards documents were calculated and averaged. The same was done for assessments. Then, the averages of the marginals for teachers in 2002-03 and 2003-04 for mathematics and science and 2003-04 for ELAR were calculated⁵. Finally, the three sets of marginals (first year instruction, content of standards, content of assessments) were compared against the significant coefficients from Tables 19, 20, and 21 to determine whether teachers' changes in their instruction moved their instruction in the direction of increased agreement with the standards and assessments or decreased agreement with the standards and assessments. The results of these analyses are presented in Table 22 for mathematics, Table 23 for ELAR, and Table 24 for science.

In mathematics, the results in Table 22 suggest that teachers' decisions about the cognitive demand levels to emphasize in their instruction were highly aligned with the messages provided by the standards and assessment documents. Across all three grade bands and for both standards and assessments, all but one of the significant changes in instruction was in the direction of increased agreement with the standards and assessments. The one exception was an increase in the proportion of time K-2 teachers spend focusing on cognitive demand level D—this change brought instruction closer to alignment with standards, but farther from alignment with assessments. Also, at each grade level, the largest change in magnitude (from Table 18) was associated with level C,

⁵ The first two years of data were chosen for mathematics and science because certain combinations of years and grade bands, such as grades K-2 mathematics in 2003, had small samples (less than 10 teachers).

the level in which the 2003/4 average instruction and the targets (the content of standards and assessments) were farthest apart.

Table 22
Comparison of Change in Cognitive Demand for Teachers' Mathematics Instruction with Cognitive Demand Emphasis of Standards and Assessments, 2003-2009

Cognitive demand level	2003/4 proportion	Coefficient sign	Proportion in standards	Proportion in assessments
Grades K-2				
C	0.257	+	0.408*	0.634*
D	0.233	+	0.330*	0.107
E	0.148	-	0.073*	0.053*
F	0.132	-	0.022*	0.007*
Grades 3-8				
C	0.251	+	0.508*	0.599*
E	0.149	-	0.088*	0.066*
F	0.137	-	0.036*	0.015*
Grades 9-12				
B	0.236	-	0.064*	0.095*
C	0.275	+	0.514*	0.741*

Note. + indicates a positive coefficient on year of administration in fixed effects regression equation. - indicates a negative coefficient. * indicates change in instruction that is in the direction of increased alignment with standards or assessments.

The results for ELAR, presented in Table 23, are again largely supportive of the contention that teachers have moved the cognitive demand emphases of their instruction toward greater alignment with standards and assessments. At grades K-2, four of the five changes are in the direction of increased agreement with content standards (no K-2 assessments are in the database). At grades 9-12, all of the changes are in the direction of increased agreement with standards, and all but one for assessments. The results are less strong for grades 3-8, where three of four changes are in the direction of increased

agreement with standards, but only one of four is in the direction of increased agreement with assessments. Still, across the many cognitive demand levels where there were significant changes in instruction, only one, level E for grades 3-8, had a change that was in the direction of decreased agreement with both standards and assessments. Notably, perhaps the strongest results for ELAR were in grades 9-12, where the overall trend of increased emphasis on lower-level cognitive demand was not true. That is, whether the standards and assessments favored lower-level thinking or higher-level thinking, teachers seem to have responded in the direction of increased agreement.

Table 23
Comparison of Change in Cognitive Demand for Teachers' ELAR Instruction with Cognitive Demand Emphasis of Standards and Assessments, 2004-2009

Cognitive demand level	2004 proportion	Coefficient sign	Proportion in standards	Proportion in assessments
Grades K-2				
B	0.246	+	0.219	n/a
C	0.238	+	0.318*	n/a
D	0.176	+	0.322*	n/a
E	0.178	-	0.100*	n/a
F	0.163	-	0.041*	n/a
Grades 3-8				
B	0.225	-	0.110*	0.313
D	0.189	+	0.365*	0.143
E	0.183	-	0.192	0.281
F	0.176	-	0.083*	0.035*
Grades 9-12				
B	0.216	-	0.040*	0.180*
C	0.208	-	0.130*	0.272
D	0.187	+	0.404*	0.233*
E	0.204	+	0.279*	0.278*

Note. + indicates a positive coefficient on year of administration in fixed effects regression equation. - indicates a negative coefficient. * indicates change in instruction that is in the direction of increased alignment with standards or assessments.

Finally, the results presented in Table 24 compare teachers' instructional changes in science in terms of cognitive demand emphasis to the content of standards and assessments. For both grade bands, a majority of significant changes were in the direction of increased agreement. For grades 3-8, all four significant changes led to increased agreement with standards, but only one of the changes led to increased agreement with assessments. For grades 9-12, the opposite was true—all three significant changes were in the direction of increased agreement with assessments, but only one of the changes led to increased agreement with standards.

Cognitive demand level	2003/4 proportion	Coefficient sign	Proportion in standards	Proportion in assessments
Grades 3-8				
C	0.196	+	0.233*	0.098
D	0.211	+	0.335*	0.177
E	0.184	-	0.171*	0.202
F	0.191	-	0.038*	0.047*
Grades 9-12				
B	0.209	+	0.189	0.443*
E	0.191	-	0.198	0.171*
F	0.204	-	0.052*	0.073*

Note. + indicates a positive coefficient on year of administration in fixed effects regression equation. - indicates a negative coefficient. * indicates change in instruction that is in the direction of increased alignment with standards or assessments.

Certainly, across the three subjects, there is strong evidence that teachers have changed their instruction, largely in the direction of increasing emphasis on memorization

or procedural thinking and away from higher levels of cognitive demand. In many cases, particularly in mathematics and ELAR, the magnitude of these changes is large, especially when aggregated across the study period. Furthermore, in the majority of cases, the changes in teachers' cognitive demand emphases seem to have brought instruction toward increased agreement with standards and assessments. Of the 24 cognitive demand levels analyzed here, across the three subjects (excluding ELAR K-2), 14 of the changes saw increased agreement with both standards and assessments, 9 saw increased agreement with one of the two, and just 1 saw increased agreement with neither.

Topics. A similar approach as the one used to analyze levels of cognitive demand was used to analyze coarse-grained topics in the SEC framework. However, the changes in the lists of topics between 2006 and 2007 necessitated a slightly different analytical plan. As was done with cognitive demands, the data were first separated into grade bands. Next, fixed-effects regressions, with fixed effects for grade and state, were run using the re-centered data. For any coarse-grained topic identified as having a significant change in coverage using the re-centered data, the main data were then used, again with fixed-effects regression. Here, however, the main data were first split into two halves—one for 2003 or 2004 to 2006, and one for 2007 to 2009. If a topic had a significant change in the same direction in either group of years for the main data as it did for the re-centered data, it was identified as having a significant change overall.

Results. The results for mathematics, shown in Table 25, indicate that no coarse-grained topics at grades K-2 or 9-12 had significant changes using the conservative

criteria chosen here. However, nine of the 16 topics had significant changes over time at grades 3-8. Three coarse-grained topics had negative shifts—number sense/properties/relationships, operations, and measurement. Of these, number sense and operations had larger shifts, the equivalent of .6% to .7% of total instruction per year. Over the six-year span, the shifts were 3.6% to 4.0% using the re-centered data, quite close to the actual shifts based on marginal means of 4.0% each. In contrast, six coarse-grained topics had positive shifts. The largest shifts were increases in coverage of instructional technology and basic algebra. But there were also significant increases in coverage of data displays, statistics, and probability, each on the order of .2% of total instruction per year, or roughly 1.2% over the six-year span.

Table 25
Changes in Emphasis on Coarse-Grained Topics in Mathematics Instruction

Topic	Recentered data 2003-2009		Main data 2003-2006		Main data 2007-2009	
	B	SE(B)	B	SE(B)	B	SE(B)
Grades 3-8						
Number sense/properties/relationships	-0.669***	0.074	-0.357†	0.212	-0.612*	0.237
Operations	-0.592***	0.061	-0.645***	0.177	-0.574**	0.180
Measurement	-0.170***	0.042	-0.074	0.122	-0.355**	0.131
Basic algebra	0.384***	0.049	0.669***	0.132	0.255†	0.132
Advanced algebra	0.031**	0.011	0.010	0.032	0.174***	0.048
Data displays	0.184***	0.029	0.193*	0.077	0.255**	0.099
Statistics	0.234***	0.010	0.036	0.027	0.115**	0.036
Probability	0.207***	0.008	0.070*	0.027	0.061	0.037
Instructional technology	0.556***	0.022	-0.018	0.039	0.199*	0.086
n	10536		5238		5298	

Note. Values are regression coefficients for fixed-effects regression of marginal proportion on year of administration. † $p < .10$, * $p < .05$, ** $p < .01$, *** $p < .001$

Table 26
Changes in Emphasis on Coarse-Grained Topics in ELAR Instruction

Topic	Recentered data 2004-2009		Main data 2004-2006		Main data 2007-2009	
	B	SE(B)	B	SE(B)	B	SE(B)
Grades K-2						
Reading fluency	0.052*	0.023	-0.354***	0.056	0.053	0.061
Reading comprehension	0.624***	0.039	0.255**	0.089	-0.056	0.105
Author's craft	-0.057**	0.022	-0.124*	0.058	0.025	0.051
Language study	-0.380***	0.028	-0.233**	0.072	-0.213**	0.066
n	6571		3279		3292	
Grades 3-8						
Phonemic awareness	-0.229**	0.071	-0.326†	0.168	0.090	0.190
Reading comprehension	0.995***	0.063	0.554***	0.167	0.087	0.157
Language study	-0.582***	0.039	-0.401***	0.115	-0.311***	0.084
n	6807		2861		3946	
Grades 9-12						
Phonics	-0.237*	0.010	-0.004	0.381	-0.410*	0.210
n	1614		554		1060	

Note. Values are regression coefficients for fixed-effects regression of marginal proportion on year of administration. † $p < .10$, * $p < .05$, ** $p < .01$, *** $p < .001$

The results for ELAR, shown in Table 26, illustrate a number of significant shifts, and at least one significant shift for each grade band. At both grades K-2 and 3-8, the largest significant increase was for reading comprehension. At grades K-2, the increase was approximately .6% per year, or 3.0% of instruction over the five-year study period. At grades 3-8, the increase was 1.0% per year, or 5.0% of instruction over the five-year study period. To counteract these increases in reading comprehension, there were decreases in coverage of language study. In the SEC framework, language study includes such topics as spelling, capitalization, syntax, and grammar. The decrease was .4% per year for grades K-2 and .6% per year for grades 3-8, changes of 1.9% and 2.9%,

respectively, over the five-year period. There was a small increase in coverage of reading fluency and decrease in coverage of author's craft (ideas such as plot, theme, and literary devices) at grades K-2, along with a decrease in coverage of phonemic awareness at grades 3-8 and phonics at grades 9-12. By far the topic that received the most increased attention over the study period was reading comprehension.

Finally, there were few significant changes in topics of science instruction across the study period. There were no changes at grades K-2 or 9-12. At grades 3-8, there was a small increase of .2% per year in coverage of science and technology and a corresponding decrease of .2% per year in coverage of properties of matter. There was also an additional small decrease in coverage of nuclear chemistry, but nuclear chemistry was a topic that was sparsely covered at any time point. As was the case with the analyses for levels of cognitive demand, these analyses of coarse-grained topics indicate that the shifts were much smaller in science than in mathematics or ELAR, perhaps owing to the fact that science is tested at fewer grades than mathematics and reading.

Table 27
Changes in Emphasis on Coarse-Grained Topics in Science Instruction

Topic	Recentered data 2003-2009		Main data 2003-2006		Main data 2007-2009	
	B	SE(B)	B	SE(B)	B	SE(B)
Grades 3-8						
Science & technology	0.230***	0.039	0.190†	0.109	0.567***	0.121
Properties of matter	-0.216***	0.047	-0.333*	0.140	0.071	0.135
Nuclear chemistry	-0.048***	0.010	-0.033	0.030	-0.053*	0.024
n	5952		2915		3037	

Note. Values are regression coefficients for fixed-effects regression of marginal proportion on year of administration. † $p < .10$, * $p < .05$, ** $p < .01$, *** $p < .001$

Comparison with Standards and Assessments. As was done for levels of cognitive demand, the significant changes in coverage of coarse-grained topics were compared with the prevalence of those topics on standards and assessments. Only the re-centered data were used for these analyses. The analyses were parallel to the analyses for cognitive demand levels. First, the average proportions of teachers' instruction focusing on each topic in 2002-03 and 2003-04 for mathematics and science and 2003-04 for ELAR were calculated. Next, the available standards and assessments were averaged separately, and the marginal proportion of each topic for the average standard and average assessment was calculated. The direction of the significant change in coverage was compared with the teachers' average 2002-03/2003-04 instruction and the average standard and assessment to see whether teachers' instructional changes brought their instruction more in line with standards and assessments.

The results, shown in Table 28 for mathematics, Table 29 for ELAR, and Table 30 for science, illustrate again that the majority of teachers' instructional changes over the period 2003 to 2009 brought their instruction into closer agreement with standards and, in most cases, assessments. In mathematics, teachers' instructional changes brought their instruction into closer agreement with the standards in eight of nine cases. For assessments, the instructional changes were only in the direction of increased agreement in four of nine cases. In this instance, teachers appear to be somewhat more responsive to the content specified in the standards than in the assessments.

Two notable shifts are the increases in teachers' coverage of statistics and probability. These were topics that were rarely covered by teachers in 2003/2004,

representing 1% or less of the curriculum. However, the two topics are both represented in content standards—statistics at 2.9% of the total content and probability at 4.1% of the total content. Teachers have increased their coverage of these topics, perhaps owing to their inclusion in the standards. However, neither topic was tested at all on any of the grades 3-8 assessments in the database. Instead, the topic of operations was far overrepresented in the assessments as compared to the standards. These findings are suggestive that the assessments in grades 3-8 may not be a fair representation of the content specified in the standards.

Table 28
Comparison of Change in Topic Coverage for Teachers' Mathematics Instruction with Topic Emphasis of Standards and Assessments, 2003-2009

Topic	2003/4 proportion	Coefficient sign	Proportion in standards	Proportion in assessments
Grades 3-8				
Number sense/properties/relationships	0.278	-	0.241*	0.260*
Operations	0.234	-	0.127*	0.302
Measurement	0.146	-	0.138*	0.135*
Basic algebra	0.097	+	0.155*	0.046*
Advanced algebra	0.005	+	0.003	0.000
Data displays	0.066	+	0.086*	0.073*
Statistics	0.010	+	0.029*	0.000
Probability	0.006	+	0.041*	0.000
Instructional technology	0.007	+	0.014*	0.000

Note. A + indicates a positive coefficient on year of administration in fixed effects regression equation. A - indicates a negative coefficient. A * indicates change in instruction that is in the direction of increased alignment with standards or assessments.

In ELAR, shown in Table 29, teachers' instructional shifts were again largely in the direction of increased agreement with the standards. Especially the largest shifts—reading comprehension and language study at grades K-2 and 3-8—brought teachers' instruction into closer agreement with the content of standards. Unlike in mathematics,

however, the instructional shifts also brought instruction into closer agreement with assessments. There were no K-2 ELAR assessments in the database. Of the 12 total comparisons represented in Table 29, the only two instances where the instructional change was not in the direction of increased agreement was for reading fluency and author’s craft at grades K-2, the two smallest significant shifts identified. Even though there was no significant increase in instructional alignment in ELAR, these results suggest that ELAR teachers were responsive to the content of standards and assessments during the period 2004-2009.

Table 29
Comparison of Change in Topic Coverage for Teachers' ELAR Instruction with Topic Emphasis of Standards and Assessments, 2004-2009

Topic	2004 proportion	Coefficient sign	Proportion in standards	Proportion in assessments
Grades K-2				
Reading fluency	0.047	+	0.044	
Reading comprehension	0.113	+	0.209*	
Author's craft	0.043	-	0.046	
Language study	0.090	-	0.056*	
Grades 3-8				
Phonemic awareness	0.046	-	0.006*	0.001*
Reading comprehension	0.146	+	0.244*	0.484*
Language study	0.080	-	0.050*	0.032*
Grades 9-12				
Phonics	0.013	-	0.007*	0.008*

Note. A + indicates a positive coefficient on year of administration in fixed effects regression equation. A - indicates a negative coefficient. A * indicates change in instruction that is in the direction of increased alignment with standards or assessments.

Finally, the few science topics showing significant changes are presented in Table 30, alongside the corresponding proportions in the standards and assessments. For both science & technology and nuclear chemistry, teachers' changes brought their instruction into closer agreement with standards and assessments. For properties of matter, however, teachers' decreasing coverage was in the direction of decreased agreement with standards and assessments. In four of six cases, science teachers' significant shifts in the topics of their instruction were indicative of responsiveness to the content of standards and assessments.

Table 30
Comparison of Change in Topic Coverage for Teachers' Science Instruction with Topic Emphasis of Standards and Assessments, 2003-2009

Topic	2003/4 proportion	Coefficient sign	Proportion in standards	Proportion in assessments
Grades 3-8				
Science & technology	0.030	+	0.070*	0.033*
Properties of matter	0.054	-	0.069	0.070
Nuclear chemistry	0.005	-	0.000*	0.000*

Note. A + indicates a positive coefficient on year of administration in fixed effects regression equation. A - indicates a negative coefficient. A * indicates change in instruction that is in the direction of increased alignment with standards or assessments.

As was the case for levels of cognitive demand, there is evidence that teachers have made several significant shifts in their instruction over the years 2003 to 2009. The most notable shifts are (a) a decrease in coverage of numbers and operations and in grades 3-8 mathematics, (b) an increase in coverage of basic algebra, instructional technology, statistics, and probability in grades 3-8 mathematics, (c) an increase in coverage of reading comprehension in grades K-2 and 3-8 ELAR, and (d) a decrease in coverage of language study in grades K-2 and 3-8 ELAR. In all of these cases, the shifts

were in the direction of increased agreement with standards. It was only when the assessments were unrepresentative of the content of the standards, as in the case of the large emphasis on operations and the lack of coverage of statistics or probability, that the shifts were not also in the direction of increased agreement with the assessments.

The Relationship between State Policy Attributes and Instructional Alignment

The final research question to be addressed is the extent to which state policy attributes and features of the standards and assessments themselves are related to the alignment of teachers' instruction with standards and assessments. To address this question, three-level hierarchical linear models (HLMs) are used. The dependent variables for the HLMs are alignment indices for the alignment of teachers' instruction with standards or assessments.

Because there are many fewer teachers with data on assessments than standards, two models were run for standards—one including all teachers with data on standards, and one including only those teachers with data on both standards and assessments. The former had the benefit of increased power due to larger sample sizes (particularly at levels 2 and 3) but it did not include the CONSISTENCY variable indicating the alignment of standards and assessments. The latter contained this key variable but had reduced samples. In the end, only the models using data on standards and assessments together were deemed satisfactory, because of the significant effect of CONSISTENCY in most models. Thus, the omission of this variable in the models including alignment to standards only led to concerns about omitted variable bias and a lack of interpretability of the level-2 coefficients for that model.

The predictors at level 1 include professional development, professional community, instructional influences, and other teacher-level variables. The predictors at level 2 are the policy attributes and a set of dummy variables for grade and year. While the use of fixed effects for year was seen as important in order to allow the inclusion of all years of data despite the slight changes in the SEC surveys and content analysis frameworks between 2006 and 2007, examination of deviance statistics indicated that the addition of fixed effects for year did not significantly improve model fit for any of the six models. Thus, fixed effects for year were not included in the final models presented below.

Results. The results are grouped by model and then by academic content area. The first models are fully unconditional models, which allow the decomposition of variance across the three levels. The second models are “full models” at levels 1 and 2, which provide the estimates of main effects for the policy attributes. Further models were investigated that included interaction terms among level-1 and level-2 variables, as well as random slopes and intercepts. These exploratory models are described and their results discussed at the end of this section.

Fully Unconditional Models. There are a total of six models to be presented—dependent variables of alignment to standards and assessments in each of the three content areas. Table 31 summarizes the variance decomposition for the six models. Asterisks indicate whether the associated variance components (not presented) are significantly different from zero.

Table 31
Summary of Fully Unconditional Models

Dependent Variable	Variance Decomposition			Intercept
	Level 1	Level 2	Level 3	
Mathematics				
Standards	26.5%	21.7%***	51.8%***	0.182
Assessments	64.7%	25.8%***	9.5%***	0.203
ELAR				
Standards	42.0%	7.5%*	50.5%***	0.222
Assessments	42.9%	46.0%***	11.0%***	0.167
Science				
Standards	69.2%	15.8%***	15.1%***	0.163
Assessments	33.6%	45.2%***	21.2%***	0.165

Note. Significance levels indicate significance of associated variance component. *** $p < .001$, ** $p < .01$, * $p < .05$.

The results of the fully unconditional models suggest that, across all models, there is a moderate to large amount of variance to be explained at levels 2 and 3. The model with the least variance to be explained at level 2 is the model predicting alignment to standards in ELAR. This model has just 7.5% of the total variance to be explained at level 2, though the variance component for this model is still significantly greater than 0. In contrast, the largest variance components at level 2 are in the models predicting alignment to assessments for science and ELAR, both of which have more than 45% of the variance at level 2.

Across all the models, certain patterns in the distribution of variance are evident. In all three subjects, a higher proportion of the variance lies at level 2 in the models for

assessments than in the models for standards. These differences are more pronounced for ELAR and science than for mathematics. The proportion of variance at level 1 is less than half in four of six models. The two exceptions are the model for standards in science and the model for assessments in mathematics, indicating that state, grade, and year effects account for majority of the variance in most cases. Clearly, there is significant variance to be explained by state policy attributes and features of standards and assessments at level 2, though the proportion to be explained varies considerably across models.

Table 32
Full HLM of Alignment to Standards and Assessments in Mathematics

Fixed Effect	Alignment to standards			Alignment to assessments		
	B	se	t ratio	Coefficient	se	t ratio
Level 2						
Intercept	19.613***	0.885	22.155	18.708***	0.707	26.454
Focus	-4.905***	0.477	-10.274	-2.206***	0.412	-5.350
Stability	-0.048	0.146	-0.326	0.477***	0.122	3.904
Consistency	17.766***	2.708	6.559	10.254***	2.367	4.332
Power	0.569†	0.307	1.851	-0.345	0.246	-1.403
Authority	-0.505	0.307	-1.645	-0.119	0.234	-0.509
Level 1						
Experience	-0.011	0.011	-1.012	-0.015	0.011	-1.458
Challenge	0.243	0.236	1.030	0.367	0.235	1.561
SBRinfluence	0.528***	0.108	4.894	0.568***	0.107	5.293
Otherinfluence	-0.765***	0.143	-5.341	-0.831***	0.143	-5.830
Profcomm	0.020	0.109	0.180	-0.160	0.109	-1.475
SBRpd	-0.003***	0.001	-3.833	-0.005***	0.001	-5.533
Random Effect	Variance Component	df	X ²	Variance Component	df	X ²
Level 2	0.00057***	190	484.43	0.00042***	190	519.58
Level 3	0.00029***	6	89.90	0.00015***	6	63.36

Note. † $p < .10$, * $p < .05$, ** $p < .01$, *** $p < .001$. Level 1 $n = 4197$, Level 2 $n = 211$, Level 3 $n = 7$. Alignment index re-scaled to 0 to 100.

Full Models. The results of the full models for mathematics, ELAR, and science are presented in Tables 32, 33, and 34, respectively. The dependent variable alignment index was re-scaled to a 0-to-100 scale for ease of presentation and interpretation of regression coefficients. The values presented in the table are unstandardized regression coefficients. However, standardized coefficients are discussed in the text where applicable. Also, while fixed-effects for grade were included in the models, these fixed effects are not presented in the tables.

Mathematics. The models for mathematics, presented in Table 32, identify several large and significant policy effects on teachers' alignment. Both models show that the degree of FOCUS of the standards is significantly negatively associated with teachers' alignment (for standards, $B = -4.905, p < .001$; for assessments, $B = -2.206, p < .001$). Both models also show that the CONSISTENCY of the standards and assessments—their alignment with one another—is positive and significantly related to alignment (for standards, $B = 17.766, p < .001$; for assessments, $B = 10.254, p < .001$). In addition, the model for standards suggests that the POWER of state standards-based reform policy is positively associated with alignment ($B = .569, p = .065$), while the model for assessments suggests that STABILITY—the number of years the particular standards document has been in place—is positively associated with alignment ($B = .477, p < .001$). There is no relationship between AUTHORITY and alignment in either model. The model for standards explains 63% of the variance at level 2 and 92% at level 3, while the model for assessments explains 45% of the variance at level 2 and 46% at level 3. In both models there remains statistically significant variance to be explained at both levels.

The magnitude of the coefficients for certain policy attributes is moderate to large. For instance, the standardized regression coefficients for FOCUS are -.47 for standards and -.31 for assessments, indicating that an increase of 1σ in FOCUS (.79 on the FOCUS index) is associated with a decrease of between one-third and one-half of a standard deviation in the alignment index. The standardized regression coefficients for CONSISTENCY are .25 for standards and .21 for assessments. The significant effects for POWER and STABILITY have smaller effect sizes.

There are also three significant level-1 predictors of alignment, and all three are common across the two models. These common significant predictors are SBRPD (for standards, $B = -.003, p < .001$; for assessments, $B = -.005, p < .001$), SBRINFLUENCE (for standards, $B = .528, p < .001$; for assessments, $B = .568, p < .001$), and OTHERINFLUENCE (for standards, $B = -.765, p < .001$; for assessments, $B = -.831, p < .001$). There is no effect of EXPERIENCE, CHALLENGE or PROFCOMM on alignment to standards or assessments. The coefficients indicate that an increase in the taking of professional development focused on standards-based reform principles is associated with lower alignment of instruction with standards and assessments. Furthermore, the degree to which teachers report being influenced by standards-based reform-type sources is positively associated with alignment, while the degree to which teachers report being influenced by other sources is negatively associated with alignment.

ELAR. The models for ELAR in Table 33 show fewer consistent policy effects on alignment than the mathematics models. As was the case for mathematics, there is a significant negative effect of FOCUS on alignment to standards ($B = -5.975, p < .001$);

however, there is no significant effect for assessments ($B = -.175, p = .814$). There is also a significant positive effect of POWER on alignment to standards ($B = .571, p = .032$), but no effect for assessments ($B = .326, p = .300$). There is a significant effect of CONSISTENCY on alignment to assessments ($B = 19.919, p = .004$), but no significant effect on alignment to standards ($B = -4.041, p = .413$). There are no significant effects of STABILITY or AUTHORITY on alignment to either standards or assessments. The model for standards explains 67% of the variance at level 2 and 92% at level 3, while the model for assessments explains 31% of the variance at level 2 and 61% at level 3.

Fixed Effect	Alignment to Standards			Alignment to Assessments		
	Coefficient	se	t ratio	Coefficient	se	t ratio
Level 2						
Intercept	21.729***	0.938	23.164	17.422***	1.100	15.852
Focus	-5.975***	0.709	-8.429	-0.175	0.740	-0.236
Stability	-0.284	0.183	-0.155	-0.258	0.221	-1.167
Consistency	-4.041	4.913	-0.822	19.919**	6.663	2.990
Power	0.571*	0.263	2.174	0.326	0.313	1.043
Authority	-0.044	0.363	-0.122	-0.071	0.303	-0.235
Level 1						
Experience	-0.050**	0.017	-2.895	-0.046***	0.011	-4.144
Challenge	0.074	0.357	0.207	0.196	0.228	0.859
SBRinfluence	0.656***	0.172	3.804	0.368**	0.110	3.339
Otherinfluence	0.014	0.246	0.058	-0.456**	0.157	-2.900
Profcomm	0.392*	0.171	2.291	0.026	0.109	0.237
SBRpd	0.003**	0.001	2.835	-0.001	0.001	-1.453
Variance						
Random Effect	Component	df	X ²	Component	df	X ²
Level 2	0.00026***	104	163.37	0.00132***	104	583.04
Level 3	0.00041***	10	85.28	0.00018*	10	22.85

Note. † $p < .10$, * $p < .05$, ** $p < .01$, *** $p < .001$. Level 1 $n = 3741$, Level 2 $n = 128$, Level 3 $n = 11$. Alignment index re-scaled to 0 to 100.

In addition to revealing fewer significant policy effects on alignment, the ELAR models generally identify smaller effects of policy attributes. The standardized effect of FOCUS on alignment to standards is $-.52$, indicating that a 1σ increase in the degree of FOCUS of a state's standards is associated with lower instructional alignment to standards by approximately $\frac{1}{2}\sigma$. The standardized effect size for the positive effect of POWER on alignment to standards is a more modest $.09\sigma$. The effect size for the positive effect of CONSISTENCY is $.17\sigma$, somewhat lower than the effect sizes in mathematics.

As in mathematics, there are also level-1 variables that are significantly related to instructional alignment. For one, teacher report of SBRINFLUENCE is positively related to alignment to both standards and assessments (for standards, $B = .656, p < .001$; for assessments, $B = .368, p < .001$). Teacher report of OTHERINFLUENCE is negatively related to alignment to assessments, but not standards (for standards, $B = .014, p = .954$; for assessments, $B = -.456, p = .004$). In contrast to the mathematics results, where teacher EXPERIENCE was not a significant predictor in either model, EXPERIENCE is negatively related to alignment to both ELAR standards and assessments (for standards, $B = -.050, p = .004$; for assessments, $B = -.046, p < .001$). Also positively related to alignment with standards are PROFCOMM ($B = .392, p = .022$) and SBRPD ($B = .003, p = .005$).

Science. Despite concerns about power, the results for science, presented in Table 34, show a number of level-1 and level-2 variables as significantly related to instructional alignment. In terms of the policy attributes, both CONSISTENCY (for standards, $B = 10.288, p = .029$; for assessments, $B = 59.121, p < .001$) and POWER are positively

related to alignment (for standards, $B = .482, p = .017$; for assessments, $B = .474, p = .091$) in both models. FOCUS is also negatively related to alignment with standards ($B = -4.094, p < .001$), but not assessments ($B = .842, p = .494$). AUTHORITY is positively related to alignment to assessments ($B = 1.465, p < .001$), but not to standards ($B = .333, p = .158$). As in most other models, STABILITY is not significantly related to alignment to either standards or assessments. The models account for 66% of the level-2 and 62% of the level-3 variance for standards and 80% of the level-2 and 58% of the level-3 variance for assessments.

Again, the magnitudes of the policy effects are in the same range as in the models for ELAR and mathematics. The largest effects in science are for CONSISTENCY, with standardized effect sizes of .65 for assessments and .13 for standards. The effect sizes for POWER are .17 for standards and .15 for assessments. The negative effect of FOCUS on alignment to standards is .32 standard deviations, and the positive effect of AUTHORITY on alignment to assessments is .52 standard deviations.

Level-1 variables in the science models follow some similar patterns to the other models. Again, teacher report of the SBRINFLUENCE on their content decisions is significant and positively related to alignment to both standards and assessments (for standards, $B = .474, p = .006$; for assessments, $B = .486, p = .001$), as is teacher participation in SBRPD (for standards, $B = .003, p = .003$; for assessments, $B = .003, p = .003$). There are also significant positive effects of class CHALLENGE ($B = .672, p = .050$) and PROFCOMM ($B = .332, p = .036$) on alignment to standards. There are no significant effects in either model for EXPERIENCE or OTHERINFLUENCE.

Table 34
Full HLM of Alignment to Standards and Assessments in Science

Fixed Effect	Alignment to Standards			Alignment to Assessments		
	Coefficient	se	t Ratio	Coefficient	se	t Ratio
Level 2						
Intercept	16.074***	1.834	8.763	13.885***	2.525	5.498
Focus	-4.094***	0.885	-4.628	0.842	1.227	0.686
Stability	0.142	0.122	1.163	-0.048	0.165	-0.290
Consistency	10.288*	4.651	2.212	59.121***	5.982	9.883
Power	0.482*	0.197	2.448	0.474†	0.277	1.708
Authority	0.333	0.234	1.423	1.465***	0.344	4.256
Level 1						
Experience	-0.017	0.018	-0.926	-0.005	0.016	-0.303
Challenge	0.672*	0.343	1.960	-0.181	0.299	-0.604
SBRinfluence	0.474**	0.170	2.783	0.486**	0.148	3.276
Otherinfluence	0.175	0.223	0.784	-0.099	0.195	-0.507
Profcomm	0.332*	0.158	2.098	0.198	0.138	1.432
SBRpd	0.003**	0.001	3.018	0.003**	0.001	3.082
Random Effect	Variance Component	df	χ ²	Variance Component	df	χ ²
Level 2	0.00013***	80	177.11	0.00039***	80	156.14
Level 3	0.00016***	9	84.17	0.00039***	9	59.35

Note. † p < .10, * p < .05, ** p < .01, *** p < .001. Level 1 n = 1549, Level 2 n = 103, Level 3 n = 10
Alignment index re-scaled to 0 to 100.

Across the six full models, certain patterns in the results emerge. In terms of policy attributes, CONSISTENCY was significant and positively related to instructional alignment in five of six models and POWER was significant and positively related to alignment in four models (marginally in two). FOCUS was significant and negatively related to instructional alignment in four of six models. In contrast, AUTHORITY and STABILITY were both positively related to alignment in just one of six models, and

were not significantly related to alignment in the other five. At level 1, the most consistent predictors were SBRINFLUENCE (positively related in all six models), and OTHERINFLUENCE (negatively related in three models). Other variables were related to alignment in some models but not others.

Sensitivity. One concern about the data used here was that the SEC surveys changed their list of topics between the 2006 and 2007 administrations. The teacher survey data and content analyses from prior to that date were translated into the new content language by staff at the University of Wisconsin. However, all available data, including data collected before and after the transition, were used for the primary analyses just described. In order to ensure that the results were not biased by the inclusion of data from the two different surveys, several checks were undertaken.

The first check of the results was to run identical models to the full models described previously, but using only data from after the 2007 transition to the new survey. The primary disadvantage of this approach is a loss of statistical power, as roughly half of the level 1 observations were lost in this analysis. The full results of the six regressions are presented in Appendix C, Tables C.2, C.3, and C.4. A statistical comparison of the regression coefficients across the models—the main models presented previously compared with the models using just data from 2007-2009—was run (A. Cohen, 1983). With five level-2 and six level-1 coefficients in each model, there were a total of 66 comparisons made. None of these comparisons indicated a significant difference in coefficients at the .05 level of significance. These results are highly

suggestive that the models are similar, and that the change in form did not impact the estimates of effects at level 1 or level 2.

Table 35
Comparison of Models: 2003-2009 Compared With 2007-2009

Policy Attribute	Math Stds.	Math Tests	ELAR Stds.	ELAR Tests	Science Stds.	Science Tests
2003-2009						
Focus	-	-	-		-	
Stability		+				
Consistency	+	+		+	+	+
Power	+		+		+	+
Authority						+
2007-2009						
Focus	-	-	-		-	
Stability		+				+
Consistency	+	+		+		+
Power	+					+
Authority	-				+	+

Note. A + indicates a positive, significant effect ($p < .10$), and a - indicates a negative, significant effect.

If the 2007-2009 data were used for the main analyses just presented, the results in terms of statistical significance of the policy effects on alignment would also be quite similar. Table 35 displays the statistical significance of the five state policy variables for the six models using all data 2003-2009 (top half) and only the data from 2007 and later (bottom half). There are 30 pairs of coefficients (e.g., the coefficient on FOCUS for mathematics standards in 2003-2009 and 2007-2009). In mathematics, 9 of the 10 pairs are the same, with only authority in the model predicting alignment to standards differing. In ELAR also, 9 of 10 pairs are the same, with only POWER in the model

predicting alignment to standards differing. In science, just 6 of 10 pairs agree, with POWER, CONSISTENCY, and AUTHORITY differing for standards and STABILITY differing for assessments. Again, this analysis is supportive of conclusion that the substantive results of the analysis presented here would be quite similar whether data from 2003-2009 or 2007-2009 were used. The change in form did not appear to impact the interpretation of policy effects on teachers' instructional alignment.

The second check of the effect of the SEC survey change on the results was to re-center the data for both teachers and standards/assessments. Again, the method was to first identify those topic areas that appeared on both the pre-2007 and post-2007 surveys. Any topics that appeared on one or the other, but not both, were deleted. The remaining cell proportions were then re-centered by dividing each proportion by the total remaining content. Alignment indices were then re-calculated for each teacher and these new re-centered alignment indices were used as the dependent variables in a set of parallel models to the full models described previously. Also modified were the values for the CONSISTENCY index, because the framework for content analysis used prior to 2007 was the reduced framework used in the teacher survey prior to 2007. Again, any topic that did not appear on both the pre- and post-2007 frameworks was deleted. The content analyses were then re-centered to sum to one, and the CONSISTENCY variable was recalculated. The procedure was done separately for math grades K-8, math high school, science grades K-8, science high school, and ELAR grades K-12, as these were the forms that pre-dated 2007.

The results of the HLMs for the re-centered data, presented in full in Appendix C, correspond closely to the full regressions presented previously. Again, there are a total of 66 possible pairwise comparisons of regression coefficients between the models in tables C.5-C.7 and the main models presented previously. Of these 66, no comparisons show statistically significant differences at the .05 level of significance. Even in terms of statistical significance of policy attributes, the two sets of models are nearly identical as seen in Table 36, with the only difference being statistically significant effects of POWER on alignment to ELAR assessments and of AUTHORITY on alignment to science standards for the models using the shortened SEC. These results are strongly suggestive that the changes made in the SEC survey are not related to the policy effects identified in the HLM analyses.

Table 36
Comparison of Models: Main Data Compared with Re-centered Data

Policy Attribute	Math Stds.	Math Tests	ELAR Stds.	ELAR Tests	Science Stds.	Science Tests
Main data						
Focus	-	-	-		-	
Stability		+				
Consistency	+	+		+	+	+
Power	+		+		+	+
Authority						+
Re-centered data						
Focus	-	-	-		-	
Stability		+				
Consistency	+	+		+	+	+
Power	+		+	+	+	+
Authority					+	+

Note. A + indicates a positive, significant effect ($p < .10$), and a - indicates a negative, significant effect.

Exploratory analysis. Using the main models presented previously as a base, a set of exploratory analyses were run to investigate interactions of policy attributes, policy effects on level 1 slopes, and other model improvements. In all cases, changes in the model were tested for improvement in model fit using a chi-squared test. The results of these exploratory analyses are summarized here.

Previous research (e.g., Desimone et al., 2004) found that certain state policy attributes had interactive effects on student or teacher outcomes. In that case, the interactive effects were between state policy attributes and prior state achievement. However, it is reasonable to think that the policy attributes studied here might have interactive effects with one another. To investigate this possibility, certain key interaction terms were added to the full models. The models were tested for improved model fit and the statistical significance of the coefficients on the interaction terms were examined.

The interaction terms studied were POWER*AUTHORITY, FOCUS*CONSISTENCY, STABILITY*FOCUS, and STABILITY*POWER. The interaction of POWER*AUTHORITY seemed natural given that these two attributes represent different ideas about how to motivate teachers. For models predicting alignment to assessments, FOCUS*CONSISTENCY seemed to be an appropriate interaction term to study because the FOCUS variable is based only on the level of specificity of the standards, while the CONSISTENCY variable indicates the alignment of the standards with the assessment. Thus, a highly focused set of standards that is also highly consistent with its corresponding assessment would likely indicate that the assessment is highly focused as well. The two interaction terms involving STABILITY were proposed because

it was believed that the length a standards document was in place might be related to the extent to which teachers might respond to the degree of FOCUS or POWER associated with state standards-based reform policies.

Across the six models and these four policy interaction terms, there were few instances where the addition of the interaction term significantly improved model fit. The only such cases were a) POWER*AUTHORITY significantly improving the model predicting alignment to mathematics assessments, b) FOCUS*CONSISTENCY significantly improving the models predicting alignment to ELAR and mathematics assessments, c) STABILITY*FOCUS significantly improving the model predicting alignment to ELAR standards, and d) STABILITY*POWER significantly improving the model predicting alignment to mathematics standards. Thus, across these four interactions and six models, only one interaction appeared as significant in more than one model—the interaction of FOCUS with CONSISTENCY. This was a significant negative predictor in the model for ELAR assessments ($B = -14.706, p = .023$) and the model for mathematics assessments ($B = -14.480, p < .001$). These coefficients would imply that alignment to assessments in these subjects is lower when the standards are highly focused and the standards and assessments are highly aligned. However, given the inconsistency with which interaction terms affected alignment, the strength of the conclusions about these interaction terms is limited.

A second exploratory investigation was to examine the extent to which the policy attributes were related to the effects of level-1 variables on alignment. The primary cross-level effects investigated were the effects of the policy attributes on the slope for

EXPERIENCE, SBRINFLUENCE, SBRPD, and PCOMM, as these were the level-1 variables seen as most likely influenced by state policy. Again, there were sporadic significant effects of any of these policy attributes on the slopes for any level 1 variables. Where significant effects appeared, they were in one model and not in the other five. There were no consistent policy effects on the slopes for any of the investigated level-1 variables, suggesting a lack of cross-level interactions in affecting instructional alignment.

The results of both sets of exploratory analyses indicated that adding complexity to the model rarely resulted in improved model fit. No consistent interaction effects were found across the models, including interactions of policy attributes with one another and interactions of level-2 variables with level-1 variables.

Chapter 5

Summary and Discussion

Since the 1990s, states and the federal government have increasingly pursued standards-based educational policies as a means to achieve improvement in the U.S. education system. Standards-based reform begins with the creation of rigorous content standards in the core academic subjects. Together with assessments of student learning and, more recently, accountability measures for students, teachers, schools, and/or districts, these standards are intended to drive teachers to modify the content of their instruction to increase alignment. Through aligned instruction, it is thought that teachers and schools will be able to bring about wide-scale improvement in student learning and narrowing of achievement gaps.

Results of several studies on the nature of SBR's effects on student achievement suggest moderate improvements averaging .2 to .35 standard deviations (Dee & Jacob, 2009; Hanushek & Raymond, 2005; Jacob, 2005), with the strongest effects in mathematics. These results are promising and suggest that SBR is likely having important effects on teachers' instruction. However, high-quality studies on the nature of SBR's effects on instruction are notably lacking, and many of the studies of instruction omit or gloss over the content of instruction. Where content is considered, it is often examined in ways that a) are not useful for estimating magnitudes of effects, b) do not allow for comparisons across states and differing policy contexts, and c) do not include comparisons of instructional change with the content messages of standards and assessments.

Furthermore, there are serious methodological concerns with much of the previous research on content effects of standards-based reform. For one, most of these studies rely on highly problematic reports of change in instruction, rather than a pre-post or comparison group design. For another, the survey studies often use one-item “scales” to construct measures of change in instruction, with obvious concerns about reliability. Even when multiple items are used, the items are not specific enough to allow description of the nature of changes and comparison of changes across settings.

Thus, the purpose of this dissertation was to use data on the content of teachers’ instruction and the content of state standards and assessments to study the effects of standards-based reform on the content of instruction. The data spanned the years 2003-2009 and covered mathematics, ELAR and science. The data included descriptions of current practice based on a survey instrument shown to be a reliable and valid way to represent teachers’ content coverage. The specific questions addressed were:

- 1) To what extent have teachers aligned their instruction with standards and assessments in ELAR, mathematics, and science, since 2003?
- 2) In what ways have teachers changed the distribution of instructional time and the topics and cognitive demand levels of their instruction in the years since 2003, and to what extent are the changes in topics and cognitive demand reflective of the content messages of standards and assessments?
- 3) To what extent are state policy contexts, features of standards and assessments, and features of local context related to alignment?

The methods used to address the research questions included fixed effects regression and hierarchical linear modeling. A sample of approximately 16,000 teachers in mathematics and ELAR and 9,500 teachers in science was used for descriptive analyses. For analyses of alignment, samples included approximately 4,000 teachers in mathematics and ELAR and 1,500 in science, scattered across 7 to 10 states. While not nationally representative, the teachers represented a wide range of states and policy contexts.

Summary and Discussion

Change in Instructional Alignment, 2003-2009. Alignment indices were calculated for each teacher in the sample to standards and assessments for all states with such data. The indices were then analyzed using fixed effects regression. In mathematics, there were increases across years in alignment to standards and assessments in most grades, with magnitudes of between .25 and .37 standard deviations. Slightly larger magnitudes were found when the SEC data was re-centered by deleting topics not included in both the pre-2007 and post-2007 versions of the survey and re-normalizing so that the remaining cell proportions summed to 1. On the re-centered data, the increases in alignment were up to .52 standard deviations, depending on the target (standards or assessments) and grade level. The only mathematics model that did not show consistent increases in alignment was grades 9-12 standards, which showed a significant decrease in alignment in one model and no change in another.

In ELAR, the results were less clear. The main models using the complete SEC data showed a decrease in alignment to standards at grades K-2 and no change in alignment for standards or assessments at any other grade level. The models using the re-

centered data showed a significant increase in alignment to standards at grades K-2 and 3-8 and assessments at grades 3-8, and no change in alignment to standards or assessments at grades 9-12. These results for ELAR were inconclusive and suggested that any changes that may have occurred in instructional alignment were likely small and possibly dependent on the particular topics in the taxonomy.

In science, there was a clear increase in alignment to standards in grades 3-8, with a magnitude of .20 to .43 standard deviations. There also appeared to be a decrease in alignment to assessments at grades 3-8, with approximately the same magnitudes. There were conflicting findings about alignment to standards for grades K-2 and standards and assessments for grades 9-12, with some models suggesting decreases in alignment and others suggesting no change. There was little evidence of a consistent increase in alignment to either standards or assessments at grades other than 3-8, however.

State-by-state analyses suggested that the trends toward increased alignment did not necessarily operate consistently across states. For instance, while alignment to standards in grades 3-8 science increased significantly over the study period, teachers in Maine significantly decreased the alignment of their instruction with both standards and assessments at that grade. In contrast, Ohio teachers significantly improved the alignment of their instruction with standards in grades 3-8 and standards and assessments in grades 9-12 in science.

Overall, the results of this analysis suggest that there are changes in instructional alignment taking place, with the largest and most consistent changes in mathematics and in grades 3-8 in science. The finding that alignment has increased in mathematics while

not changing in ELAR would fit with recent findings about the nature of changes in student achievement from Dee and Jacob (2009). They found that student achievement in mathematics increased as a result of a standards-based reform by approximately .20 standard deviations, but that it did not change in ELAR. Porter and colleagues (2007) also found that effects on aligned instruction were largest in mathematics, though their effects were for standards but not assessments.

However, the lack of consistent patterns in terms of changes in alignment was troubling, especially given the near unanimity with which teachers report efforts to improve their instructional alignment across subjects and grades (Hamilton & Berends, 2006; Koretz, Barron et al., 1996; Koretz, Mitchell et al., 1996; Pedulla et al., 2003; Stecher & Chun, 2001), and it prompted investigation of the specific types of changes made by teachers to their instruction and the ways in which those changes were, broadly speaking, reflective of the content messages of the standards and assessments.

Changes in Instructional Time, Cognitive Demand, and Topics of

Instruction. The SEC data also allowed for a finer-grained investigation of the changes made in teachers' instruction than the survey questions used to study instructional change previously. The SEC data were used to compare teachers' coverage of levels of cognitive demand, as well as their coverage of coarse-grained topics. It was also possible to use teachers' reports of the number of hours per week of instruction to examine whether instructional time has been shifted to or from tested subjects during the study period.

In terms of instructional shifts, the results suggested that shifts in instructional time during the period 2003-2009 were small in magnitude. The only significant shifts

were a small increase in hours of instruction in mathematics at grades 3-8 and a small decrease in hours of instruction in science at grades 9-12. Over the full six-year study period, these changes amounted to just 10 to 20 minutes per week. Even at early elementary grades, where teachers are in self-contained classrooms, there were no changes in instructional time during the study period. These results suggest that the large shifts in instructional time reported in earlier surveys of teachers (Koretz, Barron et al., 1996; Koretz, Mitchell et al., 1996; Stecher et al., 2000; Stecher & Chun, 2001) have largely died out. While surveys as recently as 2005 (Hamilton & Berends, 2006) found that teachers were still reporting shifts in instructional time from non-tested to tested subjects, such shifts are not supported by these results. This evidence suggests that U.S. teachers may have reached a “steady state” in terms of instructional time, with ELAR receiving approximately 7 hours of instruction per week, mathematics receiving 5, and science receiving 4.

A second question was whether teachers were shifting the cognitive demand levels of their instruction, and to what extent such changes were reflective of the cognitive demand levels emphasized in standards and assessments. In all three subjects and at all grades except science K-2, fixed-effects regressions revealed significant shifts in coverage of the various levels of cognitive demand. In mathematics at all three grade levels, the shifts were in the direction of increased emphasis on procedural skills and away from higher-order skills (at grades K-8) and memorization (at grades 9-12). The magnitudes of the shifts were largest at the lowest grades. There, teachers increased their

coverage of procedural skills by approximately 1% per year, or a shift of roughly 11 days' instructional time over the six-year period.

The cognitive demand shifts were similarly in the direction of increased emphasis on lower-level skills in science. At grades 3-8, there was a significant increase in the coverage of procedural skills. At grades 9-12, there was a significant increase in the coverage of memorization. These magnitudes were somewhat smaller than in mathematics, but they still comprised roughly one-half days' instruction per academic year. There were corresponding decreases in coverage of level F cognitive demand, which includes applying concepts and making connections.

In ELAR, the shifts in cognitive demand were more complicated. At the early grades, there was a definite shift toward the lowest three levels of cognitive demand and away from the top two levels. This was the grade and subject at which the shift was largest—approximately 1.6% per year, or almost three days' instruction. Over the course of the five-year study period, the total shift represents approximately 13 days of instruction shifted from analyzing, investigating, and evaluating to memorizing, performing procedures, explaining, and creating. At grades 3-8 and 9-12, however, the shifts were primarily toward cognitive demand level D, generate/create/communicate, which includes skills such as expressing new ideas, integrating with other topics or subjects, and organizing ideas. At grades 3-8, the shifts were away from levels B, E, and F, while at grades 9-12 the shifts were away from level B only. Again, the magnitudes of the shifts were larger in ELAR than in other subjects.

The direction of the shifts was compared with teachers' average instruction in 2003-04 and the content of average standards and assessments for each grade band. In mathematics, the shifts in cognitive demand moved instruction closer to alignment with standards and assessments in 17 of 18 cases. In ELAR, it was 15 of 21 cases, and in science it was 9 of 14 cases. Only one of the shifts—the decrease in cognitive demand level E in grades 3-8 ELAR instruction—was in a direction that decreased agreement with both standards and assessments. Even for ELAR grade 9-12 instruction, where the cognitive demand shift broke the pattern and favored higher-level skills, the evidence suggests that teachers' instruction was moving closer to agreement with standards or assessments in 7 of 8 cases.

A similar investigation was done for coarse-grained topics in the SEC framework. The analyses were done using the re-centered SEC data, and the results were confirmed using the main data, separated into two samples to correspond to the different SEC surveys. Only when there was agreement across the two models were the changes deemed significant and reported. The results indicated a high degree of teacher responsiveness to the topics emphasized in the standards and assessments in ELAR and mathematics, with not enough significant instructional changes to make a conclusion in science. In mathematics, teachers appeared to shift the topics in their instruction toward those emphasized in the standards, with eight of nine significant shifts in the direction of increased agreement. Just four of the nine shifts brought instruction into better agreement with assessments, perhaps because the assessments did not seem to adequately represent the content emphasized in the standards. In particular, though statistics and probability

together comprised roughly 7% of the typical grades 3-8 standards document, and teachers increased their coverage of these topics accordingly, neither topic was included at all in any assessment in the database.

In ELAR, the results were even more supportive of the conclusion that instructional shifts were in the direction of improved agreement, despite the earlier finding of no significant increase in ELAR alignment. At the coarse-grained level of detail, teachers at grades K-2 and 3-8 spent significantly more time covering reading comprehension and significantly less time covering language study. These shifts each brought instruction toward closer agreement with standards. The largest shift was a 1% per year increase in coverage of reading comprehension at grades 3-8. This large shift was notably in the topic area that was most emphasized across assessments—reading comprehension comprised 48% of the typical grade 3-8 assessment’s content. It seems likely that teachers have identified the prominent role of reading comprehension in their state tests and begun increasing their focus on this topic. Reading comprehension now comprises roughly one-fifth of the typical grade 3-8 ELAR instruction.

Overall, the results of the analyses of topic and cognitive demand shifts indicated that instructional shifts were most often in the direction of increased agreement with standards and assessments. For instance, most analyses at most grades showed that teachers were reducing instruction on higher-order thinking and focusing more on procedural or rote skills, and analyses of the standards and assessments at those grades suggest that these changes by teachers were, in fact, reflective of the intentions of standards-based reform. In mathematics and ELAR, teachers were shifting content

toward increased agreement with standards and assessments, except in a few cases where the assessments seemed to over- or under-emphasize certain topics.

Even when the standards and assessments called for teachers to focus on higher-order skills, as was the case for ELAR in grades 9-12, teachers responded appropriately by increasing emphasis on those skills. Thus, the results support early studies of performance-based assessments systems (Koretz, Mitchell et al., 1996; Koretz, Stecher, Klein, & McCaffrey, 1994), which indicated that teachers were, broadly speaking, responsive to the nature of the instruction emphasized in the standards-based accountability system. Perhaps if teachers' current instructional targets were more challenging and emphasized higher-level skills, teachers would increase their focus on those desired skills. Finally, when the standards emphasized the introduction of previously not-taught topics, such as probability and statistics at grades 3-8, teachers responded by increasing their instruction on those topics. Perhaps if the assessments were fairer representations of the standards, teachers would respond even more strongly by moving toward aligned instruction.

The Relationships of State Policies with Alignment. The final question was whether and to what extent state policy was related to teachers' instructional alignment. This question was motivated by a desire to inform the creation and support of high-quality policy systems that would have increased effectiveness in terms of influencing teachers' instruction. A series of HLMs were run, with alignment to standards or assessments as the dependent variables. Level-2 variables were the degree of FOCUS of state standards, the STABILITY of the state standards document, the CONSISTENCY of

the state standards with their corresponding assessments, the POWER of state standards-based reform policies, and the degree of AUTHORITY in the state standards document.

Table 37
Standardized Regression Coefficients for Predicting Alignment from Policy Attributes

	Mathematics		ELAR		Science	
	Standards	Assessments	Standards	Assessments	Standards	Assessments
Focus	-0.63 ***	-0.46 ***	-0.47 ***	-0.02	-0.41 ***	0.09
Stability	-0.01	0.19 ***	-0.06	-0.02	0.05	-0.02
Consistency	0.22 ***	0.20 ***	-0.03	0.16 **	0.11 *	0.65 ***
Power	0.11 †	-0.11	0.11 *	-0.08	0.13 *	0.13 †
Authority	-0.12	-0.07	-0.01	-0.02	0.11	0.52 ***

Note. † $p < .10$, * $p < .05$, ** $p < .01$, *** $p < .001$.

Results suggested that several policy attributes were particularly strongly related to aligned instruction. Table 37 summarizes the six models by providing standardized regression coefficients for the six policy attributes. Three of the policy attributes were significant predictors of alignment in four or more models: FOCUS, CONSISTENCY, and POWER.

More focused standards were associated with lower alignment to standards in all three subjects, with an average effect size of .5 standard deviations. More focused standards were also associated with lower alignment to assessments in mathematics, with approximately the same magnitude effect size, though there was no effect on alignment to ELAR or science assessments. These results clearly suggest that the alignment of teachers' instruction to standards is highest in states and grades where the standards cover many topics, as opposed to few. This finding fits with research suggesting that teachers are less willing to remove topics from their instruction than they are to add topics (Floden

et al., 1980). Thus, standards that are highly focused may violate teachers' norms about content coverage by requiring them to make changes they are not comfortable making. Teachers already cover large numbers of topics—in terms of cells in the SEC taxonomy, an average of 246 in mathematics (out of a possible 1085), 311 in ELAR (out of a possible 805), and 247 in science (out of a possible 1055). Thus, a standards document like Maine's grade-level expectations for fifth grade ELAR, which contain content in just 49 SEC cells, would require vast shifts in order to improve instructional alignment. In contrast, California's eighth grade standards, which contain 260 SEC cells, would presumably require much smaller shifts from teachers to improve alignment.

Another consistent finding is that alignment to standards and assessments is higher when the standards and assessments provide consistent messages about the content that is important to teach. This finding held in all models but the model for ELAR standards, with an average effect size across the six models of .22 standard deviations. This finding directly supports alignment as being an effective principle in the design of standards-based policy systems. Given that, in this sample, the average alignment of state standards with assessments was below .30 in all three subjects, there is clearly room for improvement in terms of the alignment of the system. For instance, if the average level of consistency in each subject were increased to be equal to the maximum level of consistency in each subject (.51 in mathematics, .36 in ELAR, .37 in science), the expected increase in average instructional alignment would be .039 for mathematics standards, .023 for mathematics assessments, .028 for ELAR assessments, .010 for science standards, and .059 for science assessments. These changes amount to

standardized differences of as much as one full standard deviation in alignment, for science assessments. Not surprisingly, in a system designed on the principle of system-wide change, greater coherence is associated with greater policy effectiveness.

Consistency of messages also seems to matter at the individual teacher level. In all six models, teachers reporting that their instruction was influenced by SBR-type sources, such as standards and assessments, had greater alignment of their instruction with standards and assessments. In contrast, for standards and assessments in mathematics and assessments in ELAR, teachers reporting that their instruction was influenced by other sources, such as pre-service training and parent preferences, had less aligned instruction. These findings are also supportive of the importance of consistency in terms of teachers' perceptions of their instructional influences. If teachers feel they are receiving their instructional messages from standards and assessments, they are more likely to practice aligned instruction.

The third consistent finding is that teachers in states with more standards-based reform policies emphasizing power (i.e., reward and sanctions) had higher alignment to standards. The average effect size was .12 standard deviations, somewhat smaller than the effects for consistency or focus. This evidence suggests that attaching some degree of rewards and sanctions to performance is associated with improved alignment. Put another way, even if the system is highly aligned, a standards-based reform policy system with no consequences is likely to foster less-aligned instruction than a system with consequences for students, teachers, and schools. For instance, based on the size of the coefficient, the average alignment index to ELAR standards would be expected to be approximately .034

units larger on the alignment scale (approximately .4 standard deviations) for teachers in Ohio, a high-power state, than in Montana, a low-power state. There may also be negative effects of power on teachers and students, but these findings indicate that at least one positive effect is a focusing of attention on the content in state standards. There is no apparent corresponding effect on alignment to assessments.

There are no consistent effects of policy attributes of stability or authority on teachers' instructional alignment to either standards or assessments. Even investigations of interaction terms suggest limited utility of these variables in explaining teachers' aligned instruction. There are also no consistent effects of professional development around standards-based reform or the school's professional community on aligned instruction. These findings, especially the lack of effect of stability on alignment to standards, are somewhat surprising. Of course, there is the possibility that there are alternative ways of operationalizing those policy attributes that would result in significant effects. An alternative authority variable, created from *Education Week* indicators of the extent to which the state provides supports to help teachers improve their instruction (e.g., professional development, induction) found similar low and non-significant effects. It is also possible that the facts of who was involved in the standards-development process and how the standards relate to national professional standards are less important in terms of affecting teachers' decisions about what to teach than is the actual content of those standards and assessments and the stakes attached to their implementation.

Limitations

The primary limitations of this research have to do with the composition of the sample and the generalizability of the findings. As mentioned previously, the sample was not nationally representative, and was instead a convenience sample of teachers and states based on the data that were already collected by the Wisconsin Center for Educational Research. The results would have more external validity if the sample was a population-based probability sample of teachers, but that was not possible in this case. Nonetheless, the sample was large and contained teachers from many grades and states. Furthermore, in most cases the sample descriptive statistics were similar to the population as a whole. Thus, while the results may not necessarily be perfectly representative of the population of U.S. public school teachers, there is some reason for confidence about the generalizability of the findings.

A second limitation is that the data did not allow for estimation of changes in alignment at the individual teacher level. While there were some teachers who appeared in the database more than once, there were not enough of them to be able to estimate changes in alignment or effects of policies on changes in alignment. It might also have been possible to investigate changes in alignment if the teacher samples were more stable across time. Collecting SEC data in a way that would allow for the examination of trends in content coverage across settings would be a useful endeavor, in order to investigate questions related to instructional changes.

Of course, the fact that the SEC survey changed in the middle of the study period created some difficulties for analysis and interpretation. If that change had not taken

place, there might have been somewhat more confidence in the results on the degree of changes in teachers' instructional alignment, though analyses of the re-centered data found some of the same results as the main analyses. However, there is no evidence that the changes in survey topics affected the substantive findings about policy effects or changes in cognitive demand coverage, so there is more reason to be confident as to the validity of those findings.

Finally, the analyses of the effects of policy attributes on instructional alignment were necessarily correlational, due to the nature of the data. Thus, it is impossible to support claims about causal relationships between the features of state policy and the content of teachers' instruction. It would be a worthwhile study to use data on teachers' instruction across time, alongside changes in state policy, to attempt to isolate causal impacts. However, such analyses could not be done with the available data, so the results are merely suggestive of potentially causal relationships.

Given these limitations, it is also important to emphasize what this study was able to contribute to the literature. Previous survey studies of teachers' instruction relied on problematic teacher self-report of instructional change. Furthermore, those studies did not allow for estimation of magnitudes of teacher-reported changes. The data used here are limited in their generalizability, but they are strong in terms of their validity in representing the content of instruction. Additionally, the ability to directly compare instruction with the content of standards and assessments is a strength of this study. The methods used here could easily be used with more robust, representative samples and different policy measures to conduct further analyses. Still, the results presented in this

research provide new, stronger evidence about the effects of standards-based reform on the content of teachers' instruction.

Implications for Research

The results of this study clearly suggest that some of the findings of previous research on content effects of standards-based reform were overstated or misleading. Certainly, in the years since 2003, there is evidence that teachers have increased the alignment of their instruction with standards and assessments in mathematics and with science standards in grades 3-8. However, there is no evidence of the wide-scale increases in content alignment reported by teachers in previous survey studies. Several possibilities could explain the discrepancy. For one, survey results from previous studies might represent theory-driven reconstructions of teachers' changes in their instruction (Schwartz & Oyserman, 2001). That is, since alignment of instruction to standards and assessments is an explicit purpose of standards-based reform, teachers may have identified increases in alignment that did not actually occur. A related explanation is that teachers may believe they have increased alignment because their school or district tells them so. Many districts provide their own standards or pacing guides that they say are aligned with the standards, and many textbook series claim alignment as well. Thus, teachers might receive these messages about the curriculum they are teaching and believe their instruction is well aligned, even if it is not. A third possibility is one of different interpretations of "alignment." Teachers may not think of alignment in the same way, at the intersection of fine-grained topics and levels of cognitive demand, as it is operationalized in this paper. Surely, if broader categories were used, alignment would be

higher, and this could be what teachers are thinking of. However, in the metric of the SEC alignment index it is clear that alignment to standards and assessments is no better than moderate, and that there is substantial room for improvement in alignment if that improvement is sought.

A second implication is that the findings here lend additional support for the conceptual framework, particularly alignment theories and, to a lesser extent, motivational theories, which underpin standards-based reform. Specifically, three of the policy attributes—power, focus, and consistency—are related to teachers' decisions about what to teach. Clearly, there is a great deal of variation in alignment to standards and assessments at the state and grade levels—in many cases more than 50% of the variance is found at these higher levels. The results described here indicate some of the ways that state policy can explain this higher-level variance. Indeed, the level-2 variables used here account for an average of approximately two-thirds of the total level-2 variance in alignment. However, there is additional significant variance to be explained at these higher levels, and future research should further investigate the ways in which state policies influence content decisions.

A third implication of these findings is for thinking about variation in student access to rigorous, aligned curricula. The results presented here describe what is true on average in terms of changes in alignment and cognitive demand over time. But the results say nothing about whether these effects are operating similarly in all kinds of classrooms for all kinds of students or whether the average effects mask large variations in the effects across settings. For instance, it would be reasonable to think that shifts in the cognitive

demand of instruction to match with standards and assessments would be more pronounced in schools where the assessments and the rewards and sanctions attached to them are more salient, such as schools where failing to make adequate yearly progress is a worry. The SEC data used here could be used to answer such questions, and that research will be done in the future. There is ample room for high-quality research on the differential effects of standards-based reforms on the content of instruction across settings.

Fourth, the results presented here indicate that some of the changes in the content of instruction associated with standards-based reform are non-trivially large. Especially consequential are the shifts in cognitive demand of instruction across years. Previous survey studies of the effects of standards-based reform on the content of instruction were unable to estimate the magnitude of these effects because of the ways in which they asked teachers about their instruction. Researchers studying instructional effects of educational reforms should consult the literature and ask questions that will allow them to come to valid conclusions. Suggestions include asking teachers about current practices rather than past or future, asking the same questions across surveys to allow for comparison of the results, and asking about instruction in a low-stakes context.

Finally, more research must be conducted on the connections between high-quality, aligned instruction and student learning. Standards-based reform is predicated on the notion that aligned instruction will lead to improvements in student achievement, and some research suggests that improvements in achievement have begun to materialize. But explicit connections between aligned instruction and value-added have only been made

sporadically. Research should no longer gloss over content in predicting student achievement gains, but neither should research focus on instructional change without connecting it to student learning. This study represented a first, rigorous step at studying instructional content, but student learning must be the end goal of this research.

Implications for Policy

The results presented here also have important implications for policy, as standards-based reform continues to mature. The most obvious implication is that the alignment of standards with assessments appears to be an important feature for affecting teachers' content decisions. While this idea was one of the key underlying concepts guiding the creation of standards-based reform, the low average standard-assessment alignment found here indicates that the standards-based reform system is not nearly as tightly coupled as was originally intended. In many cases, teachers are receiving vastly different messages about what to teach depending on whether they are looking at standards or assessments. To maximize the effectiveness of the standards-based reform policy, better effort needs to be made to align standards with assessments. The first step is building standards that represent the important content we desire students to know and be able to do. Then, assessments should be constructed that sample from the domain of the standards. Especially if there are to be accountability measures taken on the basis of the test results, the tests must be a fair representation of the content messages of the standards. The alignment of a particular test with a particular set of standards will never be 1.0, nor should it be. But an approach to test construction based on using the standards

as a domain to be sampled should result in improved alignment of instruction to standards.

A second implication for policy is that teachers do appear to be responsive to the content messages of standards and assessments, at least when measured at a broad level of detail. When the standards and assessments emphasized higher-order thinking skills, such as in high school ELAR, teachers responded by shifting their instruction in that direction. In contrast, when the standards and assessments emphasized procedural thinking, such as in K-2 mathematics, teachers responded by shifting their instruction in that direction. These results suggest that the effect of standards and assessments on curriculum need not be a “dumbing down” effect. Rather, a standards and assessment system emphasizing higher-order thinking skills is likely to result in teachers teaching more of that kind of thinking. While building assessments to tap into higher order thinking can be difficult, especially when there are constraints in terms of timing and scoring, policymakers should think about how standards and assessments can be better developed to provide teachers with messages that are consistent with the sort of instruction that is desired.

A third implication for policy is in thinking about the focus of teachers’ instruction. One of the persistent complaints about U.S. education, especially in mathematics (Schmidt et al., 2001), is that it is too broad and shallow. Clearly, teachers’ instruction currently supports this conclusion, as is evidenced by the large number of topics teachers cover in a year. While it might be hoped that more focused standards and assessments would drive teachers to narrow their instruction, it does not appear that

standards and assessments alone can accomplish this change. If policymakers want to focus teachers' instruction on a few core ideas, more work will have to be done in terms of district planning, textbook development, and professional development. Highly focused standards seem to be difficult for teachers to implement with fidelity based on their current practices alone.

Fourth, there are apparently differences across subjects in the size and nature of teachers' changes in instruction and alignment. The most consistent effects were in mathematics, which saw increases in alignment across the study period, changes in topics and cognitive demand levels that were reflective of the standards and assessments, and significant effects of the key policy attributes—focus, consistency, and, for standards, power—on aligned instruction. The least consistent effects were in ELAR. In that subject, there were mixed results as to increases in alignment depending on the SEC taxonomy used and smaller and inconsistent policy effects on alignment, especially for assessments. However, there were large and significant shifts in topics and cognitive demands, most notably a dramatic increase in the coverage of reading comprehension. In short, ELAR teachers seem to be responding somewhat differently than mathematics and science teachers to SBR. There are several possible explanations for the different results in ELAR. One is that the teacher samples in ELAR are worse in some way, and that the findings represent an internal validity problem. The descriptive statistics highlighted that the teacher samples were least like national averages in grades K-2 and 3-8 ELAR. Another is that ELAR teachers respond to SBR policies in fundamentally different ways than mathematics and science teachers do, perhaps because of their training or

philosophy. Or, perhaps, the content in ELAR standards and assessments is less able to be clearly understood, taught, and/or reported on the SEC than the content in mathematics and science, possibly because it may be seen as less discrete. There are other possible explanations as well. But the mixed findings for ELAR in this study, combined with the lack of SBR effects on achievement in ELAR found in other studies, suggest that different approaches may be needed to change ELAR instruction through state policy.

Finally, the results of this study lend some support to the use of rewards and sanctions to provide external motivation for educators to implement standards-based reform. What cannot be identified from this research is whether there are other, negative consequences of power on instruction. That is, the implementation of more kinds of accountability for students, teachers, and schools might result in increases in aligned instruction. Whether such increases would be a net positive for the educational enterprise is hard to say without documentation of all the effects, positive and negative, of rewards and sanctions on educators and students.

While the results described here do not provide definitive evidence for policymakers, they are strongly suggestive of ways that policies could be better designed to strengthen the impact of standards-based reform on teachers and students. Standards-based reform has been a prominent policy for roughly two decades, but there is work to be done if its theory of change is to produce the expected outcomes in terms of teacher performance and student learning.

Appendix A. Coarse-Grained Topics in the SEC Framework

Mathematics

- Number sense
- Operations
- Measurement
- Consumer applications
- Basic algebra
- Advanced algebra
- Geometric concepts
- Advanced geometry
- Data displays
- Statistics
- Probability
- Analysis
- Trigonometry
- Special topics
- Functions
- Instructional technology

Science

- Nature of science
- Science & technology
- Science, health, & environment
- Measurement and calculation in science
- Components of living systems
- Biochemistry
- Botany
- Animal biology
- Human biology
- Genetics
- Evolution
- Reproduction and development
- Ecology
- Energy
- Motion and forces
- Electricity
- Waves
- Kinetics and equilibrium
- Properties of matter
- Earth systems

- Astronomy
- Meteorology
- Elements and the periodic system
- Chemical formulas and reactions
- Acids, bases, and salts
- Organic chemistry
- Nuclear chemistry

ELAR

- Phonemic awareness
- Phonics
- Vocabulary
- Text and print features
- Fluency
- Comprehension
- Critical reasoning
- Author's craft
- Writing processes
- Elements of presentation
- Writing applications
- Language study
- Listening and viewing
- Speaking and presenting

Appendix B. *Cognitive Demands Levels in the SEC Framework*

Mathematics

B. Memorize

- Recite basic mathematics facts; Recall mathematics terms and definitions; Recall formulas and computational processes

C. Perform Procedures

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

D. Demonstrate understanding

- Communicate new mathematical ideas; Use representations to model mathematical ideas; Explain findings and results from data analysis; Develop/explain relationships between concepts; Explain relationship between models, diagrams, and other representations.

E. Conjecture, generalize, prove

- Determine the truth of a mathematical pattern or proposition; Write formal or informal proofs; Analyze data; Find a mathematical rule to generate a pattern or number sequence; Reason inductively or deductively; Use spatial reasoning.

F. Solve non-routine problems, make connections

- Apply and adapt a variety of appropriate strategies to solve problems; Apply mathematics in contexts outside of mathematics; Recognize, generate, or create patterns; Synthesize content and ideas from several sources.

English Language Arts and Reading (ELAR)

B Memorize/recall

- Reproduce sounds or words; Provide facts, terms, definitions, conventions; Locate literal answers in text; Identify relevant information; Describe.

C. Perform procedures/explain

- Follow instructions; Give examples; Check consistency; Summarize; Identify purpose, main ideas, organizational patterns; Gather information.

D. Generate/create/demonstrate

- Create/develop connections among text, self, world; Recognize relationships; Dramatize; Order, group, outline, organize ideas; Express new ideas (or express ideas newly); Develop reasonable alternatives; Integrate with other topics and subjects.

E. Analyze/investigate

- Categorize/schematize information; Distinguish fact and opinion; Compare and contrast; Identify with another's point of view; Make inferences, draw conclusions; Predict probable consequences.

F. Evaluate

- Determine relevance, coherence, internal consistence, logic; Assess adequacy, appropriateness, credibility; Text conclusions, hypotheses; Synthesize content and ideas from several sources; Generalize; Critique.

Science

B. Memorize

- Recite basic science facts; Recall science terms and definitions; Recall scientific formula.

C. Perform procedures

- Make observations; Collect and record data; Use appropriate tools; Make measurements, do computations; Organize and display data in tables or charts; Execute procedures; Conduct experiments; Generate questions, make predictions; Test effects of different variables.

D. Communicate understanding

- Explain concepts; Observe and explain teacher/student demonstrations; Explain procedures and methods of science and inquiry; Organize and display data in tables or charts; Student presentations of science information.

E. Analyze information

- Classify and compare data; Analyze data, recognize patterns; Reason inductively or deductively; Draw conclusions; Identify faulty arguments or misrepresentations of data.

F. Apply concepts/Make connections

- Apply and adapt science information to real world situations; Apply science ideas outside the context of science; Build or revise theory/plan and design experiments; Synthesize content and ideas from several sources; Use and integrate science concepts.

Appendix C. Supplementary Tables

Table C.1
Changes in Emphasis on Five Levels of Cognitive Demand Using Recentered Data

Grade	n	Level B		Level C		Level D		Level E		Level F	
		B	SE(B)	B	SE(B)	B	SE(B)	B	SE(B)	B	SE(B)
Mathematics											
K-2	1808	0.23	0.26	0.92***	0.23	0.38†	0.21	-0.88***	0.17	-0.66***	0.19
3-8	10536	-0.08	0.06	0.51***	0.06	-0.02	0.06	-0.31***	0.05	-0.11*	0.05
9-12	2411	-0.35†	0.17	0.46*	0.20	0.16	0.18	-0.14	0.14	0.02	0.18
ELAR											
K-2	6571	0.44***	0.11	0.32***	0.09	0.82***	0.07	-0.53***	0.07	-0.66***	0.07
3-8	6807	-0.42***	0.11	0.10	0.10	1.09***	0.09	-0.36***	0.03	-0.39***	0.08
9-12	1614	-1.50***	0.28	-0.54*	0.21	1.48***	0.26	0.43*	0.19	0.13	0.22
Science											
K-2	826	0.03	0.11	-0.35	0.34	-0.14	0.32	0.14	0.25	0.25	0.29
3-8	5952	-0.07	0.11	0.14*	0.06	0.18**	0.06	-0.07	0.05	-0.20**	0.06
9-12	2775	0.49**	0.17	-0.02	0.09	0.05	0.10	-0.21**	0.03	-0.30*	0.12

Note. Values are regression coefficients for fixed-effects regression of marginal proportion on year of administration.

† $p < .10$, * $p < .05$, ** $p < .01$, *** $p < .001$

Table C.2

HLM of Alignment to Standards and Assessments in Mathematics Using 2007-2009 Data

Fixed Effect	Alignment to Standards			Alignment to Assessments		
	B	se	t ratio	Coefficient	se	t ratio
Level 2						
Intercept	19.161***	0.987	19.420	19.042***	0.913	20.856
Focus	-5.328***	0.615	-8.664	-1.840**	0.571	-3.222
Stability	0.005	0.271	0.019	0.650*	0.252	2.581
Consistency	20.542***	3.537	5.808	11.507***	3.287	3.500
Power	0.908*	0.418	2.171	-0.294	0.387	-0.760
Authority	-0.777*	0.332	-2.341	0.044	0.307	0.144
Level 1						
Experience	-0.013	0.016	-0.813	-0.009	0.017	-0.528
Challenge	0.633†	0.364	1.740	0.413	0.368	1.122
SBRinfluence	0.518**	0.161	3.227	0.668***	0.163	4.110
Otherinfluence	-0.693**	0.214	-3.239	-0.905*	0.217	-4.177
Profcomm	-0.025	0.164	-0.152	-0.195	0.166	-1.173
SBRpd	-0.004**	0.001	-3.271	-0.006**	0.001	-4.448
Variance						
Random Effect	Component	df	X ²	Component	df	X ²
Level 2	0.00054***	90	270.10	0.00044***	90	253.89
Level 3	0.00021***	6	36.86	0.00018***	6	37.67

Note. † $p < .10$, * $p < .05$, ** $p < .01$, *** $p < .001$. Level 1 $n = 2115$, Level 2 $n = 111$, Level 3 $n = 7$. Alignment index re-scaled to 0 to 100.

Table C.3
HLM of Alignment to Standards and Assessments in ELAR Using 2007-2009 Data

Fixed Effect	Alignment to Standards			Alignment to Assessments		
	B	se	t ratio	Coefficient	se	t ratio
Level 2						
Intercept	24.457***	0.659	37.112	17.329***	1.124	15.415
Focus	-6.942***	0.498	-13.943	0.165	0.712	0.231
Stability	0.008	0.234	0.032	-0.507	0.341	-1.486
Consistency	-4.198	6.689	-0.628	17.178*	8.391	2.047
Power	0.322	0.316	1.020	0.494	0.488	1.013
Authority	0.097	0.240	0.405	-0.198	0.349	-0.566
Level 1						
Experience	-0.079**	0.025	-3.166	-0.053***	0.014	-3.795
Challenge	0.286	0.496	0.575	0.265	0.279	0.950
SBRinfluence	0.857**	0.239	3.588	0.456**	0.134	3.392
Otherinfluence	-0.202	0.342	-0.590	-0.639**	0.192	-3.326
Profcomm	0.176	0.242	0.727	0.054	0.136	0.398
SBRpd	0.007***	0.002	3.927	0.001	0.001	-1.412
Random Effect	Variance Component	df	X ²	Variance Component	df	X ²
Level 2	0.00024***	60	179.77	0.00119***	60	1031.03
Level 3	0	8	11.42	0.00001	8	11.62

Note. † p < .10, * p < .05, ** p < .01, *** p < .001. Level 1 n = 2278, Level 2 n = 82, Level 3 n = 9. Alignment index re-scaled to 0 to 100.

Table C.4
HLM of Alignment to Standards and Assessments in Science Using 2007-2009 Data

Fixed Effect	Alignment to Standards			Alignment to Assessments		
	B	se	t ratio	Coefficient	se	t ratio
Level 2						
Intercept	13.125***	1.559	8.420	11.645***	1.411	8.256
Focus	-5.573***	1.472	-3.786	0.696	1.331	0.523
Stability	0.358	0.264	1.352	0.428†	0.243	1.763
Consistency	8.526	7.428	1.148	73.608***	6.806	10.815
Power	0.191	0.437	0.437	0.720†	0.397	1.812
Authority	0.872†	0.515	1.693	1.792***	0.457	3.917
Level 1						
Experience	0.014	0.029	0.480	-0.007	0.026	-0.256
Challenge	1.276*	0.518	2.465	0.331	0.469	0.705
SBRinfluence	0.923**	0.270	3.424	0.830**	0.244	3.396
Otherinfluence	0.377	0.356	1.059	-0.008	0.323	-0.025
Profcomm	0.033	0.251	0.129	-0.137	0.227	-0.600
SBRpd	0.002	0.002	1.006	0.003†	0.002	1.924
Random Effect	Variance Component	df	X ²	Variance Component	df	X ²
Level 2	0.00005***	34	72.02	0.00005**	34	56.91
Level 3	0.00078***	7	304.64	0.00061***	7	890.84

Note. † $p < .10$, * $p < .05$, ** $p < .01$, *** $p < .001$. Level 1 $n = 724$, Level 2 $n = 54$, Level 3 $n = 8$. Alignment index re-scaled to 0 to 100.

Table C.5
HLM of Alignment to Standards and Assessments in Mathematics Using Re-Centered Data

Fixed Effect	Alignment to Standards			Alignment to Assessments		
	B	se	t ratio	Coefficient	se	t ratio
Level 2						
Intercept	20.453***	0.109	18.680	19.767***	0.671	29.441
Focus	-4.522***	0.504	-8.983	-1.602***	0.420	-3.816
Stability	0.198	0.165	1.195	0.574***	0.121	4.740
Consistency	14.598***	2.907	5.021	12.269***	2.509	4.890
Power	0.801*	0.366	2.187	-0.137	0.232	-0.589
Authority	-0.535	0.402	-1.330	-0.236	0.214	-1.106
Level 1						
Experience	-0.012	0.013	-0.958	-0.016	0.012	-1.324
Challenge	0.165	0.290	0.570	0.252	0.270	0.934
SBRinfluence	0.644***	0.132	4.865	0.666***	0.123	5.397
Otherinfluence	-0.670***	0.176	-3.812	-0.746***	0.164	-4.553
Profcomm	0.064	0.134	-0.477	-0.280*	0.125	-2.243
SBRpd	-0.004***	0.001	-3.758	-0.005***	0.001	-5.333
Random Effect	Variance Component	df	X ²	Variance Component	df	X ²
Level 2	0.00054***	190	342.44	0.00043***	190	544.53
Level 3	0.00057***	6	143.95	0.00011***	6	48.34

Note. † p < .10, * p < .05, ** p < .01, *** p < .001. Level 1 n = 4197, Level 2 n = 211, Level 3 n = 7. Alignment index re-scaled to 0 to 100.

Table C.6

HLM of Alignment to Standards and Assessments in ELAR Using Re-Centered Data

Fixed Effect	Alignment to Standards			Alignment to Assessments		
	B	se	t ratio	Coefficient	se	t ratio
Level 2						
Intercept	23.033***	1.019	22.603	18.007***	1.130	15.940
Focus	-6.619***	0.763	-8.681	-0.134	0.749	-0.179
Stability	-0.005	0.195	-0.026	-0.129	0.230	-0.559
Consistency	-1.383	5.048	-0.274	19.096**	6.898	2.768
Power	0.825**	0.280	2.950	0.607†	0.322	1.884
Authority	-0.213	0.405	-0.526	-0.225	0.292	-0.769
Level 1						
Experience	-0.053**	0.019	-2.784	-0.049***	0.012	-4.210
Challenge	0.094	0.391	0.239	0.196	0.237	0.826
SBRinfluence	0.700***	0.189	3.707	0.365**	0.114	3.193
Otherinfluence	0.364	0.270	0.135	-0.457**	0.163	-2.795
Profcomm	0.437*	0.188	2.328	0.021	0.114	0.182
SBRpd	0.004**	0.001	2.943	-0.001	0.001	-1.309
Random Effect	Variance Component	df	X ²	Variance Component	df	X ²
Level 2	0.00024**	104	148.97	0.00150***	104	713.15
Level 3	0.00053***	10	101.14	0.00014*	10	19.71

Note. † p < .10, * p < .05, ** p < .01, *** p < .001. Level 1 n = 3741, Level 2 n = 128, Level 3 n = 11. Alignment index re-scaled to 0 to 100.

Table C.7
HLM of Alignment to Standards and Assessments in Science Using Re-Centered Data

Fixed Effect	Alignment to Standards			Alignment to Assessments		
	B	se	t ratio	Coefficient	se	t ratio
Level 2						
Intercept	15.530***	1.805	8.602	11.409***	2.821	4.044
Focus	-3.549***	0.672	-5.285	0.340	1.225	0.278
Stability	-0.009	0.112	-0.765	-0.127	0.179	-0.712
Consistency	12.951**	4.360	2.970	58.289***	6.477	8.999
Power	0.450**	0.166	2.715	0.653*	0.284	2.298
Authority	0.366*	0.176	2.085	1.630***	0.336	4.857
Level 1						
Experience	-0.020	0.019	-1.028	-0.013	0.017	-0.795
Challenge	0.591	0.360	1.640	-0.115	0.316	-0.363
SBRinfluence	0.489**	0.179	2.735	0.545**	0.157	3.475
Otherinfluence	0.154	0.235	0.658	-0.171	0.206	-0.830
Profcomm	0.363*	0.166	2.182	0.236	0.146	1.620
SBRpd	0.004**	0.001	2.933	0.003**	0.001	2.831
Random Effect	Variance Component	df	X ²	Variance Component	df	X ²
Level 2	0.00013***	80	176.68	0.00054***	80	200.51
Level 3	0.00006**	9	26.35	0.00033***	9	45.53

Note. † p < .10, * p < .05, ** p < .01, *** p < .001. Level 1 n = 1549, Level 2 n = 103, Level 3 n = 10. Alignment index re-scaled to 0 to 100.

Bibliography

- Airasian, P. W. (1988). Symbolic validation: The case of state-mandated, high stakes testing. *Educational Evaluation and Policy Analysis, 10*(4), 301-313.
- American Federation of Teachers. (2001). *Making Standards Matter 2001: A Fifty-State Report on Efforts to Implement a Standards-Based System*. Washington, DC: Author.
- Betts, J. R., & Danenberg, A. (2002). School accountability in California: An early evaluation. In D. Ravitch (Ed.), *Brookings papers on educational policy 2001* (pp. 123-197). Washington, DC: Brookings Institution.
- Bidwell, C. (1965). The school as a formal organization. In J. G. March (Ed.), *Handbook of organizations* (Vol. 972-1019). New York: Rand McNally.
- Borko, H., & Elliott, R. (1998). *Tensions between competing pedagogical and accountability commitments for exemplary teachers of mathematics in Kentucky*. Los Angeles: Center for the Study of Evaluation.
- Borman, G. D., Hewes, G. M., Overman, L. T., & Brown, S. (2003). Comprehensive school reform and achievement: A meta-analysis. *Review of Education Research, 73*(2), 125-230.
- Boudett, K. P., City, E. A., & Murnane, R. J. (2005). *Data wise*. Cambridge, MA: Harvard Education Press.
- Carbonaro, W. (2005). Tracking, students' effort, and academic achievement. *Sociology of Education, 78*(1), 27-49.
- Carnoy, M., & Loeb, S. (2004). Does external accountability affect student outcomes? A cross-state analysis. In S. H. Fuhrman & R. F. Elmore (Eds.), *Redesigning accountability systems for education* (pp. 189-219). New York: Teachers College Press.
- Center on Education Policy. (2006). *From the capital to the classroom: Year 4 of the No Child Left Behind Act*. Washington, DC: Center on Education Policy.
- Center on Education Policy. (2009). *State high school exit exams: Trends in test programs, alternate pathways, and pass rates*. Washington, DC: Author.
- Charles, D. C. (2008). *Teachers' perspectives on the unintended consequences of high-stakes testing*. Louisiana State University, Baton Rouge, LA.

- Chubb, J. E., & Moe, T. M. (1990). *Politics, markets, and America's schools*. Washington, DC: Brookings Institution.
- Clarke, M., Shore, A., Rhoades, K., Abrams, L., Miao, J., & Li, J. (2003). *Perceived effects of state-mandated testing programs on teaching and learning: Findings from interviews with educators in low-, medium-, and high-stakes states*. Chestnut Hill, MA: National Board on Educational Testing and Public Policy.
- Clune, W. H. (1993). Systemic educational policy: A conceptual framework. In S. H. Fuhrman (Ed.), *Designing coherent educational policy* (pp. 125-140). San Francisco: 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, A. (1983). Comparing regression coefficients across subsamples: A study of the statistical test. *Sociological Methods & Research*, 12(1), 77-94.
- 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. (1990). A revolution in one classroom: The case of Mrs. Oublier. *Educational Evaluation and Policy Analysis*, 12(3), 311-329.
- Cohen, D. K., & Ball, D. L. (1990). Policy and practice: An overview. *Educational Evaluation and Policy Analysis*, 12(3), 233-239.
- Cohen, D. K., Raudenbush, S. W., & Ball, D. L. (2003). Resources, instruction, and research. *Educational Evaluation and Policy Analysis*, 25(2), 119-142.
- 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: Longman.
- Cuban, L. (1990). Reforming again, again, and again. *Educational Researcher*, 19(1), 3-13.

- Cueto, S., Ramirez, C., & Leon, J. (2006). Opportunities to learn and achievement in mathematics in a sample of sixth grade students in Lima, Peru. *Educational Studies in Mathematics*, 62(1), 25-55.
- Darling-Hammond, L. (1990). Instructional policy into practice: "The power of the bottom over the top". *Educational Evaluation and Policy Analysis*, 12(3), 339-347.
- de Mello, V. B., Blankenship, C., & McLaughlin, D. (2009). *Mapping state proficiency standards onto NAEP scales: 2005-2007*. Washington, DC: National Center for Education Statistics.
- Dee, T., & Jacob, B. A. (2009). *The impact of No Child Left Behind on student achievement*. Cambridge, MA: National Bureau of Economic Research.
- Desimone, L. M. (2002). How can comprehensive school reform models be successfully implemented? *Review of Education Research*, 72(3), 433-479.
- Desimone, L. M., Le Floch, K. C., Ansell, S., & Taylor, J. (2005). *Are we asking the right questions on national surveys about teacher's professional development? Insights from cognitive interviews with teachers*. Paper presented at the Annual Meeting of the American Educational Research Association.
- Desimone, L. M., Smith, T. M., Hayes, S. A., & Frisvold, D. (2004). Beyond accountability and average mathematics scores: Relating state education policy attributes to cognitive achievement domains. *Educational Measurement: Issues and Practice*, 24(4), 5-18.
- Desimone, L. M., Smith, T. M., & Phillips, K. J. R. (2007). Does policy influence mathematics and science teachers' participation in professional development. *Teachers College Record*, 109(5), 1086-1122.
- Education Week. (2002). *Quality counts 2002: Building blocks for success*. Bethesda, MD: Editorial Projects in Education.
- Education Week. (2003). *Quality counts 2003: If I can't learn from you*. Bethesda, MD: Editorial Projects in Education.
- Education Week. (2004). *Quality counts 2004: Count me in: Special education in an era of standards*. Bethesda, MD: Editorial Projects in Education.
- Education Week. (2005). *Quality counts 2005: No small change, targeting money toward student performance*. Bethesda, MD: Editorial Projects in Education.

- Education Week. (2006). *Quality counts 2006: Quality counts at 10: A decard of standards based education*. (Bethesda, MD), Editorial Projects in Education.
- Education Week. (2007). *Quality counts 2007: From cradle to career*. (Bethesda, MD), Editorial Projects in Education.
- Education Week. (2008). *Quality counts 2008: Tapping into teaching*. Bethesda, MD: Editorial Projects in Education.
- Education Week. (2009). *Quality counts 2009: Portrait of population*. Bethesda, MD: Editorial Projects in Education.
- Elmore, R. F. (1996). Getting to scale with successful educational practices. In S. H. Fuhrman & J. A. O'Day (Eds.), *Rewards and reform: Creating educational incentives that work* (pp. 294-329). San Francisco: Jossey-Bass.
- Elmore, R. F. (2000). *Building a new structure for school leadership*. Washington, DC: Albert Shanker Institute.
- Figlio, D. N., & Ladd, H. F. (2008). School accountability and student achievement. In H. F. Ladd & E. Fiske (Eds.), *Handbook of research in education finance and policy* (pp. 166-182). New York: Routledge.
- Finn, C. E., & Kanstoroom, M. (2001). State academic standards. In D. Ravitch (Ed.), *Brookings papers on education policy 2001* (pp. 131-179). Washington, DC: Brookings Institution.
- Finn, C. E., Petrilli, M. J., & Julian, L. (2006). *The State of State Standards 2006*. Washington, DC: Thomas B. Fordham Foundation.
- Firestone, W. A., Mayrowetz, D., & Fairman, J. (1998). Performance-based assessment and instructional change: The effects of testing in Maine and Maryland. *Educational Evaluation and Policy Analysis*, 20(2), 95-113.
- Floden, R. E., Porter, A. C., Schmidt, W. H., Freeman, D. J., & Schwille, J. R. (1980). Responses to curriculum pressures: A policy-capturing study of teacher decisions about content. *Journal of Educational Psychology*, 73, 129-141.
- Gamoran, A., & Nystrand, M. (1994). Tracking, instruction and achievement. *International Journal of Educational Research*, 21(4), 217-231.
- 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: McGraw-Hill.
- Grissmer, D., & Flanagan, A. (2001). Searching for indirect evidence of the effects of statewide reforms. In D. Ravitch (Ed.), *Brookings papers on educational policy 2001* (pp. 181-229). Washington, DC: Brookings Institution.
- Guthrie, J. W., Kirst, M. W., Hayward, G. C., Odden, A. R., Koppich, J. E., Adams, J. A., et al. (1989). *Conditions of education in California*. Berkeley, CA: Policy Analysis for California Education.
- Hagge, S. L., & Waltman, K. (2007). *Teachers' perceptions of NCLB: A multi-year study*. Paper presented at the Annual Meeting of the Iowa Educational Research and Evaluation Association.
- Hamilton, L. S., & Berends, M. (2006). *Instructional practices related to standards and assessments (RAND Working Paper WR-374-EDU)*. Santa Monica, CA: RAND.
- Hanushek, E. A., & Raymond, M. A. (2005). Does school accountability lead to improved student performance? *Journal of Policy Analysis and Management*, 24(2), 297-327.
- Hess, F. M. (2002). Reform, resistance, ... retreat? The predictable politics of accountability in Virginia. In D. Ravitch (Ed.), *Brookings papers on educational policy 2001* (pp. 69-122). Washington, DC: Brookings Institution.
- Hill, H. C. (2001). Policy is not enough: Language and the interpretation of state standards. *American Educational Research Journal*, 38(2), 289-318.
- Hoffer, T. B. (1994). Middle school ability grouping and student achievement in science and mathematics. *Educational Evaluation and Policy Analysis*, 14(3), 205-227.
- Holcomb, E. (1999). *Getting excited about data: How to combine people, passion, and proof*. Thousand Oaks, CA: Corwin Press.
- Jacob, B. A. (2005). Accountability, incentives, and behavior: The impact of high-stakes testing in the Chicago Public Schools. *Journal of Public Economics*, 89, 761-796.
- Jennings, J. H. (1998). *Why national standards and tests? Politics and the quest for better schools*. Thousand Oaks, CA: SAGE.
- Jones, M. G., Jones, B. D., Hardin, B., Chapman, L., Yarbrough, T., & Davis, M. (1999). The impact of high-stakes testing on teachers and students in North Carolina. *Phi Delta Kappan*, 81, 199-203.

- Karr, K. A., Marsh, J. A., Ikemoto, G. S., Darilek, H., & Barney, H. (2006). Strategies to promote data use for instructional improvement: Actions, outcomes, and lessons from three urban districts. *American Journal of Education*, 112(4), 496-520.
- Kliebard, H. M. (2004). *The struggle for the American curriculum: 1893-1958* (3rd ed.). New York: Taylor & Francis.
- 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.
- Koretz, D. M., Stecher, B. M., Klein, S., & McCaffrey, D. (1994). The Vermont Portfolio Assessment Program: Findings and implications. *Educational Measurement: Issues and Practice*, 13(3), 5-16.
- Lam, T. C. M., & Bengo, P. (2003). A comparison of three retrospective self-reporting methods of measuring change in instructional practice. *American Journal of Evaluation*, 24(1), 65-80.
- Lane, S., Stone, C. A., Parke, C. S., Hansen, M. A., & Cerrillo, T. L. (2000). *Consequential evidence for MSPAP from the teacher, principal, and student perspective*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.
- Little, J. W. (1990). The persistence of privacy: Autonomy and initiative in teachers' professional relations.
- Lortie, D. C. (1975). *Schoolteacher: A Sociological Study*. Chicago: The University of Chicago Press.
- Luna, C., & Turner, C. L. (2001). The impact of the MCAS: Teachers talk about high-stakes testing. *The English Journal*, 91(1), 79-87.
- Mayer, D. P. (1999). Measuring instructional practice: Can policymakers trust survey data? *Educational Evaluation and Policy Analysis*, 21(1), 29-45.
- McDonnell, L. M. (2004). *Politics, persuasion, and educational testing*. Cambridge, MA: Harvard University Press.

- McDonnell, L. M. (2005). Assessment and accountability from the policymaker's perspective. In J. L. Herman & E. H. Haertel (Eds.), *Uses and misuses of data for educational accountability and improvement* (pp. 35-54). Malden, MA: Blackwell.
- McDonnell, L. M., & Choisser, C. (1997). *Testing and teaching: Local implementation of new state assessments*. Los Angeles, CA: RAND.
- McLaughlin, M. W. (1990). The RAND Change Agent Study revisited: Macro perspectives and micro realities. *Educational Researcher*, 19(9), 11-16.
- McMillan, J. H., Myran, S., & Workman, D. (1999). *The impact of mandated statewide testing on teachers' classroom assessment and instructional practices*. Paper presented at the Annual Meeting of the American Educational Research Association.
- Meyer, J. W., & Rowan, B. (1980). The structure of educational organizations. In M. W. Meyer (Ed.), *Environments and organizations* (pp. 78-109). San Francisco, CA: Jossey-Bass.
- Minnesota Academic Standards Committee. (2003). *Minnesota Academic Standards: Language arts K-12*. St. Paul, MN: Minnesota Department of Education.
- National Commission on Excellence in Education. (1983). *A Nation at Risk: The Imperative for Education Reform*. Washington, DC: U.S. Government Printing Office.
- National Council of Teachers of Mathematics. (1989). *Curriculum and evaluation 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: Jossey-Bass.
- Ohio Department of Education. (2001). *Academic content standards: K-12 English language arts*. Columbus, OH: Author.
- Pedulla, J. J., Abrams, L. M., Madaus, G. F., Russell, M. K., Ramos, M. A., & Miao, J. (2003). *Perceived effects of state-mandated testing programs on teaching and learning: Findings from a national survey of teachers*. Chestnut Hill, MA: National Board on Educational Testing and Public Policy.

- Polikoff, M. S., Porter, A. C., & Smithson, J. (2009). The role of state student achievement tests in standards-based reform. University of Pennsylvania.
- Porter, A. C. (2002). 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. J. (2005). 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., Floden, R., Freeman, D., Schmidt, W., & Schwille, J. (1988). Content determinants in elementary school mathematics. In D. A. Grouws & T. J. Cooney (Eds.), *Perspectives on research on effective mathematical teaching*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- 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., Linn, R. L., & Trimble, C. S. (2005). The effects of state decisions about NCLB adequate yearly progress targets. *Educational Measurement: Issues and Practice*, 24(4), 32-39.
- Porter, A. C., & Polikoff, M. S. (2007). NCLB: State interpretations, early effects, and suggestions for reauthorization. *Social Policy Report*, 21(4).
- Porter, A. C., Polikoff, M. S., & Smithson, J. (2009). Is there a de facto national intended curriculum? Evidence from state content standards. *Educational Evaluation and Policy Analysis*, 31(3), 238-268.
- 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.
- Porter, A. C., Smithson, J., Blank, R., & Zeidner, J. (2007). Alignment as a teacher variable. *Applied Measurement in Education*, 20(1), 27-51.
- Raizen, S. A. (1988). *Increasing educational productivity through improving the science curriculum*. Washington, DC: OERI.
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage Publications.

- Ravitch, D. (1995). *National standards in American education: A citizen's guide*. Washington, DC: Brookings Institution.
- Reys, B. (Ed.). (2006). *The Intended Mathematics Curriculum as Represented in State-Level Curriculum Standards: Consensus or Confusion*. Charlotte, NC: Information Age Publishing.
- Reys, B., Chval, K., Dingman, S., McNaught, M., Regis, T. P., & Togashi, J. (2007). Grade-level learning expectations: A new challenge for elementary mathematics teachers. *Teaching Children Mathematics*, 14(1), 6-11.
- Ross, M. (1989). The relation of implicit theories to the construction of personal histories. *Psychological Review*, 96, 341-357.
- Ross, M., & Conway, M. (1986). Remembering one's own past: The construction of personal histories. In R. M. Sorrentino & E. T. Higgins (Eds.), *Handbook of motivation and cognition* (pp. 122-124). New York: Guilford.
- Rowan, B. (1996). Standards as incentives for instructional reform. In S. H. Fuhrman & J. A. O'Day (Eds.), *Rewards and reform: Creating educational incentives that work* (pp. 195-225). San Francisco: Jossey-Bass.
- Schmidt, W. H., McKnight, C. C., Houang, R. T., Wang, H., Wiley, D. E., Cogan, L. S., et al. (2001). *Why Schools Matter: A Cross-National Comparison of Curriculum and Learning*. San Francisco, CA: Jossey-Bass.
- Schulz, B. C. (2008). *Teachers' perspectives of how high-stakes testing influences instructional decisions and professionalism*. University of Georgia, Athens, GA.
- Schwartz, N., & Oyserman, D. (2001). Asking questions about behavior: Cognition, communication, and questionnaire construction. *American Journal of Evaluation*, 22(2), 127-160.
- Schwille, J., Porter, A. C., Alford, L., Floden, R., Freeman, D., Irwin, S., et al. (1988). State policy and the control of curriculum decisions. *Educational Policy*, 2(1), 29-50.
- Schwille, J., Porter, A. C., Belli, G., Floden, R., Freeman, D., Knappen, L., et al. (1982). *Teachers as policy brokers in the content of elementary school mathematics: Research series no. 113*. East Lansing, MI: Institute for Research on Teaching, Michigan State University.
- Schwille, J., Porter, A. C., & Gant, M. (1980). Content decision making and the politics of education. *Educational Administration Quarterly*, 16(2), 21-40.

- Sebring, P. A. (1987). Consequences of differential amounts of high school coursework: Will the new graduation requirements help? *Educational Evaluation and Policy Analysis*, 9(3), 257-273.
- Shepard, L. A., & Dougherty, K. C. (1991). *Effects of high-stakes testing on instruction*. Paper presented at the Annual Conference of the American Educational Research Association, Chicago, IL.
- Smith, M. L. (1991). Put to the test: The effects of external testing on teachers. *Educational Evaluation and Policy Analysis*, 20(2), 8-11.
- Smith, M. L. (1997). *Reforming schools by reforming assessment: Consequences of the Arizona Student Assessment Program (ASAP): Equity and teacher capacity-building*. Los Angeles, CA: CRESST.
- Smith, M. L., Edelsky, C., Draper, K., Rottenberg, C., & Cherland, M. (1990). *The role of testing in elementary schools*. Los Angeles: Center for Research on Evaluation, Standards, and Student Testing.
- Smith, M. L., & Rottenberg, C. (1991). Unintended consequences of external testing in elementary schools. *Educational Measurement: Issues and Practice*, 10(4), 7-11.
- 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.
- Smithson, J. L., & Collares, A. C. (2007). *Alignment as a predictor of student achievement gains*. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL.
- Snyder, T. D., Dillow, S. A., & Hoffman, C. M. (2008). *Digest of education statistics, 2007* (No. NCES 2008-022). Washington, DC: National Center for Education Statistics.
- Spady, W. G., & Mitchell, D. E. (1979). Authority and the management of classroom activities. In D. L. Duke (Ed.), *Classroom management: The seventy-eighth yearbook of the National Society for the Study of Education* (pp. 75-115). Chicago, IL: University of Chicago Press.
- Spillane, J. P. (2004). *Standards deviation: How schools misunderstand education policy*. Cambridge, MA: Harvard University Press.

- Stecher, B. M., & Barron, S. I. (2001). Unintended consequences of test-based accountability when testing in 'milepost' grades. *Educational Assessment, 7*(4), 259-281.
- Stecher, B. M., Barron, S. I., Chun, T., & Ross, K. (2000). *The effects of the Washington state education reform on schools and classrooms*. Los Angeles, CA: CRESST.
- Stecher, B. M., & Chun, T. (2001). *School and classroom practices during two years of education reform in Washington state*. Los Angeles, CA: CRESST.
- Stecher, B. M., & Naftel, S. (2006). *Implementing standards-based accountability: Study design, state context, and accountability policies*. Santa Monica, CA: RAND.
- Sullivan, G. P. (2006). *The impact of high stakes testing on curriculum, teaching, and learning*. Unpublished doctoral dissertation, Virginia Tech, Blacksburg, VA.
- Supovitz, J. (2009). Can high stakes testing leverage educational improvement? Prospects from the last decade of testing and accountability reform. *Journal of Educational Change, 10*, 211-227.
- Sykes, G. (1990). Organizing policy into practice: Reactions to the cases. *Educational Evaluation and Policy Analysis, 12*(3), 349-353.
- Taylor, G., Shepard, L. A., Kinner, F., & Rosenthal, J. (2002). *A survey of teachers' perspectives on high-stakes testing in Colorado: What gets taught, what gets lost*. Los Angeles, CA: Center for the Study of Evaluation.
- Tyack, D. B., & Cuban, L. (1995). *Tinkering toward utopia: A century of public school reform*. Cambridge, MA: Harvard University Press.
- Wayman, J. C., & Stringfield, S. (2006). Technology-supported involvement of entire faculties in examination of student data for instructional improvement. *American Journal of Education, 112*(4), 549-571.
- Weick, K. E. (1976). Educational organizations as loosely coupled systems. *Administrative Science Quarterly, 21*(1), 1-19.
- Wong, K. K., Anagnostopoulos, D., Rutledge, S., & Edwards, C. (2003). The challenge of improving instruction in urban high schools: Case studies of the implementation of the Chicago Academic Standards. *Peabody Journal of Education, 78*(3), 39-87.
- Ysseldyke, J., Nelson, J. R., Christenson, S., Johnson, D. R., Dennison, A., Triezenberg, H., et al. (2004). What we know and need to know about the consequences of

high-stakes testing for students with disabilities. *Exceptional Children*, 71(1), 75-94.