

The Potential Effects of “High-Quality and Uniform” Standards:
Lessons from a Synthesis of Previous Research and Proposals for a
New Research Agenda

A Final Report to the National Research Council

Douglas N. Harris
Margaret Goertz

August 4, 2008

Introduction: The Emergence of the Standards Movement

The history of the American education system is one characterized by a gradual, and recently accelerating, shift in governance from local school districts to state and federal governments. As part of the larger political focus of the 1960s and 1970s on achieving equity across racial, income, and other groups, policies such as desegregation, school finance, and special education were arguably the earliest parts of the system whose control became more centralized. The economic stagnation of the 1970s and perceived limitations in labor force skills, however, led many to question the successes of the education system and to a swing in the policy pendulum from equity to excellence (Harris & Herrington, 2006). The fact that the vast majority of the labor force at the time was composed of workers who graduated from high school during the 1940s and 1950s meant that the 1970s school reforms could not have been the cause of the economic stagnation (Harris, Handel, & Mishel, 2004). Yet, the social upheaval of the 1970s made schools an obvious target as well as a possible lever for lifting the economy from its slump. The economic part of this argument was made most poignantly in *A Nation at Risk* (1983), which argued not only that schools could and should be improved, but implied that the schools were the main source of the country's economic problems and apparent failure to win the Cold War. Twenty-five years later, the actors in the play have changed—China has replaced the Soviet Union as a military threat and Japan as an economic one—but the script is largely unchanged. Our schools are failing. This is making the country economically uncompetitive. The country needs to do something about it.

A Nation at Risk (1983) was more than a critique, however. It included a call for higher academic standards, a recommendation aggressively followed by policymakers and educators. Darling-Hammond and Berry (1988) estimate that over 700 new policies were enacted by states between 1983 and 1985 alone. There were moves toward standards-based reform in earlier decades (and centuries) (Goertz, 2007), but *A Nation at Risk* marked a turning point both in the level of attention paid to the issue and the perception that this was a national problem that might require a national—and perhaps federal—solution.

The state reforms triggered by *A Nation at Risk* raised course work standards for high school graduation, implemented or expanded minimum competency testing programs, and initiated testing of aspiring teachers (Goertz, 1986). These policies induced students to take somewhat more rigorous course work (Harris & Herrington, 2006) but did little otherwise to change the content of instruction (especially its focus on basic skills), or to alter the reigning notions of teaching and learning (Cohen, 1990). The standards-based reform movement, which emerged in the late 1980's and early 1990's, was designed to address the shortcomings of input-driven education reforms. Under the theory of standards-based reform, states establish challenging content and performance standards for all students and align key state policies affecting teaching and learning—curriculum and curriculum materials, preservice and inservice teacher training, and assessment—to these standards. Then, states give schools and school districts greater flexibility to design appropriate instructional programs in exchange for holding schools accountable for student performance (Smith and O'Day, 1991). Along with the emphasis on higher standards for all students is the use of assessments to inform instruction and

outcome-based accountability systems to create incentives for students and schools to improve.

The ideas of standards-based reform were incorporated into 1994 reauthorization of the federal Elementary and Secondary Education Act (ESEA) which required states to develop standards and assessments, and imposed sanctions if students were performing poorly. All states were required to test students in reading/language arts and mathematics at the elementary, middle and high school levels, as well as to report disaggregated results for student subgroups. But states were given considerable leeway in how they met these requirements. Without strong federal direction, there emerged a predictably wide range of state models. While all states developed assessments, standards, performance reporting, and in most cases consequences for performance, states found different ways to define what it meant for schools to succeed, what indicators to include in their definition of success, and what the consequences for failure would be (Goertz and Duffy, 2001). States varied in the coverage, rigor, specificity and clarity of their standards (Cross, Rebarber & Torres, 2004; Rothman, 2004) and on the grades they tested, as well as the type of assessments they used and the extent to which they were aligned with state standards. State sanctions ranged from public scrutiny to state takeover for schools (Goertz & Duffy, 2001). As of 2002, only 19 states were fully compliant with federal assessment provisions and many states had obtained waivers (Taylor, 2002).

Outside the original ESEA in 1965, arguably the greatest shift in federal role occurred in the 2002 ESEA re-authorization, called the *No Child Left Behind Act* (NCLB). By requiring states to significantly increase the amount of standardized testing, and by enforcing the law more stringently than it had been subsequently, the federal role

was both substantially deepened and broadened. However, responsibility for setting standards remains with the states, and variability in the quality of state standards has been highlighted as the differences across states in the percentages of students reaching state and NAEP proficiency performance standards have become evident. But the controversy over the federal role has created resistance to further federal involvement in making state policies any more uniform and addressing inequities in how schools are evaluated and treated in different states.

For this and other reasons, there is now considerable interest in creating “high-quality and uniform standards” for all states. The purpose of this paper is to summarize evidence to date about the effects of standards on student achievement and mediating factors, notably curriculum and instruction, that might inform the potential impact of such a policy. Policy debates on the topic have tended to highlight studies that, by themselves, may be of limited usefulness in understanding the potential impacts of uniform standards in the U.S.. In cross-national comparisons, for example—studies that have been the basis of much of the advocacy for national standards—countries differ in many fundamental ways, making it difficult to isolate the impact of a specific policy such as standards or to generalize these impacts across contexts. However, by synthesizing these individually imperfect studies, it is possible to draw some important conclusions. Other studies have attempted to synthesize research on standards-based reforms, but this is a somewhat broader topic. We have deliberately selected studies that are particularly useful for understanding the potential impact of uniform, high-quality standards. Our synthesis also pays closer attention to issues of research methodology.

To help structure the review and discussion, we continue in the second section below by describing what we mean by “high-quality and uniform” standards. In the third section, we outline a framework that outlines the causal chain and a related theory of action regarding how standards might improve student learning. Establishing the causal effects of standards on achievement and other student outcomes is a difficult task and one reason is, as we show, the “causal chain” between standards and student learning is somewhat complex. In the fourth section, we discuss a central problem in studying standards and educational policy more generally—that the policies are adopted through political processes and within economic and social contexts that make it difficult to isolate the impacts of policies from other differences across states (and countries) that also affect student outcomes. This makes it difficult to determine whether the policies have causal impacts and whether these impacts would likely generalize to other contexts.

While there are difficulties of studying the effects of standards, as there are in essentially all other educational topics, much rigorous research has been conducted and it is essential that we learn from it in considering where standards policies should go in the future. The fifth section uses the framework and theory of action as a basis for reviewing evidence about the effects of standards on student achievement, as well as about the links in the causal chain that produce those effects. This evidence focuses on variation between the U.S. and other countries, between U.S. states, as well as a variety of state and district case studies.

In the final section, we try to use this discussion of evidence to address the principal question: What would be the effects of “high-quality and uniform” standards on student achievement? While this paper pushes the debate forward on this question, there

are inherent difficulties of studying standards and other limits in existing research studies that make our answer to this question far from definitive.

Defining the Terrain: What are “High-Quality, Uniform Content Standards”?

The term “standards” means different things to different people and this complicates any discussion of standards-based reform. As we will see below, the discussion gets even murkier when considering the possibility of standards that are “high-quality and uniform.”

At the broadest level of discussion, we identify standards as they are applied to the three main components of the educational system—inputs, processes and outcomes. As defined by Shavelson, McDonnell and Oakes (1989), inputs are the human, financial and structural resources available to education; processes encompass what is taught (curriculum) and how (instruction); and outcomes are the consequences of the system for students. Education has long had input standards, such as those regarding the training and licensure of teachers (to ensure a minimum level of teacher quality) and the number of days in the school year, and are still used to monitor and accredit schools and school districts. In the current reform context, outcome standards establish goals for what students should know and be able to do at different points in their educational careers. The current argument for outcome standards is based on the notion that input standards were insufficient to ensure that all students had a quality education.¹

Standards for student performance are typically considered outcome standards and are composed of *content standards* and *performance standards* (NRC, 1999). States are required to establish both under NCLB. *Content standards* are broad descriptions of

knowledge and skills that students should acquire and be able to do in a particular subject area. They indicate the topics and skills that should be taught at various grades or grade spans and are intended to guide public school instruction, curriculum, teacher preparation and development and assessment. Although content standards should provide a map for the development of curriculum at the school or district level, some teachers have called for increasing specificity in content standards to drive their instruction (Hamilton et al., 2007; Kannapel et al., 2001; Public Agenda, 2006; Stigler and Hiebert, 1999). As noted earlier, states differ in the coverage, rigor, specificity and clarity of their content standards and frameworks.

The call for higher standards made in *A Nation at Risk* also resulted in *course requirements*. These are in some ways conceptually similar to content standards because both specify what academic material students should experience. Course requirements, however, straddle the boundary between outcome standards and process/output standards. For this reason, and because the current policy debate is really about content standards, we do not include course requirements in the category of content standards.

Performance standards, which often go hand-in-hand with content standards in practice, are somewhat different because they provide explicit definitions and examples of what students must demonstrate in order to show they have mastered the content standards. Performance standards answer the important question, how good is good enough? Although performance standards were intended to consist of exemplars of student work (Elmore & Rothman, 1999), as a practical matter, performance standards are expressed in the form of “cut scores” and each range of scores is given a name, e.g., “proficient.” These categories in turn define the levels of performance on which

accountability sanctions are built. Students failing to reach a given performance standard might have to repeat a grade level or take summer school. Likewise, the schools whose students do not meet the necessary standards might be required to re-constitute their staffs or, if they perform well, might be rewarded with financial bonuses. While performance standards and interconnected with content standards—allowing incentives to be attached to particular levels of academic performance—our focus again is on content standards. There is little evidence on performance standards per se and nearly impossible to study performance standards without also considering the accountability attached to those standards, and evidence on this broader topic has been reviewed by many others, including Harris (2007), Harris and Herrington (2006), Jacob (2005), and Rowan (2005).

But what does it mean for those content standards to be “high-quality and uniform”? Answering the quality portion of this question creates something of a dilemma as a basic premise of this paper is that we do not yet know how standards affect student outcomes, or at least we have not yet synthesized existing research in a way that would yield a convincing answer. If we knew that a certain type of standards influenced student outcomes then we could define that type as “high-quality.” This is putting the cart before the horse and, in this paper, we simply infer the definition of high-quality from what the advocates of high-quality standards seem to be arguing for. Smith and O’Day (1991), for example, call for standards that prize exploration and production of knowledge, rigor in thinking and sustained intellectual effort. Others such as the organization Achieve, which advocates more rigorous standards, turn to international standards as their benchmark.

Arguably the nation's strongest advocate of stronger standards has been Chester Finn of the Thomas B. Fordham Foundation, which has been publishing state-by-state grades of state standards for more than a decade. In their most recent report, Finn, Petrilli & Julian (2006a) write that their three criteria are: "clear, rigorous, and right-headed about content" (p.6), but have their own conception of what is rigorous and "right-headed" content. In discussing the standards of California, one of the states to which they give the highest grades, they use adjectives such as "clear," "specific," and "measurable" (p.1). We define high-quality standards as those that are clear, specific, and rigorous, meaning that the average teacher, after reading the standards, would understand what specific content they should be teaching and develop corresponding lessons plans that, if carried out well, would give students the opportunity to learn advanced and challenging academic content. This description, while somewhat vague, allows us to distinguish certain key differences among different types of content that standards might contain. For example, as we discuss later, one test used to compare achievement across countries focuses less on academic skills than on the application of those skills to real-world problems, while another assesses students' knowledge of specific academic topics.

By "uniform," we mean that the same standards are universally adopted by a group of states, or all 50 states. This does not necessarily mean that the standards are developed and/or required by the federal government. Indeed, some states have already begun to band together to create standards and common assessments across states, such as the New England Common Assessment Program (NECAP). Many states have chosen voluntarily to adopt content standards in line with professional organizations such as the

National Council for Teachers of Mathematics (NCTM). While truly universal adoption of any given set of standards would almost certainly require some type of federal requirement, we leave this important issue of policy adoption for other studies to pursue. We also make a clear distinction between the uniformity of standards and the uniformity of the enacted curriculum and instruction.

In short, this paper is about predicting the possible impact of high-quality and uniform content standards, as defined above, based on evidence about the implementation of standards in specific school districts, states and other countries in the past. Below, we take another step toward the review of evidence by outlining a framework which highlights both a way of identifying important research studies and of explaining the theory of action underlying the arguments for high-quality and uniform standards.

Analytic Framework and Theory of Action

For nearly two decades, researchers have sought to understand the impact of standards and standards-based reforms on teaching, learning and student achievement. The theory of action underlying standards-based reform is that rigorous standards will provide a coherent direction for education reform throughout the system. Key state or national policies affecting teaching and learning—curriculum and curriculum materials, preservice and inservice teacher training, and assessment, must be aligned with standards to provide a coherent system of instructional guidance (Smith and O’Day, 1991). In response to the development of standards in mathematics (NCTM, 1989, 2000), science (NRC, 1996) and technology education (ITEA, 2000), the National Research Council called for the development of a framework “that can be used to understand the influence

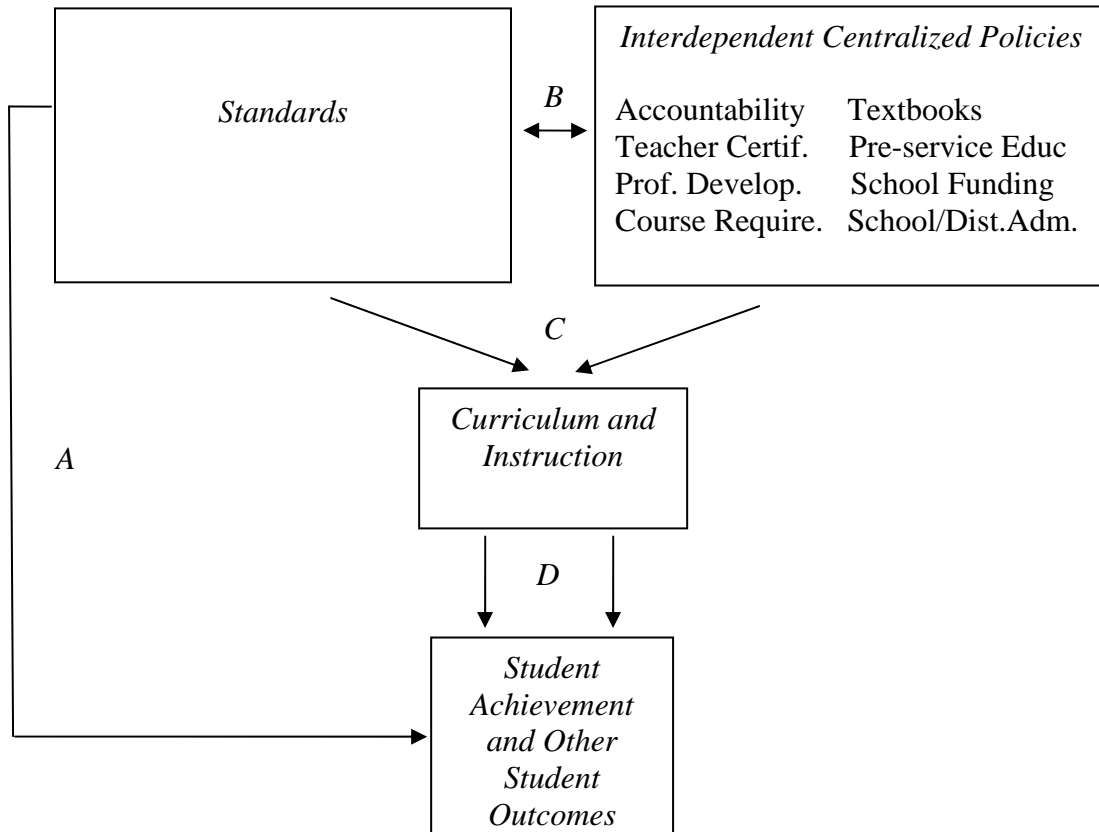
of science, mathematics and technology education standards on programs, policies and practices” in K-12 education (NRC, 2002). This is one of many frameworks used by researchers, but it provides a good starting point for our review. The NRC report recognized the important interconnection between standards and other forms of (mostly centralized) policies and the fact that the effects of policies play out through their impact on what happens in schools. Specifically, the report outlines three “channels of influence” or ways that standards impact student learning: curriculum, teacher development, and assessment and accountability. Standards are assumed to influence these three areas, in turn affecting instruction and, ultimately, student learning.

Our framework is outlined below in Figure 1. Four letters in Figure 1 indicate the main relationships of interest. Ultimately, we are interested in the impact of standards on student outcomes. To fully understand that impact, we must also consider the theory of action, including the interconnections between standards and other policies (Arrow B), the impact of standards (moderated by other policies) on the enacted curriculum and instruction (Arrow C), and the impact of standards-based curricula and instruction on student outcomes, especially student achievement (Arrow D).

This framework serves two main functions: first, to provide a logical and theoretical basis for studying effects of standards and, second, for categorizing the different types of studies that we review. For example, Arrow A indicates a direct path between standards and student outcomes. This is not based on a logical model of course—standards must impact curriculum and instruction if they are to change achievement—but Arrow A does represent an important category of analytic techniques

for understanding the effects of standards. This arrow reflects the “net effect” of standards on achievement, ignoring the ways in which the effects arise.

Figure 1: Analytic Framework for Effects of Standards



The main distinction between the “school-level practices” and “inter-dependent centralized policies” is the location at which, and therefore the way in which, action occurs. Our assumption is that centralized policies will be enacted at the national or large regional or state level. There is some ambiguity about where school districts fall in this categorization as they may make centralized decisions about, for example, curriculum, textbooks and professional development, or they may delegate some of these

responsibilities to schools. Also, we place school administration in the “centralized policy” category because we view administration as a moderator of student outcomes in the same sense as centralized policies, and administration is centralized relative to teaching.

Potentially Positive Effects of Standards

The framework above is useful partly because it provides a basis for discussing the possible theories of action through which standards might impact student learning. In particular, all four of the impacts represented by the Arrows A-D could be positive or negative. Below, we begin by presenting the arguments that have been made in support of more uniform and rigorous standards and discuss the potential positive effects of standards. The term “potential” in this case has two meanings: (a) that standards may potentially create the stated effects; and (b) that the effects are potentially positive in terms of student learning. Thus, even where standards affect instruction, they might not influence learning as theorized or in a way that is likely to improve student outcomes.

1. *To increase challenging content at all grade levels.* When advocates of standards argue for greater rigor, they appear to mean that standards students should learn more advanced academic content, such as trigonometry and statistics. To the degree that learning complex depends on having learned other content, this means that learning challenging content in early grades facilitates learning of challenging content in later grades.

2. *Reducing overlap in curricula across grades.* A specific factor that might limit exposure to challenging content is that specific academic topics are covered in the U.S. in

multiple grade levels, e.g., students in 3rd grade may study addition of single-digit numbers and students in 4th grade may review single-digit addition and then build on it with two- and three-digit numbers. Some overlap may be considered positive so long as the material is being covered with increasing depth and complexity, such as “spiraling.” However, critics argue that this approach prevents students from reaching more challenging content and that eliminating the redundancy across grades will allow schools to teach more challenging content (Coral & Schmitt, 1999).

3. Reducing variation in enacted curricula across classrooms within grades.

There is no debate that the students in any given grade in the U.S. experience very different curricula. A related problem is that there is considerable mobility of students across schools so that, in the absence of standards, mobile students will experience incoherent curricula. Uniform content standards could reduce this variation by inducing teachers across schools, districts, and even states to adopt similar curricula. Indeed the existence of national standards in many countries outside the U.S. may account for Stigler and Hiebert’s (1999) finding that there is more variation between countries than within them.

4. Facilitating coordination with related elements of educational policy (e.g., accountability). Education can be viewed as a complex system with interconnected parts, many of which are shown in Figure 1. The variation in curricula described in the first two points above make it difficult, for example, to create targeted training opportunities for teachers. One particular area where such coordination might be useful is teacher professional development (PD) related both to the content of the standards and ways of improving curriculum and instruction. If content were more standardized, high-quality

PD could be developed and provided to much larger numbers of teachers. Because the fixed costs of creating PD can be spread out over more teachers, this would yield better PD at lower costs than the current uncoordinated model allows.² The same argument applies to university-based teacher education.

5. *Reducing inequity in curricula across racial and income groups.* Curricula and instruction vary not only across grades and classrooms, but systematically according to the types of students in classrooms. A vast body of research shows that minority students and those from low-income backgrounds experience less rigorous curricula and shallow instruction (Gamoran, 1986; Oakes, 1985; Ogbu, 2003).

6. *Focusing instruction on concepts over procedures and definitions.* Evidence from the TIMSS video studies suggest that teachers in the U.S. teach students procedures for solving mathematical problems, in contrast to teachers in Japan and Germany, who focus on conceptual understanding (Stigler & Hiebert, 1999).³ American teachers include more than twice as many definitions of terms in their classes as Japanese and German teachers (Stigler & Hiebert, 1999, p.58), and none of the sampled American classrooms included mathematical proofs, compared with 10 and 53 percent of German and Japanese classes, respectively (Stigler & Hiebert, 1999, p.59).⁴ The instruction in roughly one-quarter of U.S. classes included both stating and developing concepts compared with three-quarters of Japanese and German classrooms. However, because standards are directed more at the curriculum than instruction, standards may have relatively little impact on the conceptual focus of instruction.

The above six arguments have been the most common over the past two decades. More recently, NCLB's accountability requirements have shed new light on the

considerable variation in state standards and assessments. Many are concerned about this variation and argue that uniform standards would eventually make it possible to have uniform assessments and performance standards and therefore facilitate legitimate comparisons of student achievement across states.

Potential Negative Effects

While our review of evidence below focuses on the potential positive effects of standards, it is important to acknowledge these potential negative consequences, against which any benefits eventually must be weighed.

1. Reducing productive experimentation. If standards were federally imposed, then it would restrict experimentation by states. This is especially important given the lack of evidence that any one set of standards is likely to be better than others.

2. Narrowing the curriculum. If certain content is required (and gets tested), then this will take away from non-required material; this is particularly problematic for content such as social studies, art and physical education that are frequently untested and, even then rarely included on equal par with math and reading in accountability systems.

3. Driving out effective teachers. Standards may further de-professionalize teaching (and administration) by taking control of content, and perhaps pedagogy out of the hands of teachers. This may drive out of the profession the teachers who are most effective in raising student achievement.

4. *Reducing student engagement and other non-achievement outcomes.* Further standardizing the curriculum may result in more scripted curricula and lower student engagement, resulting in higher drop-out rates and reduced motivation.

Again, we do not attend as closely to these negative side effects. With its focus on positive effects, this paper is really aimed at a “proof of concept” that high-quality and uniform standards could improve student learning. If there are no positive effects, then the standards cannot be justified even in the absence of negative consequences. We also discuss evidence later that instead of positive and negative effects, standards may have no impact on instruction and achievement.

Introduction to the Methodological Issues

When researchers try to estimate the causal impacts of any “treatment,” there is broad agreement that the random assignment experiment is the preferred approach. Random assignment allows us to reasonably assume that the control and treatment groups differ only in the fact that the latter group experienced the treatment. In medicine, for example, studies frequently test the effects of new medications by randomly assigning patients to the medication (treatment) or a placebo (control). In the absence of random assignment, researchers become concerned that participants who receive treatments are different from others in important and perhaps unmeasured ways that may influence their outcomes—a problem referred to as “selection bias.” The USDOE Institute for Education Sciences has recently made a strong push toward increasing the use of random assignment in educational research. As critics rightly point out, random assignment experiments are not perfect—for example, they often require creating unrealistic

conditions which call into question the generalizability of the findings—but they are still generally better than the alternatives (Shadish, Cook, & Campbell, 2002).

In addition to being imperfect, some types of treatments simply do not lend themselves to random assignment and, in these cases, “quasi-experimental” methods may be used. A quasi-experiment is a general category of studies in which the differences between control and treatment groups are still partly taken into account, but through means other than random assignment. In this case, the most plausible approach is to observe curriculum, instruction, and/or student achievement before a change in standards and then again at various points after the change in standards. This approach also requires trying to identify other possible changes (e.g., changes in policies such as accountability) that might have occurred around the same time that might also explain the before-and-after differences. Finally, to ensure that the changes in outcomes were not due to pre-existing trends in outcomes, it is important to measure outcomes for several periods before and several periods after the policy is implemented.

Even in a nearly ideal quasi-experimental study, it is still possible that the government (school district, state, etc.) that changed the standards did so because the circumstances were ripe for such a change. Conversely, other districts might not adopt the standards because their circumstances are less conducive. The point is that, even if the impact in a particular study location is definitely positive, the impacts may not generalize to other settings where standards might be imposed by a state or federal government. In particular, there are reasons to expect that the effects at scale (e.g., the effects of nationally mandated standards) would be smaller than in places that adopted the policies on their own accord.⁵

Educational policy is just such a case where quasi-experimental methods are not only useful, but arguably necessary. State and national governments choose policies through political and legislative processes which, in turn, reflect social, cultural, and economic conditions, in addition to previous decisions about related forms of policy. It is generally unreasonable to assume that these processes are unrelated to other important aspects of the educational system. Carnoy and Loeb (2003) find, for instance, that states with higher 4th grade test scores later adopted less aggressive accountability systems, presumably because their relative success on student test scores led them to see a smaller need for such policies. Further, it is likely that the effects of standards on achievement vary based on the initial level of achievement in the state. This makes it difficult to determine whether the effects would also arise in other states which face different conditions.

While educational policies by their nature cannot be randomly assigned to participants (nations, states, school districts, schools), there are some circumstances where random assignment experiments of related types of interventions might be useful. For example, the Tennessee STAR class size reduction initiative involved a sample of schools that volunteered to participate in a study. It was not a “policy” per se because it applied only to a small portion of schools, but the results of this study clearly do inform debates about the costs and benefits of class size policies. As it turns out, the general scarcity of random assignment experiments in education means that there is little such indirect evidence to draw on in the review of evidence on standards, but this discussion highlights a potential avenue for future research that we return to later.

There are three other methodological issues that apply specifically to the study of standards-based reform. First, standards are, almost by definition, centralized so that only small numbers of “participants”—that is, states—can be studied. This is important because one of the basic rules of research is that having more participants is always better. One reason is that having more examples provides greater confidence that the impact is not a fluke (i.e., it provides greater statistical power). Also, it allows the researcher to find comparisons between participants that are more similar than they would be with a small sample.

A second issue is that the effects of standards-based reforms, as an example of “systemic reform” (Smith and O’Day, 2001), are likely to be delayed. One principal actor in the development of standards writes that the effects of standards might not be visible for a decade or longer (Collins, 1997). Thus, even, if high-quality and uniform standards were put in place, and even if there really were an impact, it might not show up as changes in student outcomes until many years later. The key point here is that identifying delayed effects is more difficult, especially if the researcher is relying on a before-and-after approach. The greater the gap in time between the policy implementation and policy effect, the more likely it is that intervening factors will come into to play that muddle the measured effect. Some examples are given below that highlight this point.

A third methodological issue is that because content standards define the achievement that is of interest, the measured “impacts” on student achievement can be difficult to interpret (Harris & Herrington, 2006). For example, suppose that teachers do change their instruction and enacted curricula as a result of standards, but that the

achievement test is not aligned with the standards or therefore with the changes teachers make. In this case, even if the “true” impact were large and positive, the measured impact would probably be small, or even negative. Conversely, suppose that the new standards simply serve to make the curriculum align better with the test. This might be seen as a success in itself, but it also seems self-evident that spending more time teaching the material on the test will result in higher test scores. This is a fundamental problem in understanding the impacts of curriculum-based policies, such as standards, on student test scores. Some have proposed solving this problem with “audit tests” to confirm whether impacts on state assessments show up on other tests. In fact, research on accountability suggests that the impacts on high-stakes tests do not show up in low-stakes tests. This is hardly surprising as the low-stakes tests are, almost by definition, less well aligned with the curriculum. We would not expect impacts of content standards on tests that do not measure what is being tested. For a test to truly serve as an audit, it must cover exactly the same content with different types of questions, which rarely if ever occurs in practice.

We do not intend to leave the impression that using empirical evidence to learn about the potential impact of standards is an impossible feat, but rather to highlight the complexities and help identify ways to address them. For this reason, we shy away from drawing firm conclusions about any individual study, or even whole categories of studies discussed below, and instead try to paint a coherent picture from the overall collection of findings.

Review of Previous Research

One of the main reasons that the U.S. is considering implementing common standards is the perception that the country is falling behind other countries economically and educationally. The nation's relatively low ranking in international test scores comparisons is a particular point of concern of the U.S. system. As we will see below, some of the same data used to show that U.S. students are behind those in other countries has also been used to try to determine why these differences arise. Similar evidence exists regarding differences in student achievement across U.S. states. Other research expands the picture by focusing on state-level standards-based reforms and some specific school districts provide intensive case studies.

International Evidence (TIMSS)

The Third International Math and Science Study (TIMSS) is a potentially useful form of evidence because it includes data from many countries, including fairly rich survey data regarding the degree of centralization and a wide variety of other institutional characteristics about who controls what educational decisions and about the nature of the educational marketplace. Two researchers, the psychometrician William Schmidt and economist Ludger Woessman, have led separate research projects on the effects of educational policies on student achievement using the TIMSS data. Of particular interest are the TIMSS measures of the centralization of decisions about curriculum content and goals. Seventy-five percent of the countries in TIMSS controlled curricular goals at the national level—the U.S. being one of the noteworthy exceptions (Schmidt et al., 2001).

Note that the TIMSS does not track individual students over time and is not designed to calculate scale scores. In this respect, the achievement scores are not ideal for measuring changes in achievement over time; however, because nationally representative samples are tested at different ages, it is possible to identify the trajectory of learning as students progress through each national system, and to use this trajectory to examine policy effectiveness.

Research by Schmidt et al. (2001). Schmidt et al. (2001) use structural equation modeling to examine the relationship between content standards and textbook coverage (comprising the “intended curriculum”) and the relationship between each of these variables and teacher coverage of material (the “enacted curriculum”). Finally, they consider how the enacted curriculum influences achievement gain between seventh and eighth grade in math and science.⁶

It is important to point out that the SEM statistical methods used by Schmidt et al., while certainly more sophisticated than simple correlations, are not designed to identify causal impacts. Schmidt et al. do not control for any student demographics in the within-country analysis, which are typically included to at least partly address the selection bias problem, and it is possible that the choice of textbooks is based on student demographics and other factors that are also correlated with achievement gains, so that failing to control for these factors might lead to a misleading picture about the causal impacts. Below, we briefly summarize their findings when comparing whole countries and then turn to a more extensive analysis of variation within the U.S.

Looking across countries, the following results emerge from the Schmidt et al. analysis (2001, pp.170-171): (1) positive correlations between the space devoted to

particular topics in content standards and the content covered in textbooks in mathematics, but not science⁷ (see Arrow B in Figure 1); (2) positive correlations between academic standards and instructional time in math, but not science⁸ (see Arrow C in Figure 1); and (3) positive correlations between textbook coverage and instructional time in both subjects (see Arrow C in Figure 1). Interestingly, however, the results for mathematics are not completely corroborated when analyzing variation in content standards, textbook coverage, and instructional time within the U.S. The authors do not indicate why such differences emerge, or why the correlations are weaker in science.

Schmitt et al. also carried out analysis of variation within certain countries; however, in the U.S., they did not have data measuring the variation in content standards. The only relationship they could study using variation within the U.S. was that between textbook coverage and instructional time, which they find to be positive, just as it is in the typical country (see Arrow C in Figure 1).⁹ Interestingly, the magnitude of the relationship is somewhat smaller in the U.S. compared with other countries. This may be because U.S. textbooks are more comprehensive in their topic coverage so that teachers are forced to choose particular topics and ignore others. In a similar vein, Schmidt et al. report that the degree of consistency between academic standards and textbook coverage is highest in those countries that have more centralized control over the curriculum. This reinforces the importance of the interrelated policies shown in Figure 1. Again, standards probably do not by themselves have much direct impact on instruction—or therefore on student achievement—but standards may enable other types of beneficial changes.

Finally, Schmidt et al. consider whether the differences in content standards, textbook coverage, and instructional time were related to the achievement gains that

students made between 7th and 8th grade. When considering all the countries together, these three mediating factors are all positively associated with students' math achievement gains. When considering variation within the U.S., however, the only positive relationship again involves textbook coverage. Earlier, we described how textbook coverage was associated with instructional and here we are focused on textbook coverage and achievement gains. Surprisingly, there is no relationship between instructional time and achievement gain, which suggests that they may be a problem with the way that one of these two variables was measured. Finally, as in the other analyses, the relationships are weaker in science. Instructional time and teacher coverage are actually negatively associated with science achievement gains.

Setting aside for the moment the reasons why these might not reflect causal relationships, it appears that the relationships indicated by Arrows B, C, and D in Figure 1 are often reflected in the data in the expected ways, at least in math. The topics in content standards are associated with the content of textbooks and these in turn are associated with the instructional time devoted to specific topics and then to student learning. In terms of the theory of action, this is consistent with the idea that standards improve achievement by “facilitating coordination with related elements of educational policies.” While this might not seem surprising, subsequent analysis of other studies will show that content standards do not always have these intended effects.

Research by Woessman. While the underlying data are similar, the research questions and analyses used by Woessman and colleagues differ. Woessman is primarily interested in the net effect of the control over curriculum and textbooks (Arrow A) on student achievement compared with Schmidt et al's focus on the mediating factors

(Arrows B-D). The strengths of the Woessman analyses is that he controls for student demographic differences and considers (control for) a wide range of institutional factors. However, he only uses data at a single grade and year, in contrast to Schmitt et al. who use a more preferred approach of changes in scores from 7th to 8th grade.

From his analyses, Woessman (2003) concludes:

“Among the many institutions which combine to yield major positive effects on student performance are centralized examinations and control mechanisms [including control over curriculum], school autonomy in personnel and process decisions, individual teacher influence over teaching methods, [and] limits to teacher unions' influence on curriculum scope.” (Woessman, 2003, p.117).

More specifically, he concludes that it is best to have curriculum choices determined by a centralized authority *and* by each individual teacher. These twin findings might seem paradoxical, but they can be reconciled in two ways: First, the findings may imply that the optimal policy is to have standards set by the national government level and to give teachers and schools authority over how the standards get implemented. This in fact summarizes the basic logic of standards-based reform as outlined by Smith and O'Day (1991).

Second, if intermediate-level authorities—namely, schools, school districts, and other sub-national bodies—are least effective as decision-makers, then we would expect to find that decision-making at the national and teacher levels leads to better outcomes. The idea that national control would be superior is implicit in arguments in favor of uniform standards in the U.S. But, given the smaller number of countries where control is located somewhere other than the national level—the U.S. is a rare exception—

the apparent benefits of national control may largely reflect other, unmeasured disadvantages of the few countries where control is more localized. It is not immediately obvious why the shift from the state to the national level would be so important.

Another possible explanation is that the model is misspecified. In particular, these regression estimates are based on the assumption that the benefits of teacher autonomy are the same in countries that, for example, have highly centralized control over the curriculum as it is where control is at the school level. This seems implausible. It is also something that is testable with Woessman's data, though the small number of observations (countries) does complicate matters.

In addition, the study finds that schools in which teachers acting collectively to influence instruction (either through a teacher's union or otherwise) lead to a negative impact on achievement. Woessman theorizes that this is because control over such matters results in "opportunistic" behavior as teachers make decisions that are more in their own interests than in the interests of students. If this is the case, however, it is difficult to see how giving more autonomy to teachers—a key finding in Woessman's analysis—improves matters. If teachers are behaving opportunistically, why would they do so only when they are acting collectively and not when they are acting individually?

Fuchs and Woessman (2004) also conducted analyses of developed nations that are members of the Organization for Economic Cooperation and Development (OECD), including mostly Western Europe and North America (nearly a subset of the TIMSS countries) where the Program for International Student Assessment (PISA) was administered. Unlike TIMSS, the PISA database includes many fewer variables regarding the characteristics of the national educational system, instead focusing on the

school level.¹⁰ They find that increased school-level control over textbooks is associated with higher achievement, but control over curriculum is not. This is somewhat perplexing because the control over textbooks is a form of control over the curriculum, and because Woessman's analysis of the TIMSS data show a positive relationship between achievement and local control over the curriculum. In any event, we give less attention to these findings than the TIMSS because the student achievement test used in PISA is intended to measure the practical applications of knowledge much more than either the TIMSS or the curricula actually intended in most countries. As discussed earlier, the current focus on rigor is more consistent with the advanced conceptual understanding than practical knowledge.

Discussion and Conclusions. These studies using cross-country comparisons suggest that centralized control over curriculum and textbooks will end up being reflected in instruction and in higher student test scores even on low-stakes exams such as TIMSS. Just like the other subsequent types of research, however, they also have their flaws and leave some questions unanswered.

While both sets of studies intend to analyze the effect of "centralization," the very meaning of this term is ambiguous when comparing countries that are of such vastly different sizes. Most of the countries in TIMSS are about the size of large U.S. states, and the U.S. as a whole is about as large as the entire European Union. We do not question that the U.S. system is among the most decentralized overall, but, when taking the overall geography and demographics into account, the differences in centralization between the U.S. and other countries seem somewhat exaggerated. This is even more

true given that the U.S. system has become considerably more centralized since the TIMSS data were collected, most obviously with the recent adoption of NCLB.

Also, both sets of studies require the implicit assumption that the adoption of particular sets of policies are unrelated to other factors, some perhaps unmeasured, that also influence student achievement. Woessman (2003) is most explicit about this assumption when he writes that “[educational] institutions . . . may reasonably be viewed as exogenous to students' educational performance” (p.121). While this may be true, it is a fairly strong assumption and, as noted earlier, a study of a similar type studies suggests that it is untrue with regard to policy adoption within the U.S. (Carnoy & Loeb, 2003). Also, the fact that the relationships between content standards, textbook coverage, and instructional time vary across countries could suggest that the U.S. is different in important ways so that adopting the policies of other countries may not have the same impact. Policy adoption may also depend on the national culture, which itself may influence student achievement. Given Hiebert and Stigler’s discussion of teaching as a cultural activity, this seems plausible. In any event, institutions clearly do not arise randomly and this is problematic for identifying causal impacts in these types of international analyses.¹¹

The international analyses raise other questions as well. Why might the national level of government be a more effective source of control over content and textbooks relative to regional bodies? Is this result driven by the U.S. which is one of the few nations with at least a moderate degree of regional authority? Why do the expected relationships seem to arise in math, but not science? Why do many of Schmidt et al’s key

results depend so highly on the unit of analysis? We cannot answer these questions at this time but do point to them as areas for future research.

U.S. Evidence from NAEP

Interestingly, analysis of cross-state differences is less common than international comparisons even though the quality of the data from the National Assessment of Educational Progress (NAEP) is arguably better than the TIMSS and the potential for important confounding factors is lessened by the fact that states are more similar than whole countries in terms of their social, economic, and educational policy contexts.¹² This is not to say that states do not vary, but that states do at least operate within somewhat coherent national contexts that involve many similarities. North Dakota is different than Texas, but probably more similar to Texas than to Japan.

To our knowledge, there is only one study that has examined the impact of standards per se on student achievement. In their unpublished study, Harris and Herrington (2004) perform regression analysis similar to the studies above that used regression analysis of state-level data on student achievement and various measures of centralized policies. They specifically use the Fordham Foundation's index of the quality of standards, while controlling for other demographic and policy measures, along the lines of Carnoy and Loeb (2003).¹³ They found no significant relationship between the quality of standards and gains in NAEP between 1992-2000 for any racial/ethnic group of students in either grade 4 or grade 8.¹⁴ While this study is unpublished, their results corroborated other findings of Carnoy and Loeb.

A larger set of studies have used NAEP to examine the relationship between teachers' self-reported instructional practices and student achievement. For example, Raudenbush, Fotiu, and Cheong (1998) found that teachers' emphasis on reasoning was positively associated with NAEP math achievement, while Wenglinsky (2004) found an instructional focus on memorizing facts was negatively associated with learning. This line of research can be thought of as part of Arrow D in Figure 1, though like most research on instructional practices, the analyses are correlational and cannot be used to draw clear causal inferences. The focus on reasoning as opposed to fact memorization is a key principle of the NCTM standards, a topic which we take up again later.

State Case Studies

Several U.S. states have implemented comprehensive standards-based reforms. California and Washington made major changes in their content standards in environments of low-accountability, while Kentucky implemented new standards as part of a comprehensive standards-based reform effort.

California. Much of the discussion of the California case is borrowed from two excellent books on the subject: Wilson's *California Dreaming* (2003), which focused on how the story of reform unfolded, and Cohen and Hill's *Learning Policy* (2001), which focused on teachers' perceptions of and responses to the reforms and student achievement. (See also McDonnell and Weatherford (1999).)

Beginning in the mid-1980s, California embarked on a comprehensive standards-based reform. The first step was a change in content standards intended to focus curriculum and instruction less on the traditional rote memorization, algorithms,

procedure, and teacher-led instruction and more on application and student-constructed (constructivist) problem-solving. In simple terms, this was viewed as a shift from conservative to liberal education although, as Wilson emphasizes, the policies implemented always included a balance of each.

Cohen and Hill (2001) report on a survey they conducted of elementary school (grades 2-5) teachers in California in 1994. They first considered teachers' awareness of standards-related documents.¹⁵ They found that 30-40 percent of teachers had read substantial portions of the state curriculum frameworks and nearly all were aware of them. (Teachers reported much less awareness and influence from the NCTM and other related standards documents, a point we return to later.) They also found that teachers were well aware of the basic ideas and broad philosophies espoused by the reforms, but were less familiar with the specific instructional strategies. There was also considerable agreement among teachers that the new philosophy was the right one, but much less agreement about the instructional techniques associated with those philosophies.

This last point is interesting in light of other evidence that teachers receive many, often conflicting, messages about what they should be doing—from centralized policies all the way to school colleagues and parents (e.g., Cohen and Spillane, 1993). What the evidence in California may suggest is that it is possible to get teachers to embrace a philosophy, even one coming from a single centralized source, but it may be much harder to translate into instruction. Changing ideas is relatively easy, while changing instruction requires a lot of time and energy—not just among the teachers but also those in charge of textbook design and adoption, professional development, and so on. Without that coordination, teachers are left on their own to choose among competing goals and

messages. Further, teachers might perceive that making small changes (e.g., having students spend a little more time doing group work) represents the expected response, whereas the advocates of the ideas might intend something more sweeping. While Cohen and Hill are appropriately cautious about any conclusions from their data about changes in instruction, their work does suggest that instruction did change and in ways consistent with the reforms. It simply did not change as much as reformers might have hoped and, in some cases, not in the exact ways that they had intended.

At the same time, almost no teachers reported that the broad curriculum frameworks were useful in developing lesson plans. This is not surprising, given that these were only frameworks, but teachers were also largely unaware of the documents the state had produced to provide more concrete instructional guidance, including documentation of the content on state standardized assessments and recommendations for instruction. This is somewhat ironic as teachers who participated in grading open-ended questions on the state standardized test, which can be viewed as a form of concrete guidance, saw it as a powerful opportunity for professional development.

Cohen and Hill place particular emphasis on the professional development—or lack thereof—experienced by teachers in general and especially those in California. They note the small amount of time teachers spend in professional and the general lack of focus that these opportunities involve. They also directly address the selection bias problem noted earlier. One might be concerned, for example, that the correlation between the amount of professional development taken by teachers and their instructional practices might simply reflect the fact that teachers who were already interested in and/or adopting reform-type practices were the teachers most likely to take the professional

development—essentially reversing the cause-effect relationship. They address this problem using an approach that economists call an “instrumental variables” (specifically, two-stage least squares). Intuitively, this method relies on the fact that some characteristics of teachers (e.g., teachers’ previous participation in general forms of professional development) were likely to influence the probability of taking subsequent reform-oriented PD, but not have much influence on the frequency of reform-oriented practices. While this approach is far from perfect, it goes much further than previous studies in showing the importance of standards-based professional development in causing standards-based practice and, overall, presents a compelling case that professional development was an important moderator of the effects of standards on instruction.¹⁶

Cohen and Hill also studied the effects of the California reforms on student achievement and found evidence that suggests a positive effect of reform-oriented instruction. Specifically, they analyzed school-level scores on the California state achievement test (CLAS), controlling for students’ eligibility for free lunches (an indicator of income and family background) and other school conditions. Their results suggest that student achievement was positively associated with instructional practices aligned to reform principles, whether teachers had learned about the state achievement test, and perhaps teachers’ use of specially designed “replacement units,” also aligned with the reform curriculum. The fact that there is a strong relationship between understanding of the state assessment and students’ scores on that assessment are not surprising. (As indicated earlier, what is more surprising is that, despite this obvious relationship, teachers seemed relatively unaware of the tools that might have helped them

better understand the assessment.) Likewise, there was no relationship between student achievement and the “conventional” instruction that characterized the pre-reform era.

While the authors present a careful analysis, they acknowledge many of its limitations.¹⁷ These results can best be viewed as suggestive that the reforms instituted in California influenced not only what teachers did in their classrooms, but that the changes in instruction also seemed to have a positive impact on student learning. Given this apparent success in raising achievement and the support for the reforms found by teachers, it might seem ironic that the policy so quickly unraveled. But a large number of political shifts were occurring in the state at about the same time that were largely unrelated to any objective indicator of the success of the reforms. This highlights the political difficulty of standards-based reforms, as well as a tendency in education debates to make big decisions about the success of policies based on short-term trends in student test scores (e.g., NAEP). As we argued earlier, the effects of content standards are likely to be delayed, and might not be reflected immediately in student test scores. Also, unless the effects are very large, there are any number of other factors influences student test scores at any given point in time that might impact achievement trends—changes in test instrument, changes in other policies, changes in student demographics. These other changes could easily overwhelm any real positive impact that a policy like standards might have. For this reason, it is important to pay attention to rigorous research studies that accounts for these potential problems.

Washington state and Kentucky. California has received more attention from researchers than other states, but is far from the only one to have made significant changes in content standards. Like the California, educational policy changes in

Washington State centered on changes in standards, with little emphasis on accountability. The research findings are also broadly similar. Stecher, Barron, Chun, and Ross (2000) found that teachers in Washington believed the standards were reasonable and attainable and that teachers did make changes in their instruction. While the test-based accountability system was relatively weak, the results suggested that schools in which instruction was more aligned to the curriculum made greater gains on the state's standardized tests. However, the authors also note that teachers seemed to be responding more to the tests than to the standards. This highlights the fact that public reporting of results alone constitutes accountability and can drive instruction, complicating the analysis of standards.

Unlike the California and Washington state cases, the Kentucky Education Reform Act (KERA) was a more comprehensiveness standards-based reform. Koretz, Barron, Mitchell, and Stecher (1996) found strong opposition to the accountability provisions in the reform and, apparently for this reason, the Kentucky reform as a whole was somewhat less popular among teachers than the California and Washington, although a slight majority of teachers in Kentucky still expressed support. Also, more than three-quarters of principals and teachers reported that the reform had led to improvements in instruction, especially among teachers who were resistant to change. A large majority of Kentucky teachers also reported having changed their instruction in response to the policies, especially increasing instructional time on tested subjects. The authors were not able to observe actual instruction, but if others studies are any indication, we would expect that the changes in instructional techniques were much less significant than the survey results might suggest.

Koretz and Barron (1998), in a follow-up study of the program found that the effects of the reforms on the state's high-stakes exam were four times larger than the effects on NAEP, which is a form of audit test. Klein et al. (2000) reported the same pattern of results in Texas. These findings reinforce the difficulty of identifying the impacts of standards-based reforms on the basis of high-stakes tests which themselves are part of the reforms. Nevertheless, the effects on NAEP were still apparently positive in Kentucky.

The results of all three state cases collectively suggest that teachers generally support standards (more so than accountability), that they report making changes in their instruction in response, and that the actual changes they makes are less significant than what they report. While it is difficult to isolate the impact of standards from other standards-based reforms, the results also suggest that standards raise student achievement, more so on state standardized tests than on NAEP. It is more reasonable to interpret this finding as being a result of standards themselves in California and Washington, compared with Kentucky where standards were part of a comprehensive reform package. The similarity in results is noteworthy given the apparent variation in contexts and in the policies themselves.¹⁸

NCTM Standards

In 1989, the National Council of Teachers of Mathematics (NCTM) published its own set of content standards. Like the California reforms, these emphasized greater focus on concepts and strategies to promote active learning. As Cohen and Hill (2001) note, teachers in California were well relatively unaware of the NCTM standards but, as

we show below, this was probably driven by the fact that California had already embarked on its own aggressive standards-based reforms, as well as the fact that the survey on which the Cohen and Hill survey was based occurred relatively soon after the NCTM standards were released.

Lubienski (2006) summarizes evidence of small-scale studies in which students being taught with NCTM reform curricula have outperformed control groups using various forms of traditional curricula. She concludes that students receiving the NCTM-based curricula have consistently outperformed other students. It appears that only one such study involved random-assignment of treatment to control. Ginsburg-Block and Fantuzzo (1998) randomly assigned a small sample of students to either control or one of two treatments: (a) peer collaboration; (b) the treatment associated with the NCTM reforms, referred to as problem solving; or (c) control. Despite the small sample, they found positive and statistically significant impacts of the NCTM-based instruction compared with the control, and no such impacts from peer collaboration.

Other studies of NCTM have focused on the interconnection between standards and other policies (Arrow B) and the impact on instruction (Arrow C). Cauley et al. (1993) reported that the vast majority of teachers in a sample of teachers in Richmond, Virginia were well aware of the standards, though only half to 2/3 of teachers reported that they were implementing the standards. Awareness and implementation were much higher at the secondary levels. Teachers also reported that the greatest “aids” to implementing standards were administrative support, especially from the school principal, and time and resources to learn more about the standards, e.g., observing other teachers’ classrooms.

Frykholm (1995) interviewed and observed a sample of student teachers at the University of Wisconsin at Madison. These teachers reported that they felt pressure from their university faculty to implement the standards, but little pressure from the cooperating teachers or other personnel in the schools in which they taught. Also, while the student teachers reported that they were implementing the standards, the observations of their practices contradicted these reports. As with the other categories of research, the NCTM findings suggest that, overall, standards probably do have some effect on instruction and achievement.

Discussion

In this section, we attempt to integrate the findings from the international, cross-state NAEP, state cases, and NCTM studies, as well as report on larger themes of past research. We organize the discussion around the analytic framework in Figure 1.

The Net Effect of Standards on Student Achievement (Arrow A)

There is essentially no convincing evidence of a direct impact of standards on student achievement or any other student outcome. There has never been a study to your knowledge that uses a rigorous quasi-experimental method to study the impact on student achievement.

The only two studies we are aware of that even remotely fit the Arrow A type are Harris and Herrington' analysis of NAEP and Woessman's analysis of TIMSS. In the first case, the authors are looking at changes in state NAEP achievement, but they measure only the "grades" given to the standards at a single point in time, precluding the

before-and-after comparisons that are arguably necessary. This is an even bigger problem in the Woessman studies where quite different countries are being compared. The Woessman studies also do not make use of changes in student achievement.

The Connection between Standards and Other Policies (Arrow B)

There is ample evidence of interconnections between standards and centralized policies. Accountability is the most obvious centralized policy that influences how content standards play out. The role of accountability has been discussed at length by other authors (Harris, 2007; Harris and Herrington, 2006; Rowan, 2005). We would only add here that much of the evidence discussed here was carried out before the passage of NCLB. Therefore, some findings, such as the mixed messages that teachers receive (Cohen and Spillane, 1992), are probably less applicable than in the past. Teachers are now getting a fairly clear message about the importance of test scores. In this respect, the tests are becoming the standards.

Accountability is not the only other centralized factors that matter, however. Schmidt et al. found that the degree of consistency between academic standards and textbook coverage were highest in countries where there was more centralized control over the curriculum. Cohen and Hill, as well as some of the studies of NCTM, found evidence about the importance of professional development to inform teachers about standards as well as concrete steps teachers could take to translate standards into instruction. The NCTM studies also highlight the importance of support from school administrators.

The Impact of Standards on Curriculum and Instruction (Arrow C)

Standards have caught the attention of educators at all levels of the system. Although the public is divided in its support of the No Child Left Behind Act (Rose and Gallup, 2007), there is general acceptance of the concept of higher academic standards among the public, educators and policymakers (Public Agenda, 2006). Teachers feel that state standards identify what their students should know and be able to do, that standards are compatible with good educational practice, and that public should hold students and educators to account for meeting certain outcomes. Teachers like common measures to calibrate teachers' expectations, and find standards useful for bringing focus and consistency of instruction within and across schools. They also find standards helpful for guiding their own instruction and align their instruction to them, although they feel that standards include more content than they can cover in a year. Also, some teachers believe that standards are, in some cases, too vague to provide useful guidance (Hamilton et al., 2007; Kannapel et al., 2001; Public Agenda, 2006; Stigler and Hiebert, 1999) and in other cases too specific so that they intrude on teacher discretion (Wilson, 2003). This, along with Woessman's finding that control at the national level and the teacher level are both positive, suggests a need to strike a balance by providing specific ideas to teachers about how to improve their instruction, but still leaving the decisions about those changes up to the teachers themselves.

Nevertheless, there is considerable evidence that even when teachers are supportive and aware of standards, it does not necessarily translate into meaningful changes in instruction (Fairman and Firestone, 2001; Spillane, 2004; Stigler and Hiebert, 1999; Wilson, 2003; Wilson & Floden, 2001). Stigler and Hiebert (1999) report that 95

percent of U.S. teachers surveyed in TIMSS were aware of the NCTM and state standards movements and 70 reported that they had made changes in their instruction and “most” teachers reported that they had read standards-related documents (e.g. curricular frameworks). However, their analyses of the videotapes of instruction of those same teachers suggest, as Cohen (1996) reports, that these change were at best minor and perhaps superficial.

Standards might also fail to take root because, as Stigler and Hiebert (1999) note, teaching is a “cultural activity,” learned informally over the many years that teachers previously spent in classrooms as students. One aspect of the American cultural perspective on education is the focus on learning “skills,” in contrast to the Japanese view of trying to help students learn “to think about things in a new way” (p.90). This perspective results in the “scripts” of teaching lessons that, as described earlier, are more focused on learning the procedures necessary to solve problems than learning concepts. The activity of teaching, including the underlying perspective and guiding scripts, are deeply ingrained through teachers own experiences as students, reinforced by their own observations of their colleagues. Instruction, therefore, might not be easily influenced by government policy, even through required formal training. Because teachers’ perspectives are rooted in their experiences and cultural expectations, and therefore hard to change, teachers respond to pressures for change by altering the surface but keeping their previous instructional approaches largely intact (Cohen, 1996).

A central lesson of this work is that academic standards, even if they are successful in changing what is taught, may do little to change how teaching and instruction actually take place. If the real problem is that American teachers need to

teach more for conceptual understanding, the impact of standards may not go far in defining the structure of those concepts. On the other hand, with standards, teachers have a narrower, common focus that might facilitate a common conversation and process for improvement.¹⁹ The curricular diversity now ingrained in the U.S. system (and documented, for example, by Schmitt et al., 2001), makes such continuous improvement more difficult to accomplish.

Who sets standards can affect the legitimacy of standards among educators and the public and therefore the degree of implementation. Although professional organizations like NCTM have used consensus processes to develop standards, consensus over the content of standards remains elusive both within and outside the education community. Wilson (2003), for example, describes the long history of the “math wars” in California. Kim (2007) has argued that without the imprimatur of exemplary classroom teachers, the National Reading Panel’s recommendations lacked legitimacy with some professional organizations and practitioners, slowing their adoption in the classroom. In this respect, each content area might face a different set of problems in implementing standards.

The Effects of Curriculum and Instruction on Achievement (Arrow D)

The final and most elusive step in the process is improving what happens in classrooms. The most convincing evidence comes from Schmidt et al.’s analysis of the TIMSS and Cohen and Hill’s analysis of California. Both studies find that instruction that is consistent with standards has an impact on student learning, but only in mathematics. Some of the results from the international TIMSS data provide mixed

results, even about the relationship between instructional time and achievement, but the relatively correlational nature of this analysis, and the problems involved with international comparisons discussed above, means that these contrary findings should be not be given too much weight. There is little rigorous evidence regarding the impact of NCTM, though these studies yield conclusions similar to Cohen and Hill.

Conclusion

Unlike many areas of educational research, there is no single landmark study that can be cited as evidence—for or against—the adoption of high-quality and uniform standards. There is no Perry Preschool Project or Tennessee class size initiative to point toward. Instead, we must rely on piecing together different types of evidence to draw a reasonable conclusion. Doing so is urgent given the growing national pressure to create common standards.

The evidence here suggests that taking recent movements toward national standards to the next level—to the national level—would likely improve student achievement. While each study cited here is imperfect, they all point in the same general direction—that teachers will make some changes in their instruction in response to standards and that these will probably lead to somewhat higher achievement. But the potential impacts are not as large as advocates might imply, for a variety of reasons. Even when teachers are aware of standards and accept their legitimacy, this often leads to little impact on instruction. Only roughly half of teachers who are aware of standards report that they changed their instruction and we know that many of those teachers have not truly aligned their instruction with the intent of the standards. Teachers receive many

conflicting messages about standards and have difficulty escaping the deeply ingrained notions of teaching that they grew up with (Stigler & Hiebert, 1999).

Existing research is nevertheless far from convincing and there is tremendous potential to improve research on standards so that clearer conclusions can be drawn. As described earlier, it is always difficult to identify causal impacts of educational interventions and this is even more true in the case of educational policies such as standards. Yet, with the possible exception of Cohen and Hill (2001), there has not been a single quasi-experimental approach used. This is not due to a lack of data. Below, we outline several research studies, many of which could be conducted with existing data:

Study Proposal (1): Analysis of data from the Northwest Evaluation Association (NWEA). The NWEA has student achievement data on nearly one million students in 10,000 schools located in 1,600 school districts and 45 states, including nine academic years. Because the data are based on a common testing instrument, they can be compared across states. The fact that the data are available in some schools over time, may make it possible to compare instruction and achievement before-and-after a change in standards at the district level. The NWEA also collects data on instructional programs that may be detailed enough to identify important linkages between classroom instruction and achievement—to help understand the causal pathway represented in Figure 1. See Petrilli (2008) for additional discussion of these data.

The NWEA data could be used to study either state or district-level changes in standards. It is important to note, however, that analyses at the school district level may not generalize to high-quality and uniform standards across the country. The school districts that choose to change standards may be the ones most likely to benefit from such

changes, so the impacts might be smaller on other districts that are forced to change by state or national policies might be smaller. Also, school districts that adopt standards on their own are likely to commit more resources that will facilitate implementation and changes in instruction.

Study Proposal (2): NAEP and changes in state standards. Control over standards in the U.S. at present resides at the state level. Many states have implemented standards in recent decades, although it is not clear that anyone has documented these changes. If the precise nature and timing of past changes in standards could be documented, and if these changes lined up with the measurement of student achievement (NAEP is not administered every year) in a substantial number of states, then it may be possible to estimate effects that could reasonably be interpreted as causal influences. It is important to note, as do Harris and Herrington (2006), that changes in any one policy (in this case, standards) might occur at roughly the same time as other educational policies, potentially confounding the analyses. This analysis might not be possible if only a small number of states experience changes in standards that happen to line up with the timing of NAEP testing. In this case, it would be necessary to rely completely on state high-stakes tests (without an audit test), bearing in mind the caveat that the results may be inflated by test preparation and other undesirable changes in instruction.

The NAEP data for such an analysis might be combined with data collected by the Wisconsin Center for Education Research (WCER) at UW-Madison, which has been collecting data over many years in conjunction with conferences held with the Council of Chief State School Officers where researchers have conducted analyses of state content standards. WCER now has data from 31 states. Porter, Polikoff, and Smithson (2008)

describe these data in greater detail and use them to describing variation in state standards and other topics. Combining the NAEP results with these existing data on state standards may produce the type of longitudinal data that facilitates quasi-experimental analysis.

Study Proposal (3): District-level randomized experiments. We discussed earlier the difficulty of carrying out randomized trials with regard to policies, but they are not impossible. There are more than 16,000 school districts in the U.S., many of which are located in states that have not aggressively pursued high-quality standards and/or the interrelated policies necessary to make them work. It might also be possible for a state or non-profit group to implement a pilot program, like the Tennessee STAR class size initiative, in which districts could be recruited to participate in changes in content standards and associated programs. Of those volunteers, one sub-group could be randomly selected to go first, and the remainder could receive the reforms in the next year. This phased implementation approach is politically feasible because all the districts requesting the reforms would receive it, albeit at different times. This approach would also have the advantage of allowing the researchers to provide audit tests, to go along with the high-stakes test.

Again, while the collection of evidence here is generally supportive of the potential of uniform and high-quality content standards, it remains obvious that such a reform would not have much of an impact on its own, nor would the short-term effects likely be very large. Such a policy, if it is to be implemented at all, is only likely to show a demonstrable impact on achievement if it is seen as credible by the educators who would have to implement them and if it were accompanied by supports to help them do so. This policy approach might be more welcomed than ever given the growing

frustration of teachers with the pressure to raise test scores, but it could also backfire if, as in previous state-based reforms efforts, the public expects rapid or large increases in test scores. By setting reasonable expectations, and commissioning rigorous research, it may be possible to learn more about which standards policies work, which do not, and whether uniform and high-quality standards should be built into the long-term landscape of U.S. education policy.

References

- Cauley, K.M., Van de Walle, J., & Hoyt, W. (1993) *The NCTM standards: Implementation*. Metropolitan Education Research Consortium: Richmond, VA.
- Cogan, L. S., & Schmidt, W. H. (1999). Middle school math reform. *Middle Matters*, 8, 2-3.
- Cohen, D. (1990). *The classroom of state and federal education policy*. East Lansing, MI: School of Education, Michigan State University.
- Cohen, D. & Hill, H. (2001). *Learning policy*. New Haven, CT: Yale University Press.
- Cohen, D. & Spillane, J. (1992). Policy and practice: The relations between governance and instruction. In Gerald Grant (Ed.), *Review of Research in Education*, 18, 3-49.
- Collins, A. (1997). National science education standards: Looking backward and forward. *The Elementary School Journal*, 97(4), 299-313.
- Elmore, R. F., & Rothman, R. (1999). *Testing, teaching, and learning: A guide for states and school districts*. Washington, DC: Academy Press.
- Fairman, J. C., & Firestone, W. A. (2001). The district role in state assessment policy: An exploratory study. In S. H. Fuhrman (Ed.). *One Hundredth Yearbook of the National Society for the Study of Education. Part II: From the Capitol to the Classroom: Standards-Based Reform in the States* (pp. 124-147). Chicago: University of Chicago Press.
- Finn, C., Petrilli, M., & Julian, L. (2006a). *The state of state standards: Executive summary*. Downloaded April 8, 2008 from <http://www.edexcellence.net/institute/publication/publication.cfm?id=358&pubsubid=1317#1317>.
- Finn, C., Petrilli, M., & Julian, L. (2006b). *The state of state standards: California*. Downloaded April 8, 2008 from <http://www.edexcellence.net/institute/publication/publication.cfm?id=358&pubsubid=1317#1317>.
- Frykholm, J.A. (1995). *Impact of the NCTM standards on pre-service teacher beliefs and practices*. Paper presented at the annual meeting of the American Education Research Association.
- Gamoran, A. (1986). Instructional and institutional effects of ability grouping. *Sociology of Education*, 59, 185-198.

- Ginsburg-Block, M.D. & Fantuzzo, J.W. (1998). An evaluation of the relative effectiveness of NCTM standards-based interventions for low-achieving urban elementary students. *Journal of Educational Psychology*, 90(3), 560-569
- Goertz, M. E. (1986). *State educational standards: A 50-State survey*. Princeton, NJ: Educational Testing Service.
- Goertz, M.E. (2007). Standards-based reform: Lessons from the past, directions for the future. Paper presented at Clio at the Table: A Conference on the Uses of History to Inform and Improve Education Policy, Brown University, Providence RI, June 2007.
- Goertz, M. E., & Duffy, M. C. (2001). Assessment and accountability systems in the 50 states: 1999-2000. RR-046. Philadelphia, PA: University of Pennsylvania, Consortium for Policy Research in Education.
- Hamilton, L. S., Stecher, B. M., Marsh, J. A., McCombs, J. S., Robyn, A., Russell, J. L., Naftel, S., & Barney, H. (2007). Standards-based accountability under No Child Left Behind: Experiences of teachers and administrators in three states. MG-589-NSF. Santa Monica, CA: RAND.
- Harris, D.N. & Herrington, C.D. (2004). The effects of accountability and standards. Tallahassee, FL: Florida State University.
- Harris, D.N. & Herrington, C.D. (2006). Accountability, standards, and the growing achievement gap: Lessons from the past half-century. *American Journal of Education*, 112(2), 209-238.
- International Technology Educators Association (2000). A guide to develop standards-based K-12 technology education. Reston, VA.
- Kannapel, P. J., Aagaard, L., Coe, P., & Reeves, C. A. (2001). The Impact of Standards and Accountability on Teaching and Learning in Kentucky. In S. H. Fuhrman (Ed.). *One Hundredth Yearbook of the National Society for the Study of Education. Part II: From the Capitol to the Classroom: Standards-Based Reform in the States* (pp. 242-262). Chicago: University of Chicago Press.
- Koretz, D.M., Barron, S., Mitchell, K.J., & Stecher, B.M. (1996). Perceived effects of the Kentucky Instructional Results Information System (KIRIS). Washington, DC: Rand Corporation.
- Kim, J.S. (2008). Research and the reading wars. In Frederick M. Hess (Ed.) *When Research Matters: How Scholarship Influences Education Policy* (pp. 89-111). Cambridge, MA: Harvard Education Press, 2008.

- McDonnell, L. & Weatherford, M.S. (1999). State standards-setting and public deliberation: The Case of California (CSE Technical Report 506, National Center for Research on Evaluation, Standards, and Student Testing). Los Angeles, CA: UCLA.
- National Commission on Excellence in Education (NCEE). (1983). *A nation at risk: The imperative for educational reform*. Washington, DC: U.S. Department of Education.
- National Council of Teacher of Mathematics (1989). *Curriculum and evaluation standards for school mathematics*. Reston, VA.
- National Council of Teacher of Mathematics (2000). *Principles and standards for school mathematics*. Reston, VA.
- National Research Council (1996). *National science education standards*. Washington, DC: National Academy Press.
- National Research Council (1999). *Testing, teaching and learning: A guide for states and school districts*. Committee on Title I Testing and Assessment. R. F. Elmore and R. Rothman (Eds). Board on Testing and Assessment, Commission on Behavioral and Social Sciences and Education. Washington, DC: National Academy Press.
- National Research Council (2002). *Investigating the influence of standards: A Framework for research in mathematics, science, and technology education*. I.R. Weiss, M.S. Knapp, K.S. Hollweg, and G. Burrill (Eds.), Committee on Understanding the Influence of Standards on K-12 Science, Mathematics, and Technology Education., Center for Education, Division of Behavioral and Social Sciences and Education. Washington, DC: National Academy Press.
- Oakes, J. (1985). *Keeping track: How schools structure inequality*. New Haven, CT: Yale University Press.
- Ogbu, J.U. (2003). *Black American students in an affluent suburb: A study of academic disengagement*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Padilla, C., Skolnik, H., Lopez-Torkos, A., Woodworth, K., Lash, A., Shields, P. A., Laguarda, K. G., and David, J. L. (2006). *Title I Accountability and School Improvement From 2001 to 2004*. Washington, DC: U. S. Department of Education, Office of Planning, Evaluation and Policy Development, Policy and Program Studies Service.
- Petrilli, M. (2008). Presentation to the National Research Council Workshop on Assessing the Role of K-12 Academic Standards in States. Available: <http://www7.nationalacademies.org/cfe/Petrilli%20Presentation.pdf>.

- Porter, A., Polikoff, M., and Smithson, J. (2008). Is there a de facto national curriculum? Evidence from state standards. Paper prepared for the National Research Council Workshop on Assessing the Role of K-12 Academic Standards in States.
- Public Agenda (2006). Reality Check 2006: Issue No. 3: Is Support for Standards and Testing Fading?
www.publicagenda.org/specials/realitycheck06/realitycheck06_main.htm
 Accessed on August 25, 2006.
- Raudenbush, S.W., Fotiu, R.P., & Cheong, Y.F. (1998). Inequality of access to educational resources: A national report card for eighth-grade math. *Educational Evaluation and Policy Analysis*, 20(4), 253-267.
- Rose, L. C., & Gallup, A. M. (2007). The 39th annual Phi Delta Kappa/Gallup poll of the public's attitudes toward the public schools. *Phi Delta Kappan*, 89(1):33-48.
- Rothman, R. (2004). Benchmarking and alignment of state standards and assessment. In S. H. Fuhrman and R. F. Elmore (Eds.). *Redesigning Accountability Systems for Education* (pp. 96-137). NY: Teachers College Press.
- Rowan, B. (2005) Evidence and values in Research on Standards and Testing: Reflections of a researcher. Ann Arbor, MI: University of Michigan.
- Shavelson, R.J., McDonnell, L.M., & Oakes, J. (1989). *Indicators for monitoring mathematics and science education: A sourcebook*. Santa Monica, CA: RAND Corporation.
- Smith, M.S., & O'Day, J. (1991). Systemic school reform. In S. H. Fuhrman and B. Malen (Eds.). *The Politics of Curriculum and Testing* (pp. 233-267). London: Falmer Press.
- Spillane, J. P. (2004). *Standards deviation: How schools misunderstand education policy*. Cambridge, MA: Harvard University Press.
- Stecher, B., Barron, S., Chun, T., & Ross, K. (2000). The effects of the Washington State education reform on students and classrooms." CSE Technical Report #525. Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing.
- Stigler, J. & Hiebert, J. (1999). *The Teaching Gap: Best Ideas from the World's Teachers for Improving Education in the Classroom*. New York: Free Press.
- Taylor, L.L. (2002). A dose of market discipline: The new education initiatives. *Southwest Economy*, 3, 1-12.

- Wenglinsky, H. (2004). Facts of critical thinking skills? What NAEP results say. *Educational Leadership*, 62(1), 32.
- Wilson, S. M. (2003). *California Dreaming: Reforming Mathematics Education*. New Haven, CT: Yale University Press.
- Wilson, S. M., & Floden, R. E. (2001). Hedging bets: Standards-based Reform in Classrooms. In S. H. Fuhrman (Ed.). *One Hundredth Yearbook of the National Society for the Study of Education. Part II: From the Capitol to the Classroom: Standards-Based Reform in the States* (pp. 193-216). Chicago: University of Chicago Press.
- Woessmann, L. (2003). Schooling resources, educational institutions and student performance: The international evidence. *Oxford Bulletin of Economics and Statistics*, 65(2), 117-170.

Notes

¹ The term “outcomes” carries different meanings. Economists tend to treat outputs and outcomes as one in the same. Scholars of organizational behavior, in contrast, tend to treat outputs and processes as the same. We use the latter terminology.

² Formally, “economies of scale” refers to the idea that an increase in inputs leads to a more than proportional increase in output. This results in decreasing marginal costs.

³ The focus on concepts is related to what Schmidt et al. refer to as “complex performance expectations.” Asking students to define terms and learn procedures for applying terms to solve highly stylized problems is arguably less complex than asking them to learn a concept and then choose among concepts in solving problems.

⁴ It is worth noting the observation by Stigler and Hiebert that Japanese students were frequently asked to “invent” ways of solving problems. This same practice was part of the “New Math” approach that has been roundly criticized.

⁵ In one sense, this scale-up problem gets somewhat smaller when studying states as opposed to school districts—states contain a much wider range of schools. If the impacts vary little across different types of schools within the state, then the estimated impacts are more likely to generalize to other states. On the other hand, states vary with respect to the overall policy environment—the centralized policies in Figure 1—and the impact of standards may depend on these state policies.

⁶ The 7th and 8th grade scores come from a single point in time. Under the assumption that the two cohorts are the same in ways that influence student achievement, the difference between the two scores in each country can be interpreted as a change in achievement.

⁷ In defining what they mean by “content standards,” Schmidt et al. write that “official documents often provide direct statements of the content and performance levels desired for students” (p.3). They also describe the process of data collection: “First, curriculum documents—a representative sample of content standards and student textbooks—were identified and collected in each participating country” (p.24). The content of standards and textbooks was specific to the topic. It is unclear how the representativeness of the documents was established. Further, it is unclear how content standards were measured in countries, such as the U.S., that have little in the way of “national documents.”

⁸ By “topic coverage,” I refer to both the “percent teacher coverage” and “instructional time” variables. In most cases, the results are the same for both. In cases where the results differ, I defer to the instructional time variable that, in contrast to teacher coverage, refers simply to the percent of teachers covering the specific topic. Unfortunately, they study only those teachers who report instructional time greater than zero. It is somewhat unclear why they did not simply define those teachers with no instructional time as zero. It is clearly important to understand why some teachers might not cover a topic at all.

⁹ Figure 6.8 on page 209 in Schmidt et al. is misleading on this point as it implies that there is no relationship between content standards and textbook coverage or between content standards and instructional time, whereas the absence of arrows actually reflects the fact that the relationships cannot be estimated.

¹⁰ They include two national-level variables that are not of direct interest here: presence of an external exit exam and use of standardized tests. It is unclear why they did not include national level variables from the TIMSS which includes nearly all the countries in the PISA. Also, they did not conduct within-country regressions.

¹¹ Another reason to be concerned about selection bias is that the estimated effect of class size has the “wrong sign”; that is, larger classes seem to produce higher achievement. Experimental evidence on class

size in the U.S. suggests the opposite. While class size is not of direct interest here, the fact that some of the coefficients seem implausible also calls into question other aspects of the analysis.

¹² While a large number of studies have examined the effects of accountability on student achievement, such studies, because they do not account for the role of standards, are beyond the scope of the present study.

¹³ Carnoy and Loeb focused on the effects of accountability and do not consider standards as a policy variable. Independently, Harris and Herrington used the same accountability measures, but rather than use an index of measures as in Carnoy and Loeb, they focused on specific policies such as standards. The Harris and Herrington paper was never published because it largely overlapped the already published work of Carnoy and Loeb.

¹⁴ Like Carnoy and Loeb, the policy variables in this analysis are not time-varying. Also, the analyses compare 4th grade scores in 1992 with 4th grade scores in 2000 and likewise for 8th grade. While some state characteristics are controlled, it is possible that other aspects of the state student population and policy regimes changed over these periods as well and these changes might be correlated with content standards.

¹⁵ The five documents were: (1) Mathematics Framework for California Public Schools, (2) Mathematics Framework for California Public Schools, (3) A Sampler of Mathematics Assessment, Mathematics Model Curriculum Guide , K-8, (4) NCTM Curriculum and Evaluation Standards, and (5) NCTM Professional Teaching Standards.

¹⁶ The basic problem with all instrumental variables approaches is that one of the assumptions of the methods—that the instrument cannot be related to the outcome variable—is impossible to test. In this case, Cohen and Hill must assume that the professional development (PD) that teachers received was uncorrelated with their tendency to use reform-oriented practices. This is plausible, and they argue that there is little to suggest that earlier PD would have been consistent with reform principles, but it is still improvable. It is also possible that the effect they observe for those who participated would not extend to teachers who were required to take part in the PD.

¹⁷ One significant limitation not mentioned in the analysis is that the CLAS measures only attainment at a point in time, as opposed to learning over time. Controlling for student demographics is a relatively weak substitute for this, though the authors were left with no choice as the more desirable data were not available.

¹⁸ In addition to the fact that the Kentucky standards were part of a more comprehensive package, there was also variation in the standards themselves. For example, unlike California, which received high marks in the Fordham Institute rankings, Kentucky and Washington received grades of D and D-, respectively. The fact that the results are similar across the three states, despite the differences in standards on this one measure, may suggest that the Fordham ratings are not very good indicators. Or, it may mean that the quality of standards is not very important. Recall that Harris and Herrington (2004) found no relationship between state NAEP gains and the Fordham quality measures.

¹⁹ Note that this common conversation and process need not lead to a common approach to instruction for standards to be effective. It is likely that there are different approaches to instruction that are equally effective, especially when tailored to individual student needs.