Standards-Based Reform in the United States: History, Research, and Future Directions

Laura S. Hamilton, Brian M. Stecher, and Kun Yuan

RAND Corporation

Paper commissioned by the Center on Education Policy, Washington, D.C.
For its project on Rethinking the Federal Role in Education

December 2008

This work was supported by the National Science Foundation under Grant No. REC-0228295. Any opinions, findings and conclusions or recommendations expressed in this paper are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Summary

Standards-based reforms (SBR) have become widespread across the United States, particularly in the wake of No Child Left Behind (NCLB). Although there is no universally accepted definition of SBR, most discussions of standards-based reform include some or all of the following features: *academic expectations for students* (the standards are often described as indicating "what students should know and be able to do") *alignment of the key elements of the educational system* to promote attainment of these expectations, the use of *assessments of student achievement* to monitor performance, *decentralization* of responsibility for decisions relating to curriculum and instruction to schools, *support and technical assistance* to foster improvement of educational services, and *accountability* provisions that reward or sanction schools or students on the basis of measured performance. Each instance of SBR emphasizes certain components more than others.

The SBR movement reflects a confluence of policy trends—in particular, a growing emphasis on using tests to monitor progress and hold schools accountable and a belief that school reforms are most likely to be effective when all components of the education system are designed to work in alignment toward a common set of goals. Many of the SBR systems that have been adopted in response to the requirements of NCLB had their origins in state and federal initiatives from the 1980s and 1990s and in activities conducted by professional organizations such as the National Council of Teachers of Mathematics.

Although notions of what constitutes effective SBR have changed over time, the core elements mentioned above have endured; in addition, a few key ideas have emerged in recent years. These ideas include an emphasis on using information produced by the system to guide instructional decision making; an emphasis on using standards to promote instruction that is academically challenging rather than focused on low-level skills; the importance of similarly high expectations for students with different socioeconomic, racial/ethnic and linguistic backgrounds; and, perhaps most significantly, an education system in which policy and practice are driven in large part by the measurement of academic outcomes derived from large-scale assessments.

What Research Tells Us About SBR

A large body of research has been conducted over the past few decades to assess the quality and impact of various SBR systems. One line of investigation has examined the quality of the standards themselves. A review of these efforts suggests that there continues to be a lack of consensus regarding the features of high-quality standards; some states find that their standards receive positive marks from one organization and are criticized by another. Regardless of what criteria are used to evaluate quality, the existing reviews suggest that there is room for improvement; relatively few states get uniformly high marks under any set of criteria.

A second collection of research addresses the important question of how SBR affects what educators do. A few studies have attempted to examine how the creation and publication of standards, *per se,* have affected practices. The research suggests that standards accompanied by curriculum reform efforts can change the content of instruction, but that standards alone are unlikely to influence practice in a significant way. The bulk of research relevant to SBR has focused on the links between high-stakes tests and educators' practices. The preponderance of research on the impact of testing rather than the impact of standards reflects the emerging realization that "standards-based reform" has largely given way to "test-based reform," a system in which the test rather than the standards communicates expectations and drives practice.

Studies of relationships between high-stakes testing and school and classroom practices have produced one consistent finding: High-stakes testing systems influence what teachers and administrators do. Some of the changes would generally be considered beneficial (e.g., providing additional instruction to low-performing students; taking steps to align the school curriculum across grades), whereas others raise concerns about possible negative effects on the breadth and quality of instruction (e.g., shifting resources from untested subjects or topics to tested subjects or topics; focusing on specific test item styles or formats). Research also suggests that teachers have maintained a high level of autonomy in how they teach, and that SBR (or its surrogate, test-based reform) typically does not produce fundamental changes in pedagogy. Thus, while SBR does appear to influence educators' practices, it does not always do so in consistent or predictable ways.

Finally, the question that may be of greatest interest to policymakers and the public is whether SBR or the test-based reform that has predominated in recent years has improved student learning. Recent gains on state accountability tests suggest that achievement as measured by those tests has increased since the enactment of No Child Left Behind in some but not all states. However, the reason for these gains is not clear. They could be due to test-based reform, to other reforms taking place at the same time, or to a phenomenon called score inflation (i.e., score increases on high-stakes tests that do not generalize to other measures of the same content, for example, because they primarily reflect narrow test-preparation activities geared toward a specific test).

Past research that examines score gains on low-stakes (non-accountability) tests, such as the National Assessment of Educational Progress, suggests that there has been some increase in achievement associated with state accountability policies, though the gains on low-stakes tests are not as large as the gains on the high-stakes state tests, so it is difficult to know the size of the actual increase. Unfortunately, there is conflicting evidence on this point, and most of the research was conducted prior to the enactment of NCLB. As noted, it is also impossible to disentangle the effects of SBR and high-stakes testing from other initiatives taking place during the same period. Because of the difficulties inherent in trying to link achievement gains to specific policy initiatives, questions still remain about the effects of standards and assessments on student achievement.

What Have We Learned? Tensions and Challenges

Although definitive evidence regarding SBR's effects remains elusive, the available research does help us identify several lessons that point to challenges faced by those who develop SBR systems and those whose work puts them in the position of responding to these systems:

- ***When tests have high stakes, standards may take a back seat.*** As noted above, the tests rather than the standards tend to drive practice, potentially undermining some of the

benefits that are presumed to accrue from the alignment of curriculum, instruction, and other features of the education system with the state standards.

- ***Existing tests do not adequately measure all standards.*** The tests that are currently used in most states do not measure all of the knowledge and skills expressed in the standards. This is a fundamental shortcoming of tests, which are of necessity small samples of knowledge and skills from much larger domains. In addition, there is a tendency for large-scale tests to focus more on low-level skills that are easy to assess through multiple-choice items and to give short shrift to more complex problem-solving and reasoning skills.

- ***When strong sanctions are attached to specific measurable outcomes, practices tend to become distorted.*** Because the tests drive responses, the kinds of practices that teachers and administrators adopt in response to SBR tend to focus more on tested material and less on the untested content of the standards than would generally be desired. Excessive test preparation and other practices designed to raise test scores without promoting broader knowledge and understanding are another manifestation of this effect. One of the primary factors contributing to these distortions is the predictability of test content and format from one year to the next, a natural consequence of states' desires to adopt cost-effective measures that can be statistically linked to measure progress over time.

- ***SBR allocates responsibility in ways that can conflict with traditional educational governance.*** Some early proponents of SBR envisioned a trade-off in which higher-level policymakers established standards, and local educators familiar with the needs of students were given responsibility for decisions about curriculum and instruction. However, this has not always happened in practice, because state administrators, school boards, and others in leadership positions try to exert control over curriculum and instructional decisions, particularly when student performance is low. The resulting lack of clarity about who is responsible for what has led to tensions across levels of the system.

- ***Alignment and autonomy may become competing goals.*** Some of the lack of local decision-making authority stems from district or state efforts to create an aligned system by developing resources, such as pacing guides and interim assessment systems, that are designed to match the state test. Even if district officials are not telling teachers what and how to teach, the requirement that teachers adopt these tools can constrain what they do.

- ***Federalism continues to pose challenges for SBR.*** One of the most frequently heard criticisms of today's SBR systems is the wide variation in features of state accountability systems, particularly the varying meanings of "proficient." There is a clear tension between giving states the flexibility to design SBR systems that suit their needs and creating a set of systems that functions somewhat similarly across states.

Where Do We Go from Here

There are legitimate questions about the appropriate role for the federal government in SBR, especially given the mixed findings to date relating to the implementation and impact of NCLB. The research evidence does not provide definitive answers to these questions. However, if the federal government does continue its involvement in these initiatives, there are some directions that appear promising. History shows that the federal government can influence state and local education policy and practice despite its relatively small share of school and district budgets, so it is worth considering productive strategies it can take to improve the quality of education through SBR. At the same time, the history of the SBR movement demonstrates the importance of the federal government collaborating with states and other entities rather than simply issuing mandates. We identify a few ways in which federal efforts might be beneficial, many of which have been discussed by other groups, as well:

- Improve standards and assessments by convening expert panels, developing mechanisms for cross-state collaboration, and awarding grants to support the development of high-quality standards and assessments with an emphasis on promoting cognitively challenging instruction.

- Develop accountability indices that create more effective incentives by addressing the shortcomings that research has identified in current accountability metrics. For example, indices based on growth in achievement that also take into account performance all along the achievement scale (rather than just whether a student is above or below "proficient") should provide better information about performance, result in higher levels of buy-in from educators, and be the basis for a set of incentives that may be more consistent with public goals for education than the current system.

- Experiment with alternative SBR approaches to enhance our understanding of the effects of specific features, such as the strength of the incentives or the level of prescriptiveness of the standards.

- Continue to use and broaden NAEP (National Assessment of Educational Progress), which not only provides a means of comparing student performance across states, but has also allowed us to monitor achievement on subjects not typically included in state accountability systems. In addition, NAEP can provide opportunities to test on a large scale new methods for measuring student achievement, such as new, performance-based item formats.

- Support the evaluation of SBR efforts. If the federal government takes steps to require or promote SBR, it should also set aside resources for evaluating the effects of these policies. Such evaluations should measure more fully the impact of SBR on the broad goals of the education system. There is a critical need for a better understanding of a broad set of outcomes that may be associated with SBR, including graduation rates, course-taking patterns, and student learning in subjects not included in the high-stakes testing system.

A More Comprehensive Vision of SBR for the Future

This exploration into the history and effectiveness of SBR makes it clear that the original, comprehensive vision of SBR has never been given a full trial. In this final section we revisit the

key SBR elements listed at the beginning of the paper and present ideas for rethinking some of them with the goal of promoting increased opportunity for all students to receive high-quality instruction throughout their K-12 years.

- ***Standards:*** Most states have developed a broad array of standards that address far more than the subject areas for which testing is required under NCLB, including social studies and the arts, as well as topics such as career awareness that span multiple subjects. These non-tested standards should be mapped onto indicators that could be used for periodic monitoring of schools' efforts to promote attainment of these standards, as discussed below. In addition, the standards (as well as the curricula adopted to support them) should allow for multiple postsecondary pathways rather than pushing all students toward the same goal of a four-year college degree. A number of lucrative, stimulating, and rewarding careers are available to students whose interests may not be perfectly aligned with what is required to earn a bachelor's degree. It is critical to promote the necessary skills and knowledge to prepare students for these different paths while not imposing so many specific requirements that students lack the flexibility to explore fields that capture their interest.

- ***Alignment:*** Current alignment efforts are often narrow and should be expanded to reflect a more systemic view of the educational system. The current emphasis is on routine matching of the content of standards, tests, and curriculum. This view of alignment should be expanded to more closely reflect the ideas of early SBR advocates who envisioned a system in which teacher preparation, professional development, leadership, and other supports were all aligned to promote instruction toward a common set of standards. Perhaps most important and least well developed among these supports is a set of resources to model and promote high-quality, standards-based instruction. These instructional supports could include sample lesson plans or other materials to help teachers help their students meet the expectations embodied in the standards while avoiding the tendency to focus on a narrow set of skills and question formats included in a specific test. Similarly, better resources are needed to help teachers use data for decision making; such resources would eschew the test-focused use of data and foster

student-focused individualization to address students' unique needs and interests. Development of these resources should involve collaboration between the public and private sectors to take advantage of the many innovative ideas currently being generated in both sectors.

- ***Assessment:*** Advances in psychometrics and technology offer the possibility of new assessment methods that would tap a broader range of skills and knowledge than today's multiple-choice tests, and do so more efficiently and at reasonable cost. The development of high-quality assessments that require students to apply complex problem-solving and reasoning skills and are relatively immune to test-focused instruction could go a long way toward improving outcomes associated with SBR. Testing systems could be improved by increasing variety in the content and format of tests; lack of predictability is a key to reducing overly narrow test preparation and the resulting score inflation.

- ***Accountability:*** A comprehensive indicator system could replace today's test-based approach to accountability. The inclusion of non-test outcome measures would provide better information on how schools are performing in a wide range of dimensions, would reduce the pressure to focus exclusively on tested material, and could allow schools to experiment with innovative curricula and programs rather than sending the message that conformity with a narrow set of goals is desired. It is not necessary that the accountability system measure every outcome every year, nor that it provide reliable scores on every student every year in every subject. A blended indicator system in which some outcomes are measured annually and others periodically, and in which some skills are measured at the student level while others are measured at the aggregate (classroom, school, or district) level, holds promise as a way to balance breadth, burden, and cost and produce a better basis for monitoring educational performance and assessing accountability. It could also minimize the predictability that is inherent in today's narrower test-based systems and therefore reduce the likelihood that educators will narrow their instruction to focus on a small number of measured outcomes. Policymakers should consider supplementing outcome measures with a set of process indicators that provide information on what schools are actually doing. Such indicators could further support the goal of reducing

> excessive emphasis on tested content and could provide information on school practices and opportunities that are likely to promote important but hard-to-measure outcomes, such as civic-mindedness, teamwork, and creativity. Finally, accountability should be expanded beyond the school-focused, regulatory model that dominates current policy discussions and should place more emphasis on rewarding good performance, providing incentives for students, and giving parents expanded options.

SBR has been shown to be a powerful lever for changing practice at all levels of the education system. Some of the hopes of early reformers have been at least partially realized, but some of the fears of early critics have materialized. Ongoing efforts to improve our knowledge of the effects of SBR and disseminate this knowledge to decision makers at all levels will be critical for improving SBR systems that promote high-quality teaching and learning.

Introduction

Standards-based reform is one of the most prominent features of the current educational landscape. Across the nation, states have adopted standards that describe the content that schools are expected to teach and that students are expected to master. The requirement for standards and aligned assessments has been a feature of federal legislation since the Improving America's Schools Act (IASA) of 1994, and it is the centerpiece of the No Child Left Behind Act of 2001 (20 U.S.C. 6311 *et seq*.). NCLB, which was signed into law in January 2002, has exerted a strong influence on state and local decisions about education policy and practice since then. The NCLB requirements represent one important milestone in the evolution of a movement that had been in place for more than a decade prior to the law's enactment. This paper summarizes the history of the SBR movement, discusses what we know about how this movement has shaped educators' practices and student outcomes, and puts forth recommendations for improving these policies in the future.

What Is Standards-Based Reform?

Since the 1990s, the term "standards-based reform" (SBR) has been used extensively in discussions of educational policy, providing a wealth of material for us to summarize in this paper. However, "standards-based reform" has not meant the same thing to everyone who has used the term. As Wilson and Floden (2001) observed, "The slogans of standards and SBR spread widely in the 1990s, but the meaning varied across contexts" (p.195). We discuss these variations below. Further complicating the situation, educators and policymakers have used other terms, including "systemic reform," "standards-based accountability" and "curriculum alignment" to describe similar ideas that differ somewhat in emphasis or evolution.

All conceptions of standards-based reform incorporate some or all of the following six features: *academic expectations for students* (often described as indicating "what students should know and be able to do"), *alignment of the key elements of the educational system* to promote attainment of these expectations, the use of *assessments of student achievement* to measure outcomes, *decentralization* of responsibility for decisions relating to curriculum and instruction to schools, state and district *support and technical assistance* to foster improvement of educational services, and *accountability* provisions that reward or sanction schools or students on the basis of measured performance. We discuss each component below.

Academic Expectations for Students (i.e., "Standards")

One essential feature of SBR is an emphasis on managing education in terms of student outcomes rather than on inputs, such as the quality of facilities or per-pupil expenditures. As Massell and Fuhrman (1994) stated, "Standards that express desired content and performance play a central role in current reforms, so much so that many refer to prevailing efforts as *standards-based* reform." One of the most prominent advocates of standards has been former Assistant Secretary of Education Diane Ravitch, who has argued that the work of educators should be guided by common, stringent standards, much as is done in other fields, such as construction, and that the adoption of such standards can improve practice (Ravitch, 1995). Thus, the first step in SBR is the adoption of carefully framed descriptions of the knowledge and skills

students are expected to master at various stages of their education. These "academic standards" (also called "content standards") become the focal point for changing other elements of the educational system.

SBR systems typically require that academic standards be accompanied by "performance standards" (or "achievement standards," as they are called in NCLB) that indicate the *level* of attainment expected with respect to the content standards. Performance standards are usually established through a judgmental process that identifies one or more cut scores on a test that indicates whether a student has attained a specific level of performance, such as "basic" or "proficient."

In this paper we do not address performance standards in detail, but it is worth pointing out that the process of creating performance standards, setting cut scores, and attaching labels indicating the student's level of performance is fraught with technical difficulties and political controversies (Cronin et al., 2007; Linn, 2003). Most notably, the lack of a common meaning of "proficient" across states has hindered efforts to compare the performance of students in different states, and the frequent confusion of the term "proficient" with the phrase "at grade level" (Rothstein, Jacobsen, & Wilder, 2006) encourages inappropriate inferences about students' accomplishments. In addition, the common practice of comparing state proficiency results with those of NAEP is problematic because of the relatively high threshold NAEP uses for proficiency and because of technical problems in both state and NAEP standard-setting efforts (see, e.g., Pellegrino, Jones, & Mitchell, 1999; Rothstein, Jacobsen, & Wilder, 2006).

Although academic standards are central to the idea of SBR, there is disagreement on what constitutes effective statements of desired student performance. In essence, there are no standards for developing good standards. Most advocates of SBR argue that the standards should be *uniform* and apply to all students, i.e., the system should adopt common expectations for everyone it serves rather than expecting higher or lower levels of attainment from some students. Most advocates of SBR also emphasize that standards should be *challenging*; they should stretch educators' beliefs about what students can learn (see, e.g., Resnick & Resnick, 1992). This view stems in part from a belief that only by adopting challenging standards will we ensure that

students will become productive citizens and the country will maintain its international competitiveness. However, experts do not agree on how statements about student expectations should be formulated to serve their role as the linchpin of SBR. For example, how much should standards reflect our incomplete understanding of typical student learning progressions? How can standards be written to best support the desired educational reforms?

Aligned System Components

The second core element of SBR is the alignment of the rest of the instructional delivery system to promote attainment of the standards. As Clune (2001) characterized it, the central thesis of SBR is that greater alignment or coherence among the policies that affect the content and quality of instruction in schools is the only way to produce greater numbers of schools that will instill high levels of student achievement. Thus, every formulation of SBR calls for alignment of curriculum and materials, instruction, and assessment with the standards. The influence of the idea of alignment is evident today in the large number of textbooks, assessment systems, and professional development materials that are explicitly marketed as being aligned with state standards. Many advocates of SBR promote an even more expansive notion of alignment, recommending that all relevant features of the system be aligned, including pre-service professional development, teacher certification, in-service professional development, after-school programs, and teacher performance appraisals.

Although the concept of alignment is almost universally endorsed by educators, there is no widely accepted method for aligning standards, curriculum, instruction, and assessments, nor is there a consensus on how to determine if existing components are aligned. All approaches use expert judgment in some form, but methods differ in terms of the size of the conceptual units ("granularity") that are compared, the dimensions on which they are compared, and the extent of agreement that is necessary for components to be considered aligned. Too often the process of alignment consists of merely matching each element from one source (e.g., the test) to a similar representation in another source (e.g., the standards).

Student Assessments

The third major component of SBR is the inclusion of assessments of student achievement. Public and professional attention to test scores has been growing since the establishment of NAEP in the 1960s, and in the 1970s tests started being linked with consequences for individual students (Koretz & Hamilton, 2006). Spurred in part by the landmark report, *A Nation at Risk,* issued in 1983, the linking of consequences to test scores continued to grow throughout the 1980s. The role of testing in SBR was also influenced by the idea of "measurement-driven instruction" (Popham, 1987)—that is, the notion that assessment can and should shape instruction—which led to experimentation with innovative forms of assessment that would be sensitive to high-quality, cognitively challenging instruction. The SBR movement has fueled further growth in large-scale assessment, accompanied by efforts to develop data systems to track student progress.

There is more research evidence to support specific decisions about assessment than there is to inform the development of standards, yet there are still unresolved issues about assessment that pose challenges for SBR. Perhaps the greatest challenge is the tension between "richness" and "efficiency." Student achievement can be measured in many ways, including oral presentations, group projects, written essays, structured performance tasks, standardized multiple choice tests, and short constructed-response questions, to mention but a few of the possible types of assessment. Each type of assessment provides somewhat different insights into student understanding, and different combinations of types can better serve distinct purposes, e.g., diagnosis of student errors, progress in learning, instructional improvement, mastery to a given level, transfer of skills, etc. Thus, there is a rich palette of assessment options educators can use to learn about student knowledge and skill. At the same time, large-scale, standards-based reform demands a high volume of assessment geared toward a concise, quantitative summary of performance. SBR demands timely judgments about every school or every student, which puts a premium on efficient assessment. Policymakers have to balance the desire for assessments that provide richer information against the need for assessments that provide more succinct, timely information.

Decentralized Authority

Although nothing in the label "standards-based reform" suggests the need for changes in governance, many advocates believe that improvement in outcomes will only occur if those closest to the teaching and learning process have more authority to change procedures and customize practice to meet the needs of students. Thus, SBR often includes efforts to decentralize authority over the means while maintaining centralized authority over the ends of public education, i.e., SBR includes "restructured educational governance to enable local teachers and schools to decide upon the specific instructional programs they would use to achieve the standards" (Massell, Kirst, & Hoppe, 1997). Such changes in the allocation of responsibility for key curricular and instructional decisions have also been advocated under the names of "school restructuring" and "site-based management."

Policymakers who mandate SBR have frequently emphasized the trade-offs inherent in this new way of managing the education process—teachers, principals, and other educators at the local level receive increased autonomy over their day-to-day work in exchange for increased responsibility for promoting specific outcomes. As will become clear later in the section on the effects of SBR, the actual degree of autonomy experienced by local educators may not be as great as implied in the official policies because of the need for educators to tailor their decisions and actions to the specific goals embodied in the standards and assessments. In other words, the goal of alignment may temper the extent of local autonomy.

Coordinated State and District Support

Smith and O'Day (1991) are credited by many with first characterizing the elements now associated with SBR, particularly the focus on aligning all components of the education system. (They described their vision as one of "systemic reform.") Their model also emphasizes coordinated support services from districts and the state. For example, technical assistance services would be reformulated to focus on helping schools overcome obstacles to the attainment of standards. Training and consultation would not be centrally planned but would respond to identified deficiencies in local student performance, with the goal of helping schools figure out

how to move students ahead. The emphasis on support reflects a recognition that simply telling educators what is expected of them many not be sufficient to induce change; externally provided assistance may be necessary. Not all SBR efforts encompass support strategies, but state and district support clearly plays a prominent role in NCLB's system of consequences for schools that do not meet their performance targets.

Accountability

From the beginning, many advocates of SBR shared the view that assessments should be used not only to monitor progress but also to hold educators (and in some cases students) accountable (see, e.g., NCEST, 1992). This perspective was adopted by the framers of NCLB, which includes clear incentives to encourage change: Schools that repeatedly fail to meet their performance targets (called Adequate Yearly Progress or AYP) because students are not attaining proficiency on state assessments face increasingly severe sanctions, including possible reconstitution and takeover. Similar provisions were included in earlier federal legislation and in state systems that predated NCLB, but NCLB ramped up the enforcement of accountability significantly. As we will review later in this paper, evidence suggests that assigning high stakes to test results can have both motivating and corrupting influences on teaching behavior and on test scores.

As this brief overview suggests, SBR does not have a single, commonly understood definition. This vagueness may be expedient, as it is often possible to build political consensus about contentious issues by using language that obscures differences. However, in a given place and time it is difficult to know precisely what elements are present in the minds of those using or hearing the term "standards-based reform." In the next section, we discuss the evolution of SBR over the past two decades, a discussion that reinforces the complex nature of these systems and the ways in which the assumptions and goals of different stakeholder groups helped shape the elements of this reform movement.

History of Standards-Based Reform from *A Nation at Risk* to NCLB

When "standards" and "accountability" are discussed by educators and the public today, the focus tends to be on NCLB's requirements and how these are shaping state policy and local practice. However, the ideas predate NCLB by 20 or 30 years, and both federal and state governments have played an important role in shaping SBR over these past few decades. Although interest in measuring educational outcomes had been growing throughout the 1960s and 1970s, and several states began adopting elements of SBR in the 1970s, many researchers and historians view as a seminal event the publication of *A Nation at Risk* (National Commission on Excellence in Education) in 1983. That document, which used strong and colorful language to deplore the state of American education, led to policy debates about how to raise expectations for both student and teacher performance, and it emphasized the need to monitor student achievement in a systematic way (Wixson, Dutro, & Athan, 2003).

States and districts responded to this policy environment by undertaking a variety of curricular and structural reforms, including raising graduation requirements, offering more advanced courses, and adopting new textbooks or other curricular materials that were intended to improve the quality of instruction. Analyses of the changes that occurred during this time suggest that these reforms failed to produce widespread improvement, in part because they lacked coherence and failed to communicate a common understanding of what content and skills were expected to be taught (Massell, 1994). Several books and reports published around this time, such as Goodlad's influential *A Place Called School* (first published in 1984) documented a lack of clear expectations of student learning objectives that could be used to guide instruction and curriculum.

This concern for coherence and for clear communication of expectations contributed to the growing interest in reforming education through systemwide standards. As discussed earlier, the idea of "systemic reform" was articulated by Smith and O'Day (1991), who described a broad-based approach to reform that included standards for what students were expected to learn; the alignment of other components of the education system, such as assessment and teacher training, to these standards; and a restructured governance approach that emphasized the role of states and

national organizations in the standard-setting process but that delegated authority for decisions about how to meet the standards to local districts and schools. This call for a more systemic approach to improving student achievement provided an impetus for districts, states, the federal government, and several professional education organizations to engage in efforts to promote SBR. The efforts undertaken by these groups are discussed below.

A few aspects of these early SBR initiatives are worthy of specific attention. First, discussions of standards often emphasized the importance of promoting cognitively challenging instruction rather than emphasizing basic skills and factual knowledge. The idea was not simply to set *high* standards that required students to master advanced content, but to create standards that would encourage teachers to engage students in rich, complex, multi-step problem-solving activities and to hone their reasoning skills. Resnick and Resnick (1992), for example, articulated the idea of a "thinking curriculum" that would challenge and engage students and promote the development of higher-order skills and processes from an early age.

As noted earlier, discussions of the importance of going beyond basic skills and factual knowledge were grounded in concerns about the U.S. economy and the competitiveness of U.S. students relative to their peers in other nations. The descriptor "world-class" was often applied to standards in policy discussions to signal the idea that standards developed in this country should reflect what is known about educational expectations and attainment in high-achieving nations that compete with the U.S., and many standards developers used other countries' curricula as benchmarks to ensure that standards were sufficiently rigorous. There was also an equity component to these discussions—the idea that opportunities to engage in cognitively demanding activities should be provided to all students rather than being offered only to certain types of students or in certain types of schools. The equity aspect of SBR was influenced in large part by the civil rights movement (O'Day & Smith, 1993).

A second important feature of the early standards discussions was the role that assessment was intended to play. Reformers recognized that standards would be unlikely to lead to desired changes if implemented in the context of existing high-stakes testing programs that emphasized outcomes that were inconsistent with the standards. They also understood that most existing tests

were not designed to measure the higher-order skills they envisioned. So the standards development movement was accompanied by efforts to change the way student achievement was measured, particularly through the use of assessments that would require students to engage in complex thinking (Resnick & Resnick, 1992). The notion of "tests worth teaching to" led to widespread experimentation with open-ended, hands-on tasks, portfolios, and other forms of performance-based assessment. One of the most prominent and ambitious efforts to develop new assessments was led by the New Standards Project in the early 1990s and engaged teachers across the nation in developing performance-based assessments and rubrics designed to support an ambitious set of standards. As we noted earlier, SBR advocates did not necessarily argue for the attachment of incentives to test scores, but they understood that in systems in which those incentives were in place, it was necessary to ensure a good match between standards and tests.

Federal Initiatives Related to SBR

The beginning of the federal government's role in SBR is typically traced to the President's Education Summit with the Governors, held in Charlottesville, Virginia, in 1989. The meeting was convened by then-President George H. W. Bush and the nation's governors to devise a strategy to improve student achievement and to make U.S. students competitive with their peers around the globe while also promoting greater uniformity across the states. The meeting resulted in a set of six National Education Goals that were intended to guide policy and practice, and two of these goals proposed ambitious outcomes for student achievement in core academic subjects.[1] These goals were reflected in an education plan called *America 2000*, spearheaded by the first President Bush and his Secretary of Education, Lamar Alexander, to develop voluntary "world-class" standards and voluntary national tests (McDonnell, 2005). The bill never became law, but some of its ideas were included in 1994 legislation titled *Goals 2000: Educate America Act* (20 U.S.C. 5801 *et seq.*) and were supported by funding provisions included in the 1994

[1] Goal 3, *Student Achievement and Citizenship*, specified that by 2000 "American students will leave grades four, eight, and twelve having demonstrated competency in challenging subject matter including English, mathematics, science, history, and geography…" Goal 4, *Science and Mathematics*, focused on the international context: "By the year 2000, U.S. students will be first in the world in science and mathematics achievement" (NCEST, 1992, p.2).

reauthorization of the Elementary and Secondary Education Act, called *Improving America's Schools Act* (IASA) (Jennings, 1998).

It is important to note that all of these efforts were significantly shaped by input from governors and other state policymakers as well as from professional organizations; they should not be viewed as purely federal initiatives. This federal-state-professional interaction is evident in the establishment by President George H. W. Bush and the governors of a group to monitor the nation's progress toward the National Education Goals. This group, called the National Education Goals Panel (NEGP) and consisting of governors, administration officials, and members of Congress, in turn called for the creation of a group to advise government officials on whether and how to create a national system of standards and assessments that could be used to promote the goals and measure progress toward meeting them. The resulting group, the National Council on Education Standards and Testing (NCEST), included representatives from the education, business, and policy communities. Its 1992 report, *Raising Standards for American Education,* argued for increased systemwide coherence and alignment, and issued a call for the development of a national system of standards and assessments.

NCEST argued that national standards and tests would serve several purposes. They would provide students and educators with information about what was expected of them; they would give families, policymakers, the business community, and other groups information about how well students were doing; they would help teachers understand students' strengths and weaknesses; and—notably—they could be used to change the education accountability landscape from one focused on rule compliance to one focused on performance (NCEST, 1992). Although this last purpose is not the one that was emphasized in most policy debates at the time, the idea of performance-based accountability has been one of the most enduring and influential ideas from this reform movement.

The system of standards envisioned by NCEST included several components. In addition to recommending an overarching statement to provide a "guiding vision" for each set of standards, NCEST described four distinct types of standards:

- "***Content standards*** that describe the knowledge, skills, and other understandings that schools should teach…"

- "***Student performance standards*** that define various levels of competence in the challenging subject matter set out in the content standards"

- "***School delivery standards*** developed by the states collectively from which each state could select the criteria that it finds useful for the purpose of assessing a school's capacity and performance"

- "***System performance standards*** that provide evidence about the success of schools, local school systems, states, and the Nation in bringing all students, leaving no one behind, to high performance standards." (NCEST, 1992, p.13)

It is noteworthy that the vision of NCEST and other groups was a system that promoted high achievement for every student, consistent with the equity argument put forth by researchers and others who shaped the SBR debate through the 1980s and early 1990s. The "leaving no one behind" language in the NCEST report is a clear precursor to the NCLB legislation.

This description of a system of standards illustrates the broad way in which SBR was conceptualized by NCEST and other groups at the time. In particular, the notion of standards extended beyond a specification of what students were expected to learn, and included attention to what schools were providing students. The definition of *school delivery standards* alludes to the importance of examining school capacity for delivering high-quality instruction. In the Goals 2000 legislation, school delivery standards were renamed "opportunity to learn" (OTL) standards, which emphasize the monitoring of resources provided to students (in the form of materials, instructional practices, and school conditions) to promote student attainment of content standards.

OTL standards may be thought of as a way to retain some attention to educational inputs in the context of an accountability system that is primarily outcomes-based. Rather than focusing primarily on funding, most advocates of OTL standards emphasized the inputs that are most directly linked to teaching and learning, such as the materials and instructional processes used in the classroom (McDonnell, 1995). In Goals 2000, however, states' development of and compliance with OTL standards were described as voluntary, and the law did not provide for clear incentives for states to attend to OTL (20 U.S.C. 5801 *et seq.*); see also Porter, 1995). In a later section we explore some of the reasons for the failure of OTL standards to take hold.

NCEST made several other recommendations related to an SBR system. According to the 1992 report, standards should reflect high, world-class expectations rather than minimum competency; provide direction to schools but not impose a national curriculum; be created through input from a wide variety of stakeholder groups rather than being a project of the federal government; be voluntary rather than mandatory; and be open to refinement over time (NCEST, 1992). Similarly, the assessment system developed to align with the standards should be voluntary and open to change over time, and should include multiple assessments. The voluntariness was in part a concession to the desire of states to retain autonomy over their education systems. The goal, of course, was to achieve universal participation by states so that students and schools in different states would ultimately be held to similar standards.

NCEST recommended that the development of national education standards build upon work that was already being undertaken by some states and by various professional organizations, such as the curriculum standards developed by the National Council of Teachers of Mathematics (NCTM). The mathematics content frameworks developed in California in the 1980s and the 1989 NCTM *Curriculum and Evaluation Standards for School Mathematics* were viewed by the U.S. Department of Education as models of how to create standards that reflected widespread consensus at both the national and state levels (Wixson, Dutro, & Athan, 2003). In the early 1990s, the U.S. Department of Education awarded funds to various groups of educators and scholars to develop voluntary national standards in other subjects, with the expectation that the process would be similar to that used in mathematics, particularly with respect to the engagement of various constituencies (Wixson et al., 2003). The subjects for which standards were to be

developed included English language arts, science, history, civics, geography, foreign languages, and the arts. In addition, the New Standards Project, mentioned earlier, supported standards-development activity and supplemented it with the development of performance-based assessments.

As it turned out, the level of consensus around voluntary national standards anticipated by many standards advocates did not materialize. There was more disagreement about the mathematics standards than was apparent at the time, and other subject areas experienced even greater disagreement. One of the main issues on which standards developers disagreed was the level of prescriptiveness that should be built into the standards documents—whether standards should be viewed as a guide to help educators develop local curricula or whether they should be written using language that was specific enough to eliminate local discretion over curriculum. Although the latter approach would reduce local discretion, it was viewed by some SBR advocates as a way to ensure that all students received instruction consistent with the standards. A related disagreement stemmed from differing philosophies of teaching and learning in the various subject areas. A somewhat oversimplified description of this conflict is that some reformers espoused a view that knowledge in a particular subject could be seen as consisting of a discrete set of skills and content that could be clearly delineated and mastered incrementally (reflecting a primarily behaviorist perspective), whereas others believed knowledge was constructed within the social setting of the classroom and could not easily be assessed outside a specific context or decomposed into separate pieces (reflecting a primarily constructivist perspective).

Reformers also disagreed on what subjects should be the basis for a set of standards—whether, for example, standards should be written for social studies or for sub-disciplines such as history and civics. Within disciplines there were also more specific areas of disagreement, such as the inclusion of evolution in science standards or the recommended use of calculators in mathematics standards. These disagreements became more pronounced as the standards development process expanded to include a variety of stakeholder groups, such as business leaders and parents, and ultimately led to a decision by the governors at the 1996 National Education Summit to continue the state-level standards-development activities that had been

launched in response to earlier federal legislation rather than pursue the national efforts (Wixson et al., 2003).

Meanwhile, a related assessment initiative was headed down a similar path. In the mid-1990s, the Clinton administration proposed the development of "Voluntary National Tests" (VNT) in mathematics and reading that would be based on the NAEP frameworks but that, unlike NAEP, would provide individual- and school-level scores. This proposal generated a great deal of controversy, in part as a result of the plan to impose oversight by a board appointed by the secretary of education, which suggested (especially to conservatives) that the process would be politicized. Some of the areas of disagreement that characterized the standards-development efforts were applicable to the idea of a national test, as well. In addition, probably more so than the national standards development, the proposal for the Voluntary National Tests raised objections about excessive federal involvement in education (Armour-Garb, 2007). The plan to create the test was ultimately rejected by Congress. With the demise of the VNT, states continued to be the primary units responsible for developing large-scale assessments to measure students' performance, as they had become the primary entities creating academic standards.

Standards-based Reform in the States

Even before the federal government scaled back its efforts to promote SBR, the states were playing an important role in the evolution of the reforms. Many states' initiatives predated the federal efforts. States such as California, Kentucky, Maryland, Massachusetts, North Carolina, and Texas had all begun to implement SBR in the 1980s using their own funds. Later, the Goals 2000 legislation funded state efforts to develop standards, and as a result of this funding almost all of the states embarked on standards development if they had not begun this task already (Armour-Garb, 2007). In the following paragraphs, we describe SBR in three early-adopting states to illustrate the evolution of SBR that was occurring incrementally in large numbers of states during this period (Massell, Kirst & Hoppe, 1997). If we had more space, we could tell interesting stories about SBR implementation efforts in many more states, as well as a number of districts that launched their own SBR initiatives. These three brief histories are not intended to be representative but to provide examples of some of the steps states took and some of the

challenges they faced. The summaries draw on the work of the Consortium for Policy Research in Education and others (Goertz, Floden & O'Day, 1995; Massell & Fuhrman, 1994; Fuhrman, 2001).

California[2]

California was one of the first states to enact policies that incorporated key elements of SBR in the 1980s. Initially under the leadership of Superintendent Bill Honig, California adopted new curriculum frameworks that embodied more rich and rigorous content, focusing on "big ideas" and general principles rather than lengthy lists of factual information. For example, the 1992 mathematics framework called for greater attention to mathematical problem-solving and multiple representations of relationships, and was seen as more demanding than the previous state framework. As noted earlier, this framework served as a model for the national standards development efforts later promoted by NCEST. During this same period the state invested heavily in the development of new, more challenging assessments. It replaced the old California Assessment Program (CAP) with the California Learning Assessment System (CLAS). CLAS not only reflected the contents of the new frameworks; it also pushed the boundaries for large-scale assessment by including many extended-response and performance tasks.

California was also a leader in terms of controversy, and many of the efforts undertaken in the early 1990s were modified or eliminated over the next few years. For example, CLAS was abandoned by 1994-95 in favor of a more efficient, commercial, multiple-choice test. However, efficiency was not the primary motivation for the change; CLAS was ridiculed by some parents who objected to writing prompts that strayed too far from traditional content, scoring systems that were unfamiliar, and a decreasing emphasis on basic skills. During this period, the mathematics framework was also challenged by parents, mathematicians, and educators, who questioned its mathematical soundness and demanded more attention to fundamental understanding and computation. The resulting "math wars" led to a compromise revision of the framework a few years later. Nevertheless, the underlying vision of a system built on challenging standards, with aligned curriculum and assessments, remains a core element of the California

[2] This summary is drawn in large part from O'Day (1995).

educational system. Other elements of SBR were adopted by the state prior to NCLB. For example, in 1998, the Immediate Intervention/Under Performing Schools (II/USP) program was enacted to provide support to low-performing schools in the form of external evaluators and other additional resources.

Kentucky[3]

In 1989, the Kentucky Supreme Court overturned the state's system of public education, and the legislature responded by passing the Kentucky Educational Reform Act of 1990 (KERA). While the lawsuit that prompted the court's action dealt with educational finance, KERA was much broader, calling for reforms of curriculum, assessment, and governance, as well. The curriculum and assessment reforms were consistent with the growing national interest in standards-based reforms and high-stakes assessments, and Kentucky's long-term goal that all schools would be proficient predated the similar formulation in NCLB. KERA established six extremely broad learning goals for all students, including communication and mathematics for daily living; application of principles from mathematics, science, the arts, the humanities, social studies, and practical living/vocational studies; self-sufficiency, thinking and problem solving; and the acquisition, connection, and integration of new and old knowledge. Subsequently, the Department of Education elaborated these into 57 Academic Expectations that identified what students should know and be able to do as they progress through schooling. Student performance was measured by a new assessment system called the Kentucky Instructional Results Information System (KIRIS), a novel approach to assessment that included multiple-choice testing, constructed-response questions, and portfolios in the subjects of writing and mathematics. KIRIS was used to hold schools accountable for student progress, and high stakes were attached to changes in each school's biennial accountability index.

Unfortunately, there were problems using the innovative assessments included in KIRIS for accountability purposes. The mathematics and writing portfolios demanded a lot of classroom time and a lot of scoring time. Further, the state had difficulty scoring the mathematics portfolios with sufficient reliability to use the results in a high-stakes context. Efforts were made to

[3] This summary is drawn in large part from Kannapel et al. (2001).

improve the portfolios, but in 1998, the legislature decided to abandon KIRIS in favor of a more traditional assessment, called the Commonwealth Accountability Testing System (CATS) (Catterall et al., 1998). This change to a more traditional assessment is similar to what occurred in California, though concerns about technical quality were probably more important in the Kentucky case. Although there have been many changes since the enactment of KERA, the basic framework of standards, assessments, and accountability has been maintained.

Texas[4]

Educational reforms relating to standards and assessments began in Texas in the 1980s. Early on, the state adopted two minimum competency tests, the Texas Assessment of Basic Skills (TABS) and the Texas Educational Assessment of Minimal Skills (TEAMS). In 1983, Governor Mark White appointed a commission of business leaders (chaired by H. Ross Perot) to recommend educational reforms. The commission's recommendations led to the passage of House Bill 72 in 1984, which included learning standards for student achievement, assessments for teachers, and a new funding formula for schools. The standards took the form of "essential elements" teachers should try to convey in 12 areas of knowledge. In the early 1990s, the Texas Assessment of Academic Skills (TAAS) replaced the minimum competency tests.

In 1992 the legislature tried to address weaknesses in the standards by transforming them into descriptions of what students should know and be able to do. This work was put on hiatus when leadership changed in the statehouse, but it was begun again under a new governor in 1996. These new standards, called the Texas Essential Knowledge and Skills (TEKS), were completed soon thereafter, and they have been the basis for subsequent student assessment. During this same period, under further prodding from the business community, Texas created a student information system to track student attendance, performance, etc. The state also began using test scores as the basis for rewards and sanctions, including the Texas Successful Schools Award System (TSSAS).

[4] This summary is drawn from Achieve (2002), A+ Education Foundation (2003), Massell, Kirst, and Hoppe (1997), and Texas Education Agency.

Thus, at the time of the education summit in 1996, Texas was one of the few states that could claim to have an operational SBR system, with standards, assessments, and incentives. The state has continued revising its policies to promote stronger alignment and higher expectations, such as through the adoption of a new, more difficult testing system, the Texas Assessment of Knowledge and Skills (TAKS), which includes items measuring students' analytical and reasoning skills. Throughout this period, groups like the Business Education Coalition and other voices from the private sector have be instrumental in moving the SBR agenda forward in Texas.

These three state summaries illustrate the complex processes through which elements of SBR were incorporated into state educational systems during the 1990s. Despite this complexity, the speed with which standards-based reform spread among the states during the 1990s was "without historical precedent" according to Massell, Kirst and Hoppe (1997). The authors contend that the rapid uptake of these ideas was due to a coincidence of factors, including the development of model standards by professional organizations like the National Council of Teachers of Mathematics, support for national standards development from the U.S. Department of Education, the Statewide Systemic Initiatives projects launched by the National Science Foundation, and endorsement of these ideas from diverse advocacy groups, including the Business Roundtable, the National Governor's Association, and the American Federation of Teachers. Although the situation in each state was unique, it was common to find alliances among the education and business communities that sustained reform efforts in the face of competing priorities, changing administrations, and other challenges.

The model standards created by professional organizations also helped states tackle the difficult task of developing standards, and the support provided by these bodies made the process easier. These early efforts to implement standards frequently failed to address issues that would later challenge effective implementation, including building capacity among teachers and administrators to present more challenging curriculum, addressing equity concerns—particularly the needs of students with limited English proficiency and students with disabilities—and, in states that took more innovative approaches to curriculum and pedagogy, resolving the objections of small but well-organized Christian and conservative groups.

Standards-based Reform Today

By the early 2000s, every state in the U.S. had adopted a system of standards and assessments and was using this system as an accountability mechanism to promote school improvement, though fewer than one-half of these systems were in full compliance with the IASA standards and testing requirements at the time. Much of the recent SBR activity can be attributed directly to NCLB, which, like previous legislation, requires each state to establish a system of standards-based accountability that includes standards, assessments, and annual targets for performance, but which imposed stricter requirements for testing (e.g., a requirement that all students in grades 3 through 8 be tested annually) and for the creation of performance standards based on proficiency cut scores. Despite these additional requirements, today's state systems under NCLB retain many of the features of previous state SBR systems. Perhaps most importantly, "standards-based reform" has, in the wake of NCLB and other high-stakes testing policies, morphed into what might be called "test-based reform," a system in which educators and others rely primarily on the test rather than the standards to communicate expectations and to inform practice.

Beyond NCLB: SBR and High Schools

It is important to recognize that NCLB is not the only factor affecting state and local SBR efforts. In particular, there have been extensive efforts to promote increased accountability and uniformity of expectations at the high school level, but most of these efforts take place outside the realm of what is required under NCLB (especially since NCLB imposes its heaviest testing requirements on the elementary and middle school grades). Under NCLB, states are required to develop standards in core subjects at the high school level and to administer tests to all students in mathematics, English/language arts, and science in at least one high school grade. Many states have taken their own initiative to impose additional accountability requirements at the high school level, most notably through high school exit exams and end-of-course exams that students must pass in order to receive a diploma. In 2007, approximately two-thirds of high school students were being educated in states that administered high school exit exams, and this number is expected to increase (Center on Education Policy, 2007b). The stated purposes of these exams vary, but in most states the primary purpose is to evaluate student mastery of standards or

curriculum frameworks (Center on Education Policy, 2007b), which suggests that state policymakers have adopted a standards-based approach to preparing students for what comes after high school.

A few states have adopted policies requiring end-of-course (EOC) exams; for such tests states develop standards and assessments tied to the content of specific courses rather than covering the broad range of content that might be covered across high school (e.g., an EOC exam might focus on introductory chemistry compared with the NCLB science tests, which might cover all of the science content that might be taught at the 10th-or 11th-grade levels). One prominent initiative to develop EOC exams is being led by Achieve, Inc. as part of its American Diploma Project (ADP). The ADP initiative, which Achieve is undertaking in partnership with the Education Trust and the Thomas B. Fordham Foundation, involves a consortium of 33 states working with Achieve to "align high school standards and assessments with the knowledge and skills required for success after high school" (http://www.achieve.org/files/AboutADP.pdf).

The impetus behind ADP reflects the arguments that led to the SBR movement years ago: concerns about poor preparation of graduates; lack of equity in both expectations and outcomes; and a perception that the solution to these problems is to develop high-quality, challenging standards that are aligned with the skills and knowledge needed for success after high school and measure attainment of those standards with aligned assessments. One of the ADP's recent initiatives is the Algebra II Consortium, a group of 14 states that joined together to pilot a new, common Algebra II EOC exam. The exam is intended to be aligned with a set of standards that would be used to guide course development across the participating states. Its purposes are to "to improve high school Algebra II curriculum and instruction; to serve as an indicator of readiness for first-year college credit-bearing courses; and to provide a common measure of student performance across states over time (http://www.achieve.org/node/842). This effort is likely to be expanded to other courses and states over time, and reflects one of the latest initiatives in the now decades-old SBR movement intended to raise student performance by specifying clear expectations and measuring progress toward them.

Aspects of the SBR movements can also be seen outside the K-12 system. Efforts are under way to create measures of performance at the university level (Klein et al., forthcoming) and, at the other end of the spectrum, to measure academic outcomes of preschool children (Rothman, 2005). This paper focuses on the K-12 sector, but it is important to acknowledge the widespread influence that the SBR movement is exerting on education outside that grade range.

What Happened to OTL Standards?

One of the central ideas in earlier SBR efforts—the inclusion of opportunity to learn standards—has all but disappeared in current policy debates. McDonnell (1995) outlined a number of political, practical, and technical challenges to implementing and monitoring progress toward OTL standards, including the costs and measurement issues associated with trying to collect the necessary data from schools and districts. There were also concerns about excessive state or federal control over what schools do and about the costs of equalizing school and district offerings if OTL information demonstrated inequalities. As McDonnell noted shortly after Goals 2000 passed, "The political history of OTL in the NCEST and Goals 2000 deliberations and its technical limitations suggest that, at least in the short-term, the uses of OTL standards for policy purposes will be primarily hortatory. They will help define a vision of equitable, high-quality schooling, and will serve to persuade educators and the larger community to buy into that vision" (pp.317-18).

OTL never did take hold as an accountability mechanism, but the idea of trying to identify the features of high-quality schools and teachers and find ways to help unsuccessful schools and teachers adopt these features is still central in the accountability debate. In particular, under NCLB schools are directed to provide high-quality teachers for all students. In addition, the consequences for failing to meet AYP targets are intended in large part to further improve students' access to high-quality instruction, either through opportunities provided to individual students via supplemental education services and transfers to higher-performing schools, or through technical assistance and other interventions that attempt to improve the conditions in low-performing schools. In addition, one of the fundamental assumptions behind OTL standards—that holding students accountable for performance requires assurance that they did

have the opportunity to learn the material—has been a guiding idea behind a number of lawsuits concerning adequacy of education resources (O'Day & Smith, 1993).

How SBR Is Shaping Today's Debates about Public Education

Despite the extensive changes that have characterized federal and state SBR initiatives over the past few decades, a few core ideas have endured and are likely to continue to shape these reforms in the future, regardless of what happens with specific legislation or programs. In addition, a few ideas that are likely to influence the next generation of reforms have emerged more recently in the wake of widespread SBR. These ideas, which have affected not only the development of specific SBR programs and policies, but also have shaped the broader public debate on what constitutes high-quality education at the K-12 level, include the following:

- ***Emphasis on better information for improvement.*** Standards are primarily intended to provide information about what is expected of teachers and students, and they are typically linked with assessments that provide information about students' attainment of those expectations. If it can be analyzed, assimilated, and acted upon, this information has potential value for improving educational practice. The recent growth in availability of tools and services to promote "data-driven decision making" in education is an outgrowth of the SBR movement and is influencing how teachers and administrators carry out their day-to-day work and their strategic planning.

- ***Attention to academically challenging content.*** Despite the rancor that characterizes some of the debates surrounding how to teach math or reading, there is broad agreement on the value of ensuring that students participate in cognitively challenging activities. International comparisons of curriculum and achievement test scores have provided fodder for these discussions, as have numerous reports documenting the need for students to develop strong problem-solving skills to be successful in college or the workplace (see, e.g., Partnership for 21st Century Skills, 2008). These results have been used by policymakers as justification for raising standards and curricular requirements. The current system, with its reliance on standardized tests that often emphasize easy-to-

measure skills and knowledge, may hinder rather than facilitate efforts to adopt more cognitively challenging approaches, but the goal is still evident in policy discussions and in some states' efforts to improve the quality of their standards and tests.

- ***Importance of promoting equity***. Inequities in opportunity continue to plague the education system, but the SBR movement has led to increased efforts to remedy the problem by directly addressing the content of instruction and the equity of outcomes rather than simply worrying about funding. Both the requirement to calculate AYP for significant subgroups and the requirement to measure the English language development of students with limited English proficiency are manifestations of equity concerns in the context of NCLB.

- ***Significant role of the private sector in facilitating alignment of standards, assessments, and other components of the education system.*** The emphasis on ensuring that everything from textbooks to professional development is carefully designed to promote specific instructional goals has been a focus of school and district improvement efforts and has influenced not only public education but private providers of goods and services, including test developers and textbook publishers. For example, textbook publishers have marketed their texts as being aligned with specific states' standards and tests, and in some cases have published state-specific editions. They also provide resources such as benchmark assessment systems that are intended to align with state standards. While the availability of aligned materials and support systems from a single supplier may make life easier for districts trying to figure out how to raise test scores, it may also exacerbate the score inflation problem by promoting test-focused activities that drive out other curricular content. The growing injection of the private sector into educational delivery systems also raises concerns about conflicts of interest that may stem from publishers' desires to demonstrate that their materials raise scores on state tests (the same tests that those publishers are sometimes involved in developing). At the same time it is important to acknowledge that the private sector has been the source of a large number of innovations that have the potential to transform the delivery of education, such

as tutoring and distance-education programs that allow students to access help and course content outside of school.

- ***Primacy of tested outcomes.*** Perhaps most significant is the extent to which the measurement of outcomes using achievement tests drives education policy and practice today. Concerns about specific effects of high-stakes testing systems are likely to lead to changes in the details of those systems, but all signs point toward a continuing emphasis on evaluating the quality of education in terms of student attainment of knowledge and skills as measured by tests.

SBR has clearly influenced the actions of educators at all levels of the system. The broad themes listed above suggest that how we think about public education has changed as a result of SBR, but they do not provide a clear indication of whether the quality of instruction and the achievement of students have improved because of this movement. The next section summarizes the research on standards and how they influence educational practices and outcomes.

What Do We Know About the Effects of SBR?

Despite the near-ubiquity of standards and other elements of SBR today, the question of whether SBR has benefited students is still hotly debated. Concerns about possible adverse consequences of SBR date back to the movement's beginnings. Porter (1994) captured some of the main fears expressed by critics: "Those who believe that national standards in education, accompanied with student performance assessments, are not an appropriate strategy for educational reform, fear that standards will trivialize education and de-skill teaching by being too prescriptive. They fear that one-size-fits-all approach of national standards setting will create an inflexible delivery system that will be incapable of coping with differences between poor schools and rich schools, able students and weak students, well-prepared teachers and teachers teaching out-of-subject." He also noted that "virtually all of the arguments, both for and against standards, are based on beliefs and hypotheses rather than on direct empirical evidence" (p. 427). Although a large and

growing body of research has been conducted to examine the effects of SBR, the caution Porter expressed in 1994 about the lack of empirical evidence remains relevant today.

The focus of research on SBR has evolved along with the movement itself. In the early 1990s, much of the research examined the quality of the standards and the standards creation process. With the nationwide adoption and enactment of SBR and the growing availability of large-scale data on school practices and student achievement, the emphasis has shifted to investigating associations among SBR, school and classroom practices, and student achievement. Some research has also examined other components of the education system—such as professional development—that were intended to be aligned with the standards; we do not review that research here.

The ambiguity associated with defining SBR creates challenges for identifying the relevant research base. A number of studies have focused on instructional practices or curricula that are aligned with specific standards (such as the NCTM mathematics standards), but most of these studies did not systematically examine the effects of the SBR context in which the use of standards was embedded (see, e.g., Le et al., 2006). A review by Lauer et al. (2005) discusses much of this research and finds that standards-based curricula and standards-aligned instructional practice are both associated with positive achievement outcomes. Although much of the reviewed research focuses on so-called "reform-oriented" approaches that may not align well with most states' current standards, this research suggests that the publication of standards can, when accompanied by supports such as professional development, exert a small influence on what and how teachers teach. A number of these studies have examined SBR initiatives that predated today's NCLB accountability-focused SBR and that often lacked aligned assessments (see, e.g., Clune, 2001). In addition, there have been studies of the effects of SBR in other countries (e.g., Bishop, 1998; OECD, 2007). All of this research is relevant to thinking about SBR, but for the purposes of understanding the likely effects of SBR today, we focus on research that examines standards-based reforms that involved high-stakes assessment in the United States.

High-quality research on the effects of SBR is difficult to conduct for a number of reasons, including challenges associated with measuring practices and outcomes, obtaining a

representative sample and adequate data, setting up the needed experimental design to study the causal effect of SBR, and addressing the diversity in the assessment programs and accountability policies in different states and districts. Thus, the amount of rigorous analysis is limited. Nevertheless, the existing body of research provides some valuable information in three areas: studies of the quality of standards themselves, studies of links between SBR and school and classroom practices, and studies of links between SBR and student achievement.

Evaluations of Content Standards

Because standards are intended to shape what teachers do in the classroom, the quality of the standards—including breadth, clarity, and emphasis on various pedagogies or content areas—is critical for promoting improved instruction. Reviews of recent evaluations of the quality of standards suggest three overarching themes.

First, there continues to be a lack of consensus regarding the criteria that should be used to evaluate content standards, reflecting, in part, the disagreements that characterized the standards-development process discussed earlier. Discrepancies in the results of evaluations conducted by different groups of researchers can be attributed primarily to the different criteria used to evaluate the standards and to the subjectivity involved the application of these criteria (Archbald, 1998).

The language describing the features that characterize high-quality standards tends to be fairly vague; in science, for example, a National Research Council committee concluded that standards should be "clear, detailed, and complete; reasonable in scope; rigorous and scientifically correct; and built around a conceptual framework that reflects sound models of student learning" (National Research Council, 2006, p.2). Most evaluators of standards would not disagree with this description, but the process of translating these descriptors into specific criteria and assigning weights to each criterion have led to disagreements in ratings among organizations such as Achieve, the American Federation of Teachers (AFT) , the Council for Basic Education (CBE), and the Thomas B. Fordham Foundation (Valencia & Wixson, 2001). Differences in the importance attached to these dimensions of quality in the evaluation process reflect diversity of

understanding about the purposes and roles of standards and diversity of values regarding teaching and learning. Despite the discrepancies in evaluation results, however, the existing ratings of state standards provide information about the quality of standards from multiple perspectives that might be helpful in efforts to improve standards.

Second, great variability exists in the quality of standards across states, regardless of what criteria researchers used to evaluate standards. Moreover, the quality of standards differs by subject and grade. For example, according to the evaluation report released by the AFT in 2008, the percentage of standards across all grade levels and subjects which reached the criteria for "strong standards" defined by the AFT ranged from 0% for seven states to 100% for one state (Glidden, 2008). The evaluation conducted by the Fordham Foundation in 2006 also showed substantial variation in the quality of standards across states (Finn, Julian, & Petrilli, 2006).

Third, although states have been revising their standards over the years, the results of the various reviews suggest that there is still much room for improvement in the quality of content standards (Finn et al., 2006). Finn et al. reported that 37 states have revised at least one of their state standards since 2000. Yet the average quality of standards across all subjects remained "C-minus," according to the criteria defined by the Fordham Foundation. Glidden (2008) reported that only 16 states had more than 75% of their standards meeting the definition of "strong standards" defined by the AFT.

Taken together, these evaluation findings reveal the challenges inherent in trying to judge the quality of standards. Arguably the most important test of quality is whether the standards promote high-quality instruction and improved student learning, but as we discuss later, there is very little research to address that question. Moreover, the disagreements among researchers who have evaluated standards reflect much broader disagreements over what constitutes high-quality instruction and curriculum, so even if it were possible to conduct rigorous studies of the effects of standards it is unlikely that this work would lead to a consensus regarding how to create good standards.

SBR and School Practices

SBR can only attain its ultimate goal of improving student performance if it leads to improvements in educational practices. Studies of changes in school practices accompanying SBR should not only help us to understand the mechanisms through which SBR influences learning, they may also help us interpret changes in student achievement objectively (how much "teaching to the test" actually occurs) and develop a more complete picture of the outcomes of SBR. Many of the elements of SBR have been studied independently, and there have been a few studies of SBR as a comprehensive system. For example, there is some research on how the adoption of standards, *per se*, or the alignment of standards with curriculum influences school practices or student outcomes. The research on the influence of standards has tended to focus on reform-oriented approaches to curriculum and instruction, which are not always relevant to today's state SBR systems (see, e.g., Lauer et al., 2005). By contrast, the link between high-stakes tests and student outcomes has been studied extensively, and testing and incentives have become important elements of SBR, so much so that in some places tests with high stakes have been implemented without many of the other SBR elements. As a result, this review draws mostly on research examining high-stakes testing.

Because neither SBR nor high-stakes testing has been implemented in a way that allows for experimental investigations of effects (i.e., there are usually no "control" groups that include comparable students or schools not subjected to SBR and testing, and random assignment of schools or students to SBR versus a non-SBR condition is not feasible), most of the research on the impact of SBR and testing on educators' practices is based on surveys, interviews, observations, or correlational methods. Although these approaches do not support strong causal conclusions, the common findings that have emerged from a large number of studies across various SBR contexts lend credibility to the following generalizations.

High-stakes testing does seem to affect school practices; in particular, the implementation of such tests has been accompanied by efforts to improve the quality of curriculum and instruction

(Center on Education Policy, 2006; Goertz, 2007; Hamilton, 2003; Hamilton et al., 2007; Lane et al., 2002; Stecher, 2002). Teachers, as well as school and district administrators, have reported taking a number of steps to improve school performance in response to SBR initiatives that involve high-stakes testing. School and district actions include adopting programs to address the needs of low-performing students, aligning curriculum and local assessment programs to state standards, increasing the use of data to improve decision making, and providing professional development and other supports (e.g., curriculum coaches) to promote improved teaching. At the classroom level, teachers report spending more time on instruction, working harder, seeking ways to improve their practices, aligning their instruction with standards, and using interim test results to individualize instruction. We would expect the same effects to follow from the implementation of SBR policies that include accountability for test results.

However, teachers and administrators have reported other reactions to high-stakes tests that raise concerns about effects of such tests on the quality of instruction. Some of these responses suggest that schools take shortcuts to improve performance on high-stakes tests without necessarily improving student learning more broadly, a class of responses that is sometimes referred to as "test preparation" but that includes a wide variety of activities. For example, schools reported focusing on certain subjects to improve the overall school-level achievement in a short period of time or reassigning more qualified teachers to the tested subjects or grades. Schools were also found to manipulate the composition of test-taking population by assigning low-performing students to special education, retaining them in non-tested grades, allowing more absences on test days, granting exemptions from testing demanded by parents, or increasing dropout rates (see Hannaway & Hamilton, 2008, for a review of these practices). Fortunately, more recent SBR and high-stakes testing policies have been designed to mitigate some of these problems, such as by mandating a minimum test-participation rate.

At the classroom level, teachers have reported reallocating instructional time away from non-tested subjects in order to provide more instruction in tested subjects. Reallocation of instructional time was also found across tested and non-tested content and skills within subjects: teachers reported devoting more attention to material that is included in the test and skipping or de-emphasizing material that is not tested (Hamilton, 2003). Although proponents of SBR may

view the reallocation from non-tested to tested material as a positive outcome, particularly in the context of SBR systems with high-quality assessments that comprehensively cover the standards, there remains a risk that students will miss important content. This is particularly true in light of studies that indicate a lack of perfect alignment between tests and standards and a tendency for tests to focus on standards with lower cognitive demand that under-represent the more challenging standards (Rothman et al., 2002).

This reallocation effect may be exacerbated by state efforts to make it easier for teachers to "teach to" the standards and state tests. Some states have published guidelines to help teachers figure out which standards are likely to be tested; in Pennsylvania, for example, teachers have access to Assessment Anchors that specify the "eligible content," which enables them to focus on the parts of the standards that will be tested and ignore those that won't. These findings reveal one of the difficulties inherent in trying to judge the benefits of standards and testing—whether one views reallocation from social studies to math or from non-tested math concepts to tested math concepts as desirable or undesirable is largely a value judgment that is likely to be influenced by one's views regarding what outcomes schools should promote as well as by information about current student performance.

Teachers report other actions in response to high-stakes testing that in some cases may improve test scores without improving the underlying achievement the test is intended to measure. These actions range from focusing on specific test item styles and formats (which is common) to outright cheating (which seems to be less common but is difficult to measure) (see Koretz & Hamilton, 2006, for a discussion of various strategies for raising test scores). For example, several large-scale surveys indicated that many teachers designed their classroom presentations to resemble the format of the test, used instructional materials that mirrored the format of state accountability tests, drilled students on the same format of questions as those that appeared in state tests, and changed the sequences in which they presented topics to accommodate the testing schedule (Stecher, 2002). The desirability of some of these changes is related to test quality: In some cases, complex, performance-based assessment appears to lead to greater emphasis on problem solving in the classroom (Lane et al., 2002), which may be beneficial, whereas extensive practice with the multiple-choice format would generally be viewed as less desirable.

Qualitative research on teachers' responses to high-stakes tests provides additional evidence of instructional changes in response to the idiosyncrasies of test formats and accountability metrics. In a study of responses to NCLB, teachers reported actions such as removing novels from the reading curriculum in favor of short passages similar to those on the state test (Hamilton et al., 2007). Additionally, research shows that reallocation of instructional time is reported to occur not only across subjects or content, but also between different student groups. In particular, in response to NCLB's focus on the percentage of students scoring at the proficient level, teachers and other school staff reported increasing their focus on students who had the greatest potential to move from below to above the proficiency threshold (often referred to as "bubble kids"; see Booher-Jennings, 2005; Hamilton et al., 2007; Pedulla et al., 2003). Of course, this response is largely a function of a design feature of NCLB rather than something that would occur simply as a result of SBR, but it points out the importance of carefully designing the SBR system to create the desired incentives while avoiding undesirable ones.

Although most of the behavioral changes discussed so far could be viewed as either negative or positive depending on their specific features and context, an unambiguously undesirable response to high-stakes testing is cheating. Research has revealed several examples of cheating, such as providing the actual test items in advance, rephrasing test questions for students, leaving related materials in view during test administration, providing answers to students, allowing longer test time, or changing students' answers before scoring (Hannaway & Hamilton, 2008). These practices threaten the validity of results on accountability tests, and lead to worries about educators' conduct and its impact on students.

In summary, evidence drawn from studies of standards, alignment, high-stakes testing programs, and broader SBR policies suggests that federal and state SBR policies have exerted a significant effect on the actions of public school teachers and administrators. At the same time, there is evidence that teachers maintain a great deal of autonomy even as they struggle to meet the mandates of NCLB and other SBR policies (Hamilton et al., 2008), and the responses of teachers and administrators have not always been consistent with what SBR advocates envisioned. Moreover, some researchers have found that while high-stakes testing does influence practices in

ways that reflect attention to the content of the test, most testing programs have failed to induce deeper pedagogical change or fundamentally alter the *ways* teachers deliver instruction (Diamond, 2007; Firestone, Mayrowetz & Fairman, 1998). SBR has led to some beneficial changes in school practices at both the organizational and classroom levels, but it has also led to responses that are less clearly desirable, and SBR has not always produced the kinds of instructional improvement that advocates hoped for. These findings raise concerns about the generalizability of achievement gains (as we discuss in the next section), equity of educational opportunities, and ethical conduct of educators.

Effects of SBR on Student Achievement

Improved student achievement is widely viewed as the primary goal of SBR, but for a variety of reasons it is difficult to measure the impact of SBR on student achievement. One challenge is the variation in state standards and accountability tests that makes it difficult to compare achievement gains across states. Another challenge is the existence of various concurrent education reform efforts, which make it difficult to establish a direct causal relationship between SBR and achievement gains within a state or across states. In addition, achievement trends might be affected by state and district characteristics that influence both achievement and the extent to which accountability policies were enacted, which makes it more difficult to assess the pure effect of SBR on achievement.

Statistical methods can be used to minimize influences from other factors and remove confounding effects of state and district characteristics from estimates of testing effects; yet, great caution should be taken when making inferences from achievement gains to actual learning. It is important to remember that a test contains only a sample of questions designed to represent all the possible questions that could be used to assess students on a certain subject. As a result, the extent to which increases in scores on a sample of test questions support valid arguments regarding a corresponding improvement in overall achievement varies widely, depending on the representativeness of actual test questions and the scope of the inference (Koretz, 2005).

A further concern is that many of the actions administrators and teachers take in response to SBR (discussed above) can lead to a phenomenon called *score inflation,* which, as the name suggests, produces inflated gains in scores. Score inflation "refers to increases in scores that are not accompanied by commensurate increases in the proficiency scores are intended to represent" (Koretz, 2003, p. 9). To assess the extent of score inflation it is useful to examine test-score trends on a test that measures the same subject area or content as the high-stakes test but that does not have high stakes attached to it.

Research has consistently demonstrated that gains on high-stakes tests do not generalize to low-stakes tests (see Hamilton, 2003 for a review). In one recent study, Jacob (2007) demonstrated this phenomenon by examining differences in students' performance trends on state accountability tests and the NAEP tests between 1990 and 2002. Among the four states examined—Texas, North Carolina, Arkansas, and Connecticut—Jacob found the rate of improvement on state accountability tests far outpaced that on the NAEP tests. In addition, using item-level mathematics test data from Texas, Jacob analyzed possible reasons for differential performance trends on state accountability tests and the NAEP tests. He found that differences in the skills and question formats across exams might have contributed to differences in students' performance on the two types of tests, especially for 4^{th} grade. Specifically, the state accountability test focused more on basic arithmetic skills and lower-level cognitive tasks, and had lower difficulty levels than the NAEP test. The existence of score inflation poses great challenges for valid interpretation of test results and objective evaluation of the effect of SBR on student achievement.

Despite these challenges, available research does provide some insight about the association between SBR and student achievement. First, studies show that by and large students have improved their achievement on both high- and low-stakes tests since the 1990s. For instance, Jacob (2007) examined achievement trends among 4^{th} and 8^{th} grade students in Texas, North Carolina, Connecticut, and Arkansas on state accountability tests and the NAEP tests from 1990 to 2003. He found that students in all four states had improved their performance on both types of tests, especially in Texas and North Carolina, although the gains cannot necessarily be attributed to SBR. The Center on Education Policy has examined achievement trends in every

state since the enactment of NCLB and has found that achievement on state tests and the NAEP tests have both increased on average since the law's enactment. Although gains on the NAEP tests were not as large as gains on state accountability tests, the fact that NAEP trends tended to move in the same direction as state test-score trends suggests the gains are not exclusively due to score inflation (Center on Education Policy, 2007a, 2008). However, as with the other findings described above, it is impossible to determine whether they are due specifically to SBR. Moreover, there is no evidence that gains have been greater since NCLB than they were during the period preceding NCLB (when the initial SBR requirements of IASA were in effect; see, e.g., Fuller et al., 2007).

Where achievement gains have occurred, researchers have found consistent patterns in scores by subject and grade. Achievement gains in mathematics tend to be greater and more robust than those in reading, and gains at the elementary and middle school level are on average greater than those at high school level. These differences are observed in state accountability tests and on NAEP (Center on Education Policy, 2008; Jacob, 2007).

A more direct approach to examining the influence of SBR is to study relationships between accountability policies and student achievement. Though the evidence is not conclusive, several studies suggest a positive relationship. Carnoy and Loeb (2002) examined the association between the strength of state accountability and achievement gains students made on the NAEP mathematics tests in 1996-2000. They reported that 8th-graders in states with strong accountability systems improved more than their counterparts in states with weak or nonexistent accountability systems on the NAEP mathematics test between 1996 and 2000. Jacob (2005) reported that students in the Chicago Public Schools substantially increased their mathematics and reading achievement after the implementation of an accountability policy in 1996-97. Hanushek and Raymond (2005) also found a positive relationship between the implementation of accountability policy and achievement gains from 1992 to 2002 in 42 states. In addition, international research has found positive relationships between accountability exams and achievement levels (Bishop, 1998; Bishop, Mane, & Bishop, 2001).

While there have been some studies that did not find such a positive relationship (Amrein & Berliner, 2002a, 2002b; Nichols, Glass & Berliner, 2006), the earlier two studies have been criticized on methodological grounds (Braun, 2004; Hanushek & Raymond, 2003; Rosenshine, 2003). The later study, however, raises legitimate questions about the relationship. As a whole, the research suggests that there is a positive relationship between high-stakes testing policies and scores on achievement tests, though the findings are far from conclusive and the magnitude of any relationship remains unknown.

Although these findings suggest that student achievement has improved since the enactment of SBR initiatives, caution is warranted in drawing any conclusions regarding the effectiveness of these policies. Despite positive trends on both high- and low-stakes test scores, gains on high-stakes tests have been much greater than those on the NAEP or state-administered low-stakes tests, and evidence of score inflation is widespread. Moreover, there are concerns about the differential effect of SBR on achievement by performance level: Evidence from NAEP suggests differences in gains for students at different achievement levels, with initially high-scoring students making much less progress than low-performing students (Thomas B. Fordham Institute, 2008).

There are also differences in the magnitudes of gains between reading and mathematics and between elementary and secondary grade levels, despite the fact that both subjects and grade levels are influenced by SBR policies. Additionally, there is no consistent finding regarding whether SBR contributes to narrow achievement gaps among ethnic groups. Moreover, although NAEP scores provide a good alternative to state accountability tests for examining achievement gains, they have limitations; for example, NAEP scores are only available for a sample of students and for certain grades, and the lack of stakes on NAEP may reduce student motivation to perform well. These questions and measurement issues need to be addressed before any final conclusions can be made about the effect of SBR on achievement.

In summary, nationwide achievement gains since 1990s on both state accountability tests and NAEP tests indicate that students have improved their mathematics and reading achievement over the last two decades. The achievement gains that accompanied the introduction and

development of SBR suggest a positive link between SBR and achievement. However, many questions critical to fully understanding the effect of SBR on achievement remain unaddressed, including how to measure true achievement gains or identify the direct contribution of SBR. There is a need for new research methods that can address these challenges, and for including in accountability systems specific design features—such as an audit test that can be used to validate gains on the high-stakes tests—to provide better information about student learning.

How Has Federal Policy Affected Educator Practices and Student Achievement?

The research reviewed above cannot distinguish the effects of federally mandated SBR programs from the effects that would have occurred if states or districts had initiated the programs in the absence of federal requirements. Nevertheless, the investigations of responses to the various federal SBR initiatives indicate that despite the nation's tradition of local control over public education and the fact that federal funding represents a relatively small proportion of the total funds used by schools, the federal government's actions can play a substantial role in shaping education policy and practice at the state, district, and school levels, particularly if it adopts policies that include enforcement mechanisms. In particular, a comparison of NCLB with earlier incarnations of the Elementary and Secondary Education Act (ESEA) illustrates the strong influence that strictly enforced, test-based accountability policies exert on educators: Although SBR was a part of the 1994 legislation, the absence of strong enforcement mechanisms resulted in much weaker responses than we have observed since the enactment of NCLB.

The research also shows that the specific details of the SBR and accountability policies affect what happens in schools and classrooms. For example, the NCLB requirement that school performance be measured according to the percentage of students scoring proficient led some states with pre-existing accountability systems to modify their measurement strategies, regardless of whether the existing strategies were working well in that state, and it also appears to have influenced teacher and principal decisions about allocation of time, effort, and other resources. Similar consequences follow directly from other features of NCLB, such as the definition of a highly qualified teacher, or the rules regarding inclusion of subgroups. Of course, some states might have adopted some of these features in the absence of NCLB's requirements,

but in all likelihood, states would have come to different decisions based on their prior experiences, local context, and other factors. Although there is still a great deal of variability in state systems, as exemplified by the widely varying meanings of "proficient" performance, the federal law has clearly led to common features across states and illustrates the strong influence that federal initiatives can have not only on state policy but on what is taught and learned in classrooms.

What We Have Learned

At the beginning of this paper we described a vision of standards-based educational reform advocated by many educators in the 1980s and 1990s. After reviewing the evidence that has been collected by researchers over the past decade or more, it is impossible to determine whether this vision of standards-based reform can achieve its widely supported goals. The lack of evidence about the effects of SBR derives primarily from the fact that the vision has never been fully realized in practice. Yet, although there has not been a rigorous test of the efficacy of SBR, there has been considerable research on the implementation of SBR-inspired reform and the impact of various components of SBR. The evidence suggests a number of tensions and challenges that should be at the forefront when policymakers debate the future of SBR. We discuss these issues next, and then turn our attention to the future of SBR in general and the federal role in particular.

When Tests Have High Stakes, Standards May Take a Back Seat

One of the most consistent findings from research on educational testing and on performance measurement in sectors other than education is that when tests have high stakes they shape behavior. Research also indicates that under these circumstances teachers are more likely to pay attention to tests than standards, potentially undermining some of the benefits of SBR. The over-attention to tests is especially likely to occur when standards are perceived as too broad or numerous to cover. Some states have responded to teachers' concerns about overly ambitious standards by publishing information about which standards are likely to be tested, an action that may help teachers focus but that also increases the likelihood that unmeasured standards will not be taught.

If standards are intended to serve as the primary source of information about what is expected of students, it is critical that the assessments used to measure progress toward those standards are well aligned to the standards, both in content and in cognitive demand. In addition, research shows that standards in subjects that are not included in the assessment program are more likely to be ignored, so if there is a desire to promote standards-based instruction in untested subjects, some means of measuring offerings or outcomes in those subjects (such as additional tests or monitoring of instruction) need to be put in place. Policymakers should also carefully consider whether the presumed benefits of high-stakes testing outweigh the drawbacks, and determine whether the accountability consequences associated with test scores should be modified or reduced to counteract the tendency to focus only on tested content. One of the lessons that emerges from a comparison of responses to the 1994 IASA and the 2002 NCLB legislation is that attaching stakes to scores is likely to be necessary to spur states, districts, and schools to fully adopt SBR, but that stakes can also lead to adverse consequences that should be carefully monitored.

Existing Tests Do Not Adequately Measure All Standards

This problem stems in part from a fundamental shortcoming of tests—they can contain only small samples of knowledge and skill from a much larger domain. The need to select a subset of content for inclusion can lead to tests that do not represent standards well. This sampling dilemma plays out in a number of ways. Some of the discrepancies between tests and standards stem from time and budget constraints that lead states to rely heavily on multiple-choice items, and from the fact that certain types of skills and knowledge are easier to measure through large-scale, paper-and-pencil tests than others. The alignment study by Rothman et al. (2002) found that tests provided better coverage of standards with low levels of cognitive demand than standards that were more cognitively challenging. Given the tendency of educators to focus on tested content, this misalignment could have negative effects on the quality of curriculum and instruction. As we discuss later, recent innovations in technology and psychometrics may offer some solutions to this problem.

When Strong Sanctions Are Attached to Specific Measurable Outcomes, Practices Tend to Become Distorted

This consequence is so well known in the literature that it has a name—Campbell's Law (Campbell, 1975) —which states "The more any quantitative social indicator is used for social decision making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor." As we document in this paper, there is ample evidence that NCLB has led schools to ignore many adopted state standards in favor of material that is tested. Not only are standards for social studies, science, music, art, and physical education being shortchanged, but so are standards in math and reading that do not lend themselves to efficient, large-scale testing.

The research also shows that teachers are devoting a great deal of time to focused "test preparation" and other practices designed specifically to raise test scores with little regard to their broader impact. These "distorted" behaviors are due, in part, to the fact that test content and format has become predictable from one year to the next. As a result, test-based accountability, at least as implemented by many states today, is undermining some important aspects of SBR. As noted earlier, standards-based reform has become "test-based reform," and this represents a challenge to the future of SBR.

SBR Allocates Responsibility in Ways that Can Conflict with Traditional Educational Governance

Most advocates of SBR recommend allocating responsibility for curricular and instructional decisions to educators who are proximate to the students' learning environment, i.e., teachers and principals. For example, Smith and O'Day (1991) clearly specified a division of responsibility in which the task of specifying standards was centralized and the job of making instructional decisions was decentralized. This allocation of responsibility mimics the trade-off made in some businesses where the central office sets long-range goals and production decisions are made by staff close to the process. However, problems can arise when those who monitor school progress are not satisfied with the actions taken locally (or more often with the results achieved). Thus,

the district superintendent may choose to impose curricular decisions on schools, the city mayor may wrest control of schools from elected Boards of Education so she/he can require grade-level exams or particular textbooks, or the state department of education may send independent experts to offer advice. While the increased autonomy promised by standards- based reform may sound good in theory, in practice it can be difficult to create a satisfactory allocation of responsibility for decision making. As far as we can tell, this is an unresolved challenge for SBR.

Alignment and Autonomy May Become Competing Goals

One specific challenge to autonomy stems from another key feature of SBR—alignment. The idea of aligning all components of the educational system to standards was a focus of early SBR efforts, and continues to characterize district and school responses to SBR in the context of NCLB. Common district strategies for improving school performance include developing curriculum maps to ensure alignment between curriculum and standards; adopting interim assessments that resemble the state test, and providing professional development to promote standards-based teaching. These kinds of actions are likely to help promote standards-based instruction, but at the same time may lead to a reduction in the kind of autonomy that was viewed by early reformers as central to the idea of SBR, and by more recent accountability advocates as a sort of compensation for accepting greater accountability, particularly if they promote a highly prescriptive approach to instruction. Even if no one is telling teachers what pedagogical strategies to use, the confluence of professional development, curriculum materials, assessments, data systems, and other resources is likely to influence not only what is taught but how it is taught. Moreover, many teachers have expressed concern that the need to cover the content standards forces them to make certain instructional decisions, such as moving through material quickly rather than delving more deeply into specific topics.

The effect of standards on autonomy may also be influenced by features of the standards themselves—although heavily prescriptive standards may be easiest to interpret and translate into instruction, they are also likely to constrain teacher choice in how to teach. Policymakers who develop SBR systems, and district and school leaders who set local policy in response to them, need to weigh the relative importance of preserving teacher autonomy and ensuring

coverage of standards and promote a system that reflects their priorities. If the decision is to adopt a system that will strongly influence what and how teachers teach, the SBR policies should be accompanied by clear guidance regarding curriculum and instruction (such as pacing guides and sample lesson plans). If policymakers are more concerned about preserving local autonomy, they might want to issue less-prescriptive guidance, but they should still provide professional development and other support to encourage high-quality instruction.

Federalism Continues to Pose Challenges for SBR

The history of SBR is marked by shifts in the roles of the federal government, the states, and various professional organizations. NCLB was adopted during a time when most policymakers believed responsibility for setting standards and measuring progress toward them belonged with states. This led to the awkward situation in which proficiency varies widely among states, as does the percentage of schools and districts that are making adequate yearly progress. Many policymakers are annoyed at the discrepancies among states, and many educators are frustrated by the feeling that NCLB imposes unfair demands based on state performance levels. There seems little justification for having different standards for basic literacy and numeracy across the states, when graduates are going to be competing for employment in a regional, national, or international economy. There have been calls for national standards to put all schools and districts on an even footing; however, such moves upset those who wish to maintain the dominant state and local role in education.

Recently, as problems with NCLB proficiency levels became apparent, many reformers' reluctance to embrace a national model was replaced with enthusiasm for such a model, and this change has been observed across the political spectrum. Organizations such as the Thomas A. Fordham Institute and the Nelson A. Rockefeller Institute of Government have convened experts and published reports to explore national standards, and professional groups such as the National Association of Secondary School principals have released position statements advocating a more national approach to standard setting. It is important to distinguish between national and federal efforts. Most advocates of a national approach do not believe the responsibility should fall to the federal government; instead they advocate more voluntary cooperation among states. One

approach is the development of state collaboratives, such as the New England Comprehensive Assessment Program, in which Vermont, New Hampshire, and Rhode Island have been working together to develop common standards, assessments, and proficiency levels.

On a broader scale, Achieve, Inc. is involved in several efforts to create common standards and measures across states, including the American Diploma project discussed earlier. The organization recently released K-12 mathematics benchmarks that indicate the content that should be covered in various grade ranges and provide some illustrative model course sequences but leave room for local decisions about curriculum, such as whether to teach high school math using a traditional Algebra I-geometry-Algebra II sequence or a series of integrated mathematics courses. It seems clear to us that SBR cannot function effectively if thresholds for proficiency or the content of state tests differ widely among states; it is not as clear to us whether this situation is best addressed through voluntary or mandatory leveling of standards. However, SBR must address the question of reasonable, equitable standards nationwide.

Where Do We Go from Here?

This review of the history and impact of SBR leads us in two directions. The evidence summarized in the paper leads to a set of modest recommendations to improve the existing system. However, we think the original vision of SBR can be implemented in a more faithful way by dramatically rethinking the current system and introducing more extensive changes. We present the incremental changes first and then suggest the more visionary alternative. In neither case, however, should SBR be viewed as having the potential to solve the nation's education challenges all by itself; rather it should be considered as one prong of a multi-pronged approach that might include changes in school funding, school choice systems, and teacher preparation programs, to name a few.

Although improved SBR does not require federal involvement, there are ways that the federal government might contribute to this effort. The history of SBR suggests that strong federal mandates are likely to be met with resistance, and researchers and policymakers disagree on the

desirability of involving the federal government in SBR. At the same time, history also indicates that the federal government can exert a strong influence on state and local practice, particularly when working in collaboration with state governments and private entities. Several specific activities might be particularly effective as strategies to promote more consistent, higher-quality SBR across states. The problems that arose from earlier efforts to develop national standards and assessments suggest a need for caution, but the growing recognition of the limitations of the 50-state approach and a corresponding recognition of the role the federal government can play make this a good time to experiment with federal strategies. The following five recommendations represent incremental efforts that could be taken by the federal government to promote more effective SBR policies.

1. Improve standards and assessments. The federal government could convene experts representing states and private groups to help establish standards for standards (i.e., to develop criteria that outline the desired features of standards), clarify the advantages and disadvantages of framing standards more narrowly or more broadly, link standards with what is known about the skills and knowledge needed for success after high school, promote cognitively challenging instruction, and help states benchmark their standards against those of other, high-performing states or other nations. Similar groups could be commissioned to gather and disseminate information about best practices in large-scale assessments, and could provide technical assistance to individual states or groups of states as needed. Many states lack the capacity to develop high-quality assessments on their own, and the extensive testing demands imposed by NCLB have forced some states to scale back their innovative assessment practices due to time and funding constraints. Input may be particularly needed at the high school level, where there is currently a lack of high-quality, curriculum-aligned assessments[5].

Such groups could also gather and disseminate information about effective use of data for decision making. This is an activity in which many districts and schools are engaging as a way to promote school improvement, but currently most educators and administrators lack a good understanding of how to use data effectively. As a result of the trends discussed earlier, "data-

[5] Another paper prepared for this CEP series (Popham, 2008) discusses the federal role in assessment in much greater detail.

driven decision making" often translates into use of scores from tests that are designed to resemble the state test rather than encompassing a broader notion of data. In all of these efforts, the government's role should be to convene and coordinate rather than create the standards or tests. Nongovernmental organizations such as Achieve could play an important role in these activities.

Another way to promote higher-quality standards and assessments is through the awarding of grants to states or other entities for development and research. For example, the Department of Education could award grants to foster the development of mixed, balanced, assessment alternatives that serve multiple purposes: diagnosis, instructional improvement, and monitoring/reporting. States or private organizations that received these grants would be expected to share the results in a way that would allow other states to benefit from the work.

2. Develop accountability indices that create more effective incentives by addressing the shortcoming that research has identified in current accountability metrics. Evidence from high-stakes testing programs suggests that educators' actions are influenced not only by the content of the tests but by the ways in which scores are reported and synthesized into an accountability index such as the "meets AYP" designation. These indices need to be consistent with the goals of public education: If getting all students to "proficiency" is sufficient, the current use of percent-proficient as an outcome may be appropriate, but if we care more about ensuring that all students grow, regardless of where they started, a growth model that relies on a scale score rather than a proficiency cut score may be preferred. While it is not possible to achieve complete consensus on these goals, future SBR efforts should take them into consideration explicitly and should carefully weigh the pros and cons of any given approach. For example, growth models require students to be tested at successive grade levels on a test that is designed to measure growth on a specific construct, which imposes limits on the use of some innovative assessment approaches. These trade-offs require careful analysis of the various options, and decisions should be informed by input from the many groups that have a stake in how accountability systems are designed.

3. Experiment with alternative SBR approaches to enhance our understanding of the effects of specific features. The Department of Education could create a funding mechanism to support experiments to explore different, innovative approaches to SBR. For example, although many conceptions of SBR emphasize autonomy, we currently know relatively little about the effects of granting autonomy or what the right balance is between autonomy and prescriptiveness. Autonomy can be influenced through relatively minor changes, such as the adoption of a pacing plan that helps to ensure that instruction is aligned with standards but that may reduce teachers' ability to adapt to student needs. Autonomy could also be affected by broader governance changes. The government could help investigate questions about autonomy by allocating funds for experimental studies in which local educators were given different degrees of autonomy. Experimentation could also be conducted to explore the impact of variation in the strength of incentives or the use of group versus individual incentives.

4. Continue to use and broaden NAEP, which not only provides a means of comparing student performance across states, but has also allowed us to monitor achievement on subjects not typically included in state accountability systems. Although NAEP is certainly not a perfect measure of student achievement, and has been criticized in particular for its lack of incentives (which may affect whether students take the test seriously) and for the fact that it is not completely aligned with any state's standards, it is the only assessment that is administered to representative samples of students in all states. It can therefore play a role in tracking the progress of reforms, in conjunction with state accountability tests and other measures such as Advanced Placement exams. Moreover, NAEP occasionally administers assessments in subjects other than those that are the focus of state accountability policies, and data from these assessments can be used to monitor what is happening to student achievement in non-accountability subjects. NAEP can also serve as a testing ground for alternative assessment strategies, as it has in the past.

5. Support the evaluation of SBR efforts. One of the primary responsibilities of the federal government should be to ensure ongoing collection of evidence demonstrating the effects of the policies, which could be used to make decisions about whether to continue on the current course or whether small adjustments or a major overhaul are needed. Such evaluations should attempt to

measure the full impact of SBR on the broad goals of the education system, including graduation rates, course-taking patterns, and student learning in subjects not included in the high-stakes testing system. This type of evidence could also be a way to generate support and buy-in from educators and the public. In essence, this activity would constitute an evaluation of the efforts of the states to implement SBR.

The evaluation role could be carried out by an agency within the U.S. Department of Education, but the evaluation would probably be more credible if conducted by an external organization (or consortium of organizations). This effort could include an ongoing audit of state standards and assessments. It should also include reviews of existing research on SBR. These reviews should be conducted by panels of independent experts with experience evaluating and synthesizing research findings, and the results should be publicly disseminated in a timely manner. The Department of Education has emphasized the importance of using interventions that are supported by rigorous research, and SBR policies should be subject to the same level of research scrutiny. There would be great value in providing both a summary of what is known about SBR implementation and outcomes and guidelines for best practices based on the evidence. Existing organizations, such as the What Works Clearinghouse, might undertake this task, or a new approach might be tried.

A More Comprehensive Vision of SBR for the Future

This exploration into the history and effectiveness of SBR makes it clear that the original, comprehensive vision of SBR has never been given a full trial. In most instances, both practical and political considerations led jurisdictions to implement limited versions of SBR. In some cases, advocates believed that only a few components were needed to produce dramatic results; in other cases compromises were necessary to build a strong enough coalition to implement a change. In this final section we revisit the key SBR elements listed at the beginning of the paper, and we present ideas for rethinking some of them, with the goal of promoting increased opportunity for all students to receive high-quality instruction that prepares them for a variety of postsecondary paths.

Standards

While the focus of most SBR systems has been on mathematics and reading, states have adopted academic content standards for other subjects including history/social studies, science, arts, and physical education. The existence of these standards reflects a broad consensus that well-educated citizens need to know more than mathematics and reading. While some subjects might be more important than others, and some topics more essential within a subject, to completely exclude any of these standards from SBR is inconsistent with the purpose of public education. The evidence suggests that standards that are not tested are less likely to get taught than those that are tested. States and other entities responsible for the development of SBR systems should therefore acknowledge the importance of all of these standards while establishing priorities for the purpose of monitoring and reporting. In addition, there is growing interest in promoting skills and attributes that are viewed as necessary for success in today's (as well as tomorrow's) economy (see, e.g., Partnership for 21st Century Skills, 2008) but that are not necessarily measured by existing state tests. These include complex problem-solving and communication skills, as well as non-academic attributes, such as perseverance and ability to work in teams.

It is likely to be impossible to monitor all the standards every year for every student, but it is certainly possible to monitor some standards from each domain, to select a set that reflects the priorities stakeholders place on the various outcomes, and to discourage excessive emphasis on the standards that most easily lend themselves to large-scale, standardized testing. As we discuss below, one approach to promoting attention to a broader set of standards would be to use monitoring mechanisms other than tests. For example, while today's testing systems may discourage reading teachers from assigning material other than short passages, a separate monitoring system could be put in place to ensure that schools are providing access to instruction that uses novels and other extended prose materials. The selection of standards to monitor could be changed over time to achieve balance and to avoid encouraging a narrow focus. These suggestions raise some technical challenges, but many of these can be overcome if the will is present to do so. The entire process should be informed by input from a variety of stakeholders, including not only the business and industry groups that have traditionally promoted SBR, but

others whose voices have not always been well represented, including teachers, parents, and, ideally, students themselves.

It is also necessary to acknowledge individual differences among students and revisit the worthy twin goals of setting more challenging standards and applying them uniformly to all students. Equity has been an underpinning of SBR since its early days, and high expectations are essential, but we also need to acknowledge that we cannot eliminate variability in performance: All students will not achieve to the same level or succeed in the same subjects, nor will they all be attracted to the same fields of study. This recognition should influence both curriculum expectations and (as we will discuss below) the measurement of progress. Schools should provide options that lead to a variety of postsecondary paths, including preparation for the many rewarding careers that do not require a four-year degree, and the development of standards should be informed by these options.

For some subjects, such as mathematics and reading, there is evidence that many of the same skills required for successful transition to higher education are also relevant to careers that require less than a four-year degree (ACT, 2006). In these cases, standards should reflect these key skills and areas of knowledge, and their relevance to postsecondary options should be clearly communicated to students. At the same time, expectations should be set in a way that permits flexibility in the paths students pursue. It is critical to recognize the tensions between, on the one hand, holding high expectations and offering high-quality, cognitively challenging instruction to all students, and on the other, providing multiple pathways through school to facilitate individual interests and skills. The latter, if taken to extreme, can result in rigid tracking policies that systematically deny opportunities to some groups of students. But requiring all students to take high-level courses that they may not need, while possibly closing off opportunities for them to pursue fields in which they are interested, such as fine arts or technical education, is also potentially harmful to students' career opportunities and perhaps to their engagement with school.

Alignment

In recent years discussions of alignment have focused primarily on the match between state standards and assessments, as well as on the alignment of local curriculum with state standards and assessments. Going back to the original conceptions of SBR, standards should be the focal point for all aspects of the education system, including teacher preparation, resource allocation, curriculum adoption, and instruction, and many of these have received insufficient attention of late. More effort needs to be devoted to aligning teacher preparation (familiarity with standards and how to teach them), principal preparation (provision of instructional leadership), and in-service professional development (meeting the needs of individuals in classroom settings) with standards. Similarly, support structures (learning communities, teacher mentors, distinguished educators) and resource allocation (allocating teachers among schools, materials among classrooms, etc.) should be assessed in terms of their alignment with and contribution to the learning of the standards. The primary challenge to implementing these alignment activities is resource limitations, but some of the nation's largest school districts have made strong efforts to implement some aspects of these aligned systems and can provide lessons regarding what does and does not work.

Perhaps most important is the provision of resources and support that directly address instruction, such as sample lesson plans that illustrate high-quality instruction focused on a specific standard or set of standards. One area that deserves special attention is the use of data for instructional decision making. "Data-driven decision making" has become an extremely popular strategy for promoting school improvement, but to a large extent it emphasizes data from tests that are designed to resemble the state test (see, e.g., Marsh, Pane, & Hamilton, 2006). Data use, therefore, becomes another source of narrowing and test preparation. Resources to help educators think more broadly about data—including the collection, analysis, and application of data—could provide another means to improve standards-based instruction while minimizing adverse consequences.

The kinds of supports and resources provided should reflect, to the extent possible, broad consensus regarding the goals of the SBR system with respect to autonomy. Some reformers

desire a highly prescriptive system, which strongly influences what and how teachers teach; for them, alignment entails clear guidance regarding curriculum and instruction (such as detailed pacing guides). Others believe the best outcomes will be achieved if decisions are made locally and more autonomy for instruction is preserved. They might want to issue less-prescriptive guidance but still provide professional development and other support to encourage high-quality instruction. This tension is hard to resolve. Some districts compromise by offering more autonomy to schools and teachers who are successful, while imposing a more structured version of alignment on those who are not.

Efforts to develop and implement these supports should involve collaboration between the public and private sectors. The latter has been the source of a number of innovations that may improve students' access to high-quality instruction while introducing significant cost savings. Examples include intelligent tutoring systems to supplement in-class instruction; online portals that allow teachers and students to communicate and share information such as grades and homework; and video-based sample lessons that can be viewed by any teacher at any time. It would be short-sighted for federal or state governments or school districts to attempt to develop such resources on their own, but funding and regulatory mechanisms need to be in place to support the involvement of creative, innovating reformers from the private sector (see Hess, 2008, for detailed discussion of what it would take to engage the private sector productively).

Assessment

The shift from standards-based reform to test-based reform illustrates the profound effect that the content and format of large-scale tests can exert on instruction and learning. This is a concern because the skills measured by the test are a subset (and often a very limited subset) of the skills contained in the standards. One of the reasons for the discrepancies between many states' standards and assessments is that certain types of skills and knowledge are easier to measure through large-scale, paper-and-pencil tests than others. If standards are intended to serve as the primary source of information about what is expected of students, it is critical that the assessments used to measure progress toward those standards are well aligned to the standards, both in content and in cognitive demand.

Recent innovations in testing, particularly those that take advantage of advances in information technology for administering and scoring assessments, could contribute to states' efforts to develop more valid but cost-effective measures of their standards, particularly complex problem-solving and reasoning skills, and could help states tailor assessments to individual students' needs (such as through computerized adaptive testing). Another factor that could help promote better tests is a de-emphasis on the need for annual individual student scores. For instance, if tests could be administered using a matrix sampling approach such as that used for NAEP—i.e., different students take different items so that a large number of items can be included but individual student scores cannot be calculated reliably—it would be possible for states to include a wider range of item types that might do a better job of capturing the range of skills and knowledge embodied in the standards.

Reducing the amount of testing required, such as by allowing states to test only in specific grades such as 5, 8, and 11 (a common practice prior to NCLB) could also free up time and funds to support higher-quality testing. As with matrix sampling, this approach would eliminate states' ability to track the performance of individual students from one year to the next, so it is important to weigh the relative costs and benefits of these changes in light of policymakers' goals for the system.

One of the primary benefits of an assessment system that includes a broader range of content and item formats is a reduction in the predictability associated with most state tests. If it were impossible for teachers and students to predict the specific content that would be covered on the test, or how the items would be formulated, the prevalence of test-focused instruction and score inflation might diminish.

Accountability

Another way to address problems stemming from test-driven reform is to develop a broader set of indicators for use in school reporting and accountability systems. Adopting non-test outcome measures not only provides the public with more complete information on how schools are

performing, it also is likely to reduce the pressure to focus exclusively on tested material, and therefore may mitigate some of the problems discussed earlier. As suggested in the discussions of standards and assessments above, it would not be necessary to measure everything every year, but instead the system could measure a diverse set of outcomes in a way that keeps the data-collection burden to a reasonable level and minimizes the predictability inherent in today's test-based accountability systems. A variety of indicators to measure student outcomes could be incorporated into school performance reports, and some districts are already experimenting with such systems by supplementing the existing state test scores with scores on AP or college admissions tests, graduation rates, and enrollment in postsecondary institutions.

Furthermore, the indicator system does not have to rely exclusively on outcomes. Supplementing outcome measures with a set of process indicators that provide information on what schools are doing could further support the goal of reducing excessive emphasis on tested content, and could provide information on school practices and opportunities that are likely to promote important but hard-to-measure outcomes such as civic-mindedness, teamwork, and creativity. For example, indicators could be developed to measure such factors as student course-taking (e.g., access to a rigorous, college-preparatory curriculum), arts and athletic offerings (including extracurricular options), and provision of instructional support services.

Indicator systems could also be designed to allow schools to provide information on innovative programs such as mentorship or internship programs, and could therefore potentially encourage innovation rather than fostering conformity with a narrow set of criteria. Moreover, indicators could be developed to provide information on the educational progress and opportunities of groups of students for whom the state tests do not always provide valid information, such as English language learners and students with disabilities. Emphasis should be on factors that have strong evidence of association with positive student outcomes and on those that reflect local as well as national goals for public education. These indicators could be used not only for reporting and accountability purposes but also to identify schools in which further investigation and intervention are needed.

In a sense, the adoption of such measures would mark the return of opportunity to learn standards to the SBR policy debate. Porter (1995) discussed the use of OTL standards as performance indicators and noted that they would not only provide information about school offerings, but could be used to create a vision to guide schools toward more effective practice. Collecting this kind of information can be costly, but improvements in state and district data systems are making some of this information more easily available than it was during the early OTL policy debates, and not all information would need to be collected every year (Porter, 1995; Hamilton & Stecher, 2006).

It will also be important to avoid creating a system that requires local educators and administrators to devote extensive time to data collection, so a small number of relatively easy-to-collect indicators would probably be desirable. And any indicators that are used for accountability purposes should be designed so that the information collected from individual schools is comparable and not easily subject to corruption. For example, to the extent possible, indicators of outcomes such as graduation rates and postsecondary attendance should rely on a central database and should use common formulas for calculating these measures. Although the creation of a broad-based indicator system will require time, resources, and a thoughtful debate about how to measure each construct and how to use the data produced by the system, the idea is worth pursuing as the nation moves toward the next phase in the SBR policy environment.

Accountability should also be extended beyond the school-focused, sanction-heavy model that is currently widespread. This is a limited view of accountability, and SBR might benefit from more expansive thinking about who is held accountable for what. In particular, accountability should entail rewards as well as (or perhaps more than) sanctions and should give credit for accomplishments as a way to motivate similar behavior in others. These rewards could include bonuses or salary increases but should not be exclusively financial and should not be tied solely to test-score gains. Another aspect that deserves further exploration is a more balanced approach to accountability that entails some form of stakes for students, parents, as well as educators. The *Standards for Educational Accountability Systems* published by the Center for Research on Evaluation, Standards, and Student Testing (CRESST) discusses the problems associated with systems that impose stakes on educators but not on students, and calls for greater consistency in

incentives across groups (Baker, Linn, Herman, & Koretz, 2002). A more differentiated approach to accountability does not necessarily require that students be held back in a grade or denied a diploma as a result of performance on a single test (a practice that violates another commonly accepted testing standard that warns test users against using a single test score for high-stakes decisions), but that the system as a whole should identify ways to motivate all involved parties.

An additional direction to consider is to expand accountability beyond the regulatory model that has dominated recent policy debates and to empower all parents and students to contribute to debates about what outcomes are most valued. One mechanism to increase accountability to families consists of the choices made available to parents in some locations through charter schools or other offerings. If parents are given the right to walk away from schools that aren't meeting their needs—including schools that are perceived to focus excessively on narrow, test-specific content—incentives to focus on tests could be offset, at least in part, by a competing set of incentives.

Choice-based reforms require a number of conditions to work effectively, one of which is broad dissemination of high-quality information on a broad range of school practices and outcomes as well as the availability of good alternatives for all students (which has generally not been the case for districts implementing NCLB's transfer provisions). Information dissemination would be facilitated by the development of broader indicator systems, discussed above. Choice takes many forms, and the research on its effects is mixed, but experimentation with choice in the context of a broader set of indicators could contribute to our understanding of how comprehensive reform of the public education system can lead to success, and could promote the kinds of innovations that are likely to be needed to keep pace with the changing demands of the workplace in the future.

Moreover, it is widely recognized that schools are not the sole contributors to students' academic achievement and that involvement of families and social service organizations is critical to promoting student achievement. Broadening the notion of accountability to address the responsibilities of other parties could go a long way toward ensuring that all students have the

resources they need, both in and out of school, to achieve at high levels. This notion of shared accountability has taken hold in many parts of the education policy and research communities, as evidenced most recently by a statement disseminated by the "Broader, Bolder Approach to Education" initiative (see http://www.boldapproach.org/statement.html). We do not have the space here to examine these accountability options in depth, but we want to suggest that the time is ripe for exploration of innovative approaches.

Conclusion

SBR has been shown to be a powerful lever for change at all levels of the education system. Some of the hopes of early reformers have been at least partially realized, but at the same time, some of the fears of early critics have materialized. This paper attempted to summarize broadly what is known about SBR and to lay some groundwork for thinking about a revised federal role in promoting high-quality standards and assessments. Ongoing efforts to improve the knowledge base and disseminate what is known to decision makers at all levels will be critical for developing SBR systems that promote high-quality teaching and learning.

References

A+ Education Foundation. (January 2003). *Achieving Excellence: Education Reform in North Carolina and Texas.* Montgomery, AL: Author.

Achieve, Inc. (June 2002). *Aiming Higher: Meeting the Challenges of Education Reform in Texas.* Achieve's Benchmarking Initiative. Washington DC: Author.

ACT. (2006). *Ready for college and ready for work: Same or different?* Available at http://www.act.org/research/policymakers/pdf/ReadinessBrief.pdf (last accessed 9/26/08).

Amrein, A.L. & Berliner, D.C. (2002a). High-stakes testing, uncertainty, and student learning. *Education Policy Analysis Archives 10*, no.18). Retrieved July 16, 2008 from http://epaa.asu.edu/epaa/v10n18/

Amrein, A.L. & Berliner, D.C. (2002b). *The impact of high-stakes tests on student academic performance: An analysis of NAEP results in states with high-stakes tests and ACT, SAT, and AP test results in states with high school graduation exams.* Retrieved July 16, 2008 from Educational Policy Studies Laboratory, Education Policy Research Unit: http://edpolicylab.org

Archbald, D.A. (1998). *The reviews of state content standards in English language arts and mathematics: A summary and review of their methods and findings and implications for future standards development* (Report ED-98-PO-038).Washington DC: National Education Goals Panel. Retrieved July 2008 from http://eric.ed.gov/ERICDocs/data/ericdocs2sql/content_storage_01/0000019b/80/19/a6/ec.pdf

Armour-Garb, A. (2007). *Intergovernmental approaches for strengthening K-12 accountability systems.* Albany, NY: Nelson A. Rockefeller Institute of Government.

Baker, E.L., Linn, R.L., Herman, J. L., & Koretz, D. (2002). *Standards for educational accountability systems* (CRESST Policy Brief 5). Los Angeles: Center for Research on Evaluation, Standards, and Student Testing.

Bishop, J.H. (1998). The effect of curriculum-based external exit exam systems on student achievement. *The Journal of Economic Education*, 29, 2, 171-182.

Bishop, J.H., Mane, F., & Bishop, M. (2001). *Secondary education in the United States: What can others learn from our mistakes?* (Working Paper 01-07) Ithaca, NY: Center for Advanced Human Resource Studies.

Booher-Jennings, Jennifer (2005). Below the bubble: 'Educational Triage' and the Texas accountability system. *American Educational Research Journal*, 42, 2, 231–68.

Braun, H. (2004). Reconsidering the impact of high-stakes testing. *Education Policy Analysis Archives*, *12*(1). Retrieved July 16, 2008 from http://epaa.asu.edu/epaa/v12n1/

Campbell, D.T. (1975). Assessing the impact of planned social change. In G. Lyons (Ed.), Social research and public policies: The Dartmouth/OECD Conference. Dartmouth College Public Affairs Center.

Carnoy, M. & Loeb, S. (2002). Does external accountability affect student outcomes? A cross-state analysis. *Educational Evaluation and Policy Analysis*, 24, 305-31.

Catterall, J., Mehrens, W., Ryan, J., Flores, E., & Rubin, P. (1998). *Kentucky instructional results information system: A technical review*. Frankfort, KY: Office of Education Accountability, Kentucky General Assembly.

Center on Education Policy. (2006). *From the capital to the classroom: Year 4 of the No Child Left Behind Act.* Washington, DC: Author.

Center on Education Policy (2007a). *Answering the question that matters most: Has student achievement increased since No Child Left Behind?* Washington, DC: Author.

Center on Education Policy (2007b). *State high school exit exams: Working to raise test scores*. Washington, DC: Author.

Center on Education Policy (2008). *Has student achievement increased since 2002? State test score trends through 2006-07.* Washington, DC: Author.

Clune, W.H. (2001). Towards a theory of standards-based reform: The case of nine NSF state-wide systematic initiatives. In S.H. Fuhrman (Ed.), *From the capitol to the classroom: Standards-based reform in the states.* (pp.13-38). Chicago, IL: The University of Chicago Press.

Cronin, J., Dahlin, M., Adkins, D., & Kingsbury, G.G. (2007). *The proficiency illusion.* Washington, DC: Thomas B. Fordham Institute.

Diamond, J.B. (2007). Where the rubber meets the road: Rethinking the connection between high-stakes testing policy and classroom instruction. *Sociology of Education,* 80, 285-313.

Finn, C.E., Julian, L., & Petrilli, M.J. (2006). *The state of state standards 2006*. Washington, DC: Fordham Foundation.

Firestone, W.A., Mayrowetz, D., & Fairman, J. (1998). Performance-based assessments and instructional change: The effects of testing in Maine and Maryland. *Educational Evaluation and Policy Analysis*, 20, 2, 95-113.

Fuhrman, S.H. (2001). Introduction. In S.H. Fuhrman (Ed.), *From the capitol to the classroom: Standards-based reform in the states* (pp. 1-12). Chicago, IL: The University of Chicago Press.

Fuller, B., Wright, J., Gesicki, K., & Kang, E. (2007). Gauging growth: How to judge No Child Left Behind? *Educational Researcher,* 36(5), 268-78.

Glidden, H. (2008). Common ground: Clear, specific content holds teaching, texts, and tests together. *American Educator*, Spring, 13-19.

Goals 2000: Educate America Act of 1994. 20 U.S.C. 5801 *et seq*.

Goertz, M.E. (June 2007). Standards-based reform: Lessons from the past, directions for the future. Paper presented at the Conference on the Uses of History to Inform and Improve Education Policy.

Goertz, M.E., Floden, R.E., & O'Day, J. (1995). *Studies of education reform: Systemic reform, vol I: Findings and conclusions*. New Brunswick, NJ: Rutgers, the State University of New Jersey, Consortium for Policy Research in Education.

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

Hamilton, L.S. (2003). Assessment as a policy tool. *Review of Research in Education*, 27, 25-68.

Hamilton, L.S., & Stecher, B.M. (2006). A better way: Measuring charter school success and failure. In R.J. Lake & P.T. Hill (Eds.), *Hopes, fears, and reality: A balanced look at American charter schools in 2006* (pp.61-71). Seattle, WA: University of Washington, National Charter School Research Project.

Hamilton, L.S., Stecher, B.M., Marsh, J., McCombs, J.S., Robyn, A., Russell, J., Naftel, S., & Barney, H. (2007). *Implementing standards-based accountability under No Child Left Behind responses of superintendents, principals, and teachers in three states.* Santa Monica, CA: RAND.

Hamilton, L.S., Stecher, B.M., Russell, J.L., Marsh, J.A., & Miles, J. (2008). Accountability and teaching practices: School-level actions and teacher responses. In B. Fuller, M.K. Henne, & E. Hannum (Eds.), *Strong state, weak schools: The benefits and dilemmas of centralized accountability*

(Research in the Sociology of Education, Vol. 16, pp.31-66). St. Louis, MO: Emerald Group Publishing.

Hannaway, J., & Hamilton, L. (2008). *Effects of Accountability Policies on Classroom Practices*. Washington, DC: The Urban Institute.

Hanushek, E., & Raymond, M. (2003) Accountability works after all. *Education Next*, 3, 3. Retrieved July 16, 2008 from http://www.hoover.org/publications/ednext/3347781.html

Hanushek, E.A., & Raymond, M.E. (2005). Does school accountability lead to improved student performance. *Journal of Public Analysis and Management*, 24, 297-327.

Hess, F.M. (2008). *The future of educational entrepreneurship:Possibilities for school reform.* Cambridge, MA: Harvard Education Press.

Jacob, B.A. (2005). Accountability, incentives and behavior: The impact of high-stakes testing in the Chicago Public Schools. *Journal of Public Economics*, 89, 761-96.

Jacob, B.A. (2007). *Test-based accountability and student achievement: An investigation of differential performance on NAEP and state assessments* (Working Paper No. 12817). Cambridge, MA: National Bureau of Economic Research.

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

Kannapel, J.P., Aagard, L., Coe, P., & Reeves, C. (2001). The impact of standards and accountability on teaching and learning in Kentucky. In S.H. Fuhrman (Ed.), *From the capitol to the classroom: Standards-based reform in the states* (pp. 217-41). Chicago, IL: The University of Chicago Press.

Klein, S.P., Freedman, D., Shavelson, R., & Bolus, R. (forthcoming). Assessing school effectiveness. *Evaluation Review.* Available at http://www.stat.berkeley.edu/~census/finalER.pdf (retrieved 10/13/08).

Koretz, D. (2003, April). Attempting to discern the effects of the NCLB accountability provisions on learning. Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Koretz, D.M. (2005). *Alignment, high stakes, and the inflation of test scores* (CSE Report No. 655). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing.

Koretz, D. M., & Hamilton, L. S. (2006). Testing for accountability in K-12. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 531-578). Westport, CT: American Council on Education/Praeger.

Lane, S., Parke, C.S., & Stone, C.A. (2002). The impact of a state performance-based assessment and accountability program on mathematics instruction and student learning: Evidence from survey data and school performance. *Educational Assessment,* 8, 279-315.

Lauer, P.A., Snow, D., Martin-Glenn, M., Van Buhler, R.J., Stoutenmyer, & Snow-Renner, R. (2005). *The influence of standards on K-12 teaching and student learning: A research synthesis.* Denver, CO: Mid-continent Research for Education and Learning.

Le, V., Stecher, B.M., Lockwood, J.R., Hamilton, L.S., Robyn, A., Williams, V., Ryan, G., Kerr, K., Martinez, F., & Klein, S. (2006). *Improving mathematics and science education: A longitudinal investigation of the relationship between reform-oriented instruction and student achievement.* Santa Monica, CA: RAND.

Linn, R.L. (2003). Performance standards: Utility for different uses of assessments. *Education Policy Analysis Archives,* 11, 31. Retrieved October 20, 2003 from http://epaa.asu.edu/epaa/v11n31/

Marsh, J.A., Pane, J.F., & Hamilton, L.S. (2006). *Making sense of data-driven decision making in education:Evidence from recent RAND research*. Santa Monica, CA: RAND.

Massell, D. (1994). Achieving consensus: Setting the agenda for state curriculum reform. In R.F. Elmore & S.H. Fuhrman (Eds.), *The governance of curriculum: 1994 yearbook of the Association for Supervision and Curriculum Development* (pp. 84-108). Alexandria, VA: Association for Supervision and Curriculum Development.

Massell, D. & Fuhrman, S.H. (1994). *Ten years of state education reform, 1983-1993*. New Brunswick, NJ: Consortium for Policy Research in Education.

Massell, D., Kirst, M., & Hoppe, M. (1997). *Persistence and change: Standards-based reform in nine states.* New Brunswick, NJ: Consortium for Policy Research in Education.

McDonnell, L.M. (1995). Opportunity to learn as a research concept and a policy instrument. *Educational Evaluation and Policy Analysis,* 17, 3, 305-322.

McDonnell, L.M. (2005). No Child Left Behind and the federal role in education: Evolution or revolution? *Peabody Journal of Education,* 80, 2, 19-38.

National Council on Education Standards and Testing (1992). *Raising standards for American education.* Washington, DC: U.S. Government Printing Office.

National Commission on Excellence in Education (1983). *A nation at risk: The imperative for educational reform.* Washington, DC: U.S. Government Printing Office.

National Research Council (2006). *Systems for state science assessment.* Committee on test Design for K-12 Science Achievement. M.R. Wilson and M.W. Bertenthal, eds., Board on Testing and Assessment, Center for Education, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

Nichols, S.L., Glass, G.V., & Berliner, D.C. (2006). High stakes testing and student achievement: Does accountability pressure increase student learning? Educational Policy Analysis Archives, 14(1). http://epaa.asu.edu/epaa/v14n1/

No Child Left Behind Act of 2001. 20 U.S.C. 6311 *et seq*.

O'Day, J. (1995). Systemic reform in California. In Goertz, M.E., Floden, R.E., O'Day, J. (Eds.), *Studies of education reform: Systemic reform, vol II: Case studies.* New Brunswick, NJ: Rutgers, the State University of New Jersey, Consortium for Policy Research in Education.

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

Organisation for Economic Co-operation and Development (2007). *PISA 2006: Science competencies for tomorrow's world,* Vol. 1. Paris: Author.

Partnership for 21st Century Skills (2008). *21st century skills, education, and competitiveness: A resource and policy guide.* Available at http://www.21stcenturyskills.org/documents/21st_century_skills_education_and_competitiveness_guide.pdf (last accessed 9/26/08).

Pedulla, J.J., Abrams, L.M., Madaus, G.F., Russell, M.K., Ramos, M.A., & Miao, J. (2003). *Perceived effects of state-mandated testing programs on teaching and learning: Findings from a national survey of teachers.* Boston, MA: National Board on Educational Testing and Public Policy.

Pellegrino, J.W., Jones, L.R., & Mitchell, K.J. (1999). *Grading the nation's report card.* Washington, DC: National Academy Press.

Popham, W.J. (1987). The merits of measurement-driven instruction. *Phi Delta Kappan,* 68*,* 679–82.

Popham, W.J. (2008). *The role of assessment in federal education programs*. Washington, DC: Center on Education Policy.

Porter, A. (1994). National standards and school improvement in the 1990s: Issues and promise. *American Journal of Education,* 102, 421-449.

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

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

Resnick, L.B., & Resnick, D.P. (1992). Assessing the thinking curriculum: New tools for educational reform. In B.R. Gifford & M.C. O'Connor (Eds.), *Changing assessment: Alternative views of aptitude, achievement, and instruction* (pp. 37–75). Boston, MA: Kluwer.

Rothman, R. (2005). Testing goes to preschool. *Harvard Education Letter, 21*(2) (March/April). Available at http://www.edletter.org/pdfs/2005-ma-fcd.pdf (retrieved 10/13/08).

Rothman, R., Slattery, J.B., Vranek, J. L., & Resnick, L.B. (2002). *Benchmarking and alignment of standards and testing* (CSE Technical Report 566). Los Angeles: Center for Research on Evaluation, Standards, and Student Testing.

Rothstein, R., Jacobsen, R., & Wilder, T. (2006). "Proficiency for all"—An oxymoron. Paper prepared for the Symposium, "Examining America's Commitment to Closing Achievement Gaps: NCLB and Its Alternatives," sponsored by the Campaign for Educational Equity, Teachers College, Columbia University.

Rosenshine, B. (2003) High-stakes testing: Another analysis. *Education Policy Analysis Archives*, 11, 24. Retrieved July 16, 2008 from http://epaa.asu.edu/epaa/v11n24/

Smith, M.S., & O'Day, J. (1991). Systemic school reform. In S. H. Fuhrman & B. Malen (Eds.), *The politics of curriculum and testing: The 1990 yearbook of the politics of education association* (pp.233-67). New York, NY: The Falmer Press.

Stecher, B.M. (2002). Consequences of large-scale high-stakes testing on school and classroom practice. In L.S. Hamilton, B.M. Stecher & S.P. Klein (Eds.), *Making sense of test-based accountability in education*. Santa Monica, CA: RAND.

Texas Education Agency. *An Overview of the History of Public Education in Texas*. Last Updated July 13, 2004. Retrieved December 15, 2008 from http://www.tea.state.tx.us/tea/historyove rview.html

Thomas B. Fordham Institute (2008). *High-achieving students in the era of NCLB*. Washington, DC: Author.

Valencia, S.W., & Wixson, K.K. (2001). Inside English/language arts standards: What's in a grade? *Reading Research Quarterly*, 36, 202-217.

Wilson, S.M., & Floden, R.E. (2001). Hedging bets: Standards-based reform in classrooms. In S.H. Fuhrman (Ed.). *From the capitol to the classroom: Standards-based reform in the states.* Chicago, IL: The University of Chicago Press, 193-216.

Wixson, K.K., Dutro, E., & Athan, R.G. (2003). The challenge of developing content standards. *Review of Research in Education,* 27*,* 69-107.