Editorial

THEMATIC SECTION

Christina Segerholm The Quality Turn. Political and Methodological Challenges in Contemporary Educational Evaluation and Assessment
Laura S. Hamilton, Brian M. Stecher & Kun Yuan Standards-Based Accountability in the United States: Lessons Learned and Future Directions
Peter Dahler-Larsen Constitutive effects as a social accomplishment: A qualitative study of the political in testing
Barbara Crossouard Classroom assessment and education: Challenging the assumptions of socialisation and instrumentality
D. Royce Sadler Assessment, evaluation and quality assurance: Implications for integrity in reporting academic achievement in higher education
Thomas A. Schwandt Quality, Standards and Accountability: An Uneasy Alliance

OPEN SECTION

Maria Hedlin Admission policy of Swedish teacher education favouring men: Discussion in Parliament in 1962
Stig-Börje Asplund Being a skilled reader: Reception patterns in vehicle engineering students’ literature discussion
Anna-Carin Jonsson & Dennis Beach Predicting the use of praise among pre-service teachers: The influence of implicit theories of intelligence, social comparison and stereotype acceptance
Kerstin Bergqvist “Own work” in primary school – A teaching and learning practice in the context of administration and control
Education Inquiry is an international on-line, peer-reviewed journal with free access in the field of Educational Sciences and Teacher Education. It publishes original empirical and theoretical studies from a wide variety of academic disciplines. As the name of the journal suggests, one of its aims is to challenge established conventions and taken-for-granted perceptions within these fields.

Education Inquiry is looking for lucid and significant contributions to the understanding of contextual, social, organizational and individual factors affecting teaching and learning, the links between these aspects, the nature and processes of education and training as well as research in and on Teacher Education and Teacher Education policy. This includes research ranging from pre-school education to higher education, and research on formal and informal settings. Education Inquiry welcomes cross-disciplinary contributions and innovative perspectives. Of particularly interest are studies that take as their starting point, education practice and subject teaching or didactics.

Education Inquiry welcomes research from a variety of methodological and theoretical approaches, and invites studies that make the nature and use of educational research the subject of inquiry. Comparative and country-specific studies are also welcome.

Education Inquiry readers include educators, researchers, teachers and policy makers in various cultural contexts.

Every issue of Education Inquiry publishes peer-reviewed articles in one, two or three different sections. Open section: Articles sent in by authors as part of regular journal submissions and published after a blind review process. Thematic section: Articles reflecting the theme of a conference or workshop and published after a blind review process. Invited section: Articles by researchers invited by Education Inquiry to shed light on a specific theme or for a specific purpose and published after a review process.

Education Inquiry is a continuation of the Journal of Research in Teacher Education, which is available in printed copies as well as electronic versions and free access at http://www.use.umu.se/forskning/publikationer/lof/
Standards-Based Accountability in the United States: Lessons Learned and Future Directions

Laura S. Hamilton*, Brian M. Stecher** & Kun Yuan***

Abstract:
Standards-based accountability (SBA) has been a primary driver of education policy in the United States for several decades. Although definitions of SBA vary, it typically includes standards that indicate what students are expected to know and be able to do, measures of student attainment of the standards, targets for performance on those measures, and a set of consequences for schools or educations based on performance. Research on SBA indicates that these policies have led to some of the consequences its advocates had hoped to achieve, such as an emphasis on equity and alignment of curriculum within and across grade levels, but that it has also produced some less desirable outcomes. This article summarizes the research on SBA in three areas: quality of standards, ways in which SBA has shaped educators’ practices, and effects on student achievement. The article identifies lessons learned from the implication of SBA in the United States and provides guidance for developing SBA systems that could promote beneficial outcomes for students.

Keywords: Accountability, Assessment, Standards, Education policy, United States

Introduction
Education policy debates in the United States have been shaped by a growing recognition that students in the U.S. have failed to keep pace with their counterparts in other countries in tests of educational achievement (OECD, 2009) and by a related concern that too many students are either not graduating from high school or are graduating without the skills and knowledge needed for success in postsecondary education or careers (National Commission on Excellence in Education, 1983; Swanson, 2009). One of the most prominent policy responses to these concerns has been an emphasis on standards-based accountability (SBA), an approach to measuring and incentivising school performance by attaching consequences to student scores in achievement tests.

Across the U.S., states have adopted standards that describe the content schools are expected to teach and that students are expected to master in specific grade levels to prepare them for college and careers, along with tests that measure attainment of those standards. The requirement for standards and aligned assessments has been a feature of federal legislation since the Improving America’s Schools Act (IASA) of

*RAND Corporation. E-mail: laurah@rand.org
©Authors. ISSN 2000-4508, pp.149–170
1994, and it is the centrepiece of the No Child Left Behind Act of 2001 (20 U.S.C. 6311 et seq.). Signed into law in January 2002, NCLB requires each state to develop content standards and aligned student achievement tests in specific subjects and grades, along with a set of escalating sanctions that are tied to a repeated failure of schools to meet their student performance targets. These targets are tied not only to aggregated scores for all tested students in a school, but also to the performance of subgroups of students such as English language learners and low-income students (see Hamilton et al., 2007, for a more detailed description of the elements of NCLB). The law has exerted a strong influence on state and local decisions about education policy and practice since its passage. Although the specific features of NCLB are unique to the U.S., its general approach to accountability, which promotes an approach to school governance that relies heavily on outcomes, is also a topic of debate in a number of other countries. This paper is intended to share the findings of research conducted in the U.S. context, with the hope of informing accountability system design and implementation in other nations.

**Historical Context**

Both federal and state governments in the U.S. played an important role in shaping SBA during the decades prior to NCLB’s enactment. Although interest in measuring educational outcomes had been growing throughout the 1960s and 1970s, and several states began adopting elements of SBA 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 colourful language to deplore the state of American education, led to policy debates about how to raise expectations for both student and teacher performance, and emphasised 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 which content and skills were expected to be taught (Mas-sell, 1994). This concern for coherence and for clear communication of expectations contributed to the growing interest in reforming education through system-wide standards. 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 emphasised the role of states and national organisations 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 organisations to engage in efforts to promote SBA.

States played a particularly important role in the evolution of SBA prior to NCLB. In the 1970s, states began developing minimum-competency examinations in response to growing concerns about low-achieving students. Then in the 1980s and early 1990s, states such as California, Kentucky, Maryland, Massachusetts, North Carolina and Texas began to implement SBA using their own funds. Later, the 1994 Goals 2000: Educate America Act (PL 103-227) funded state efforts to develop standards, and almost all states embarked on developing standards if they had not begun this task already (Armour-Garb, 2007).

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 federal standards and testing requirements at the time. Much of the recent SBA activity can be attributed directly to NCLB which has required each state to establish a system of SBA that includes standards, assessments and annual targets for performance, but departed from earlier federal initiatives by imposing stricter requirements for testing (e.g., a requirement that all students in grades 3 to 8 be tested annually) and for the creation of proficiency-based cut scores.

It is important to recognise that NCLB is not the only factor affecting state and local SBA efforts. In particular, there have been extensive efforts to promote increased accountability and uniformity of expectations at the high school level, outside the NCLB requirements. 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 most states the primary purpose of these exams 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.

Another example of SBA-induced policy change is the growing emphasis on teacher-level accountability and the use of value-added modelling of student test scores to estimate teacher effectiveness. This approach to accountability reflects a growing body of evidence suggesting that individual teachers exert strong effects on student learning and that this effectiveness varies substantially, even within schools (Nye, Konstantopoulos, & Hedges, 2004; Rivkin, Hanushek, & Kain, 2005). Despite technical challenges associated with attributing learning gains to individual teachers (Baker et al., 2010; McCaffrey et al., 2003), the use of value-added modelling has become widespread at the state and local levels in the U.S. and is likely to play a significant role in future accountability policy.
The next wave of accountability reforms in the U.S., while expected to depart from NCLB in various ways, is likely to maintain a focus on standards and assessments, particularly in light of other federal policies that have emphasised these strategies. One of these programmes is Race to the Top, through which the U.S. Department of Education has awarded competitive grants to states that agree to implement reforms including new educator evaluation and accountability policies and data systems to track student achievement over time. More recently, the U.S. Department of Education has begun to grant states waivers from some of the most harsh NCLB requirements but, to qualify, states must agree to reforms similar to those embodied in Race to the Top. Another prominent initiative is the Common Core State Standards (CCSS), a non-federal effort to develop challenging, college- and career-ready standards in reading and mathematics, with voluntary adoption by states. A majority of states have agreed to adopt the CCSS and many of these states have also agreed to participate in one of two consortia that are developing assessments that will be aligned with the CCSS. Given the continuing importance of SBA, it is useful to examine what the research suggests about its promise as a strategy for improving student learning. The rest of this paper describes SBA, discusses what we know about how the SBA movement has shaped the current debate about public education, educators’ practices, and student outcomes in the U.S., and offers recommendations for improving these policies in the future.

What Is Standards-Based Accountability?

The term “standards-based accountability” (SBA) is used extensively in discussions of U.S. education policy, but it has not meant the same thing to everyone who has used the term. Further complicating the situation, educators and policymakers have used other terms, including “systemic reform”, “standards-based reform” and “curriculum alignment” to describe similar ideas that differ somewhat in emphasis or evolution. As Wilson and Floden (2001) observed, “The slogans of standards and standards-based reform spread widely in the 1990s, but the meaning varied across contexts” (p. 195).

Although there is no universally accepted definition of SBA, most discussions of SBA include some or all of the following features:

Academic expectations for students. SBA represents an approach to accountability that emphasises academic outcomes, which in the U.S. represents a shift from an earlier approach that emphasised school inputs and process. Thus, the first step in SBA 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 education system, including the testing system (discussed below). Academic standards are typically 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
Standards-Based Accountability in the United States: Lessons Learned and Future Directions

scores (NCLB requires at least three) for a test that indicate 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.

Although academic standards are central to the idea of SBA, 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 SBA 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 also emphasise 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 SBA. 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? In addition, there is an inherent tension in the application of standards that are simultaneously high and uniform; this model would require a substantial reduction in variability in student performance, and there is no evidence that such a reduction has ever been achieved in any education system. The tension is particularly evident in the NCLB policies on the testing of students with disabilities.

The U.S. Department of Education issued a number of guidelines regarding which students needed to be included in accountability testing and when it is acceptable to modify the testing situation (e.g., U.S. Department of Education, 2007). The complexity of these rules reflects the government’s desire to maintain an emphasis on uniform standards while acknowledging the reality that some students will almost inevitably be unsuccessful in the tests used for the majority of the student population.

Alignment of the key elements of the education system. The second core element of SBA is the alignment of the rest of the instructional delivery system to promote attainment of the standards. As Clune (2001) characterised it, the central thesis of SBA 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. 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 SBA 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 programmes, 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 the 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. 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). Excessive emphasis on alignment can also lead to some of the negative effects of SBA that we discuss later, particularly when educators use the goal of alignment as a rationale for narrowing their instruction to focus on tested content (Koretz, 2005).

Assessments of student achievement. The third major component of SBA is the inclusion of assessments of student achievement. The role of testing in SBA was 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 (Resnick & Resnick, 1992). The SBA movement has fuelled further growth in large-scale assessment, accompanied by efforts to develop data systems to track student progress. Although there is more research evidence to support specific decisions about assessment than there is to inform the development of standards, there are still unresolved issues about assessment that pose challenges for SBA.

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, standardised 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, such as 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, SBA demands a high volume of assessment geared toward a concise, quantitative summary of performance, along with timely judgments about every school or every student. These requirements put a premium...
on efficient assessment. Policymakers have to balance the desire for assessments that provide richer information against the need for assessments that provide succinct, timely information.

Support and technical assistance. As noted earlier, Smith and O’Day (1991) are credited by many with first characterising the elements now associated with SBA, particularly the focus on aligning all components of the education system. (They described their vision as one of “systemic reform”.) Their model also emphasises 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 may not be sufficient to induce change; externally provided assistance may be necessary. Not all SBA efforts encompass support strategies, but state and district support clearly plays a prominent role in NCLB’s system of consequences for schools that fail to meet their performance targets.

Accountability. Advocates of SBA have 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. Consequences for poor performance were included in earlier federal legislation and in state systems that predated NCLB, but NCLB ramped up the specificity and intensity of the consequences 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 behaviour and on test scores.

As this brief overview suggests, SBA 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. Despite this ambiguity, the shift toward SBA in the U.S. has fundamentally shaped education policy debates at all levels of the system. In the next section we discuss some of the broad themes that characterise current education policy in the U.S.

How SBA Is Shaping Today’s Debates about Public Education in the U.S.

In response to NCLB, every state has adopted an SBA system but, as noted above, these systems vary along key dimensions such as the specificity of the content standards, the content and format of state assessments, and the level of difficulty at which the
“proficient” cut score is set. Initiatives such as the Common Core State Standards are intended to address some of this variation but we do not yet know what the outcomes of those efforts will be. Regardless of which provisions are incorporated into future accountability legislation and programmes, several core ideas have emerged and are likely to influence the next generation of education reforms in the U.S. These ideas, which have affected not only the development of specific SBA programmes 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 analysed, assimilated and acted upon, this information has potential value for improving educational practice by clarifying precisely which outcomes should be the focus of instruction and providing information about how much progress each student is making in achieving them. The recent growth in the availability of tools and services to promote “data-driven decision making” in education (Coburn & Turner, 2011) is an outgrowth of the SBA 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 lack of consensus that characterises some of the debates surrounding how to teach maths 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 at college or the workplace (see, e.g., Partnership for 21st Century Skills, 2008). These results have been used by policymakers in the U.S. as justification for raising standards and curricular requirements. The current system, with its reliance on standardised tests that often emphasise 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. The two consortia developing new assessments linked to the Common Core State Standards have stated that they are working to ensure that these assessments measure the challenging skills contained in the standards.

**Importance of promoting equity.** Inequities in opportunity continue to plague the education system, but the SBA 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 focusing on funding. Both the requirement to link accountability requirements to the performance of 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 the 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 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 (see below) 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 what is 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 (e.g., narrowing of curriculum) 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.

The broad themes listed above suggest that how we think about public education has changed as a result of SBA, 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 summarises the research on standards and how they influence educational practices and outcomes.

**What Research Tells Us about SBA in the U.S.**² Despite the near-ubiquity of standards and other elements of SBA today, the question of whether SBA has benefited students in the U.S. is still hotly debated. Concerns about possible adverse consequences of SBA 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”. 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 SBA, the caution Porter expressed in 1994 about the lack of empirical evidence remains relevant today.

High-quality research on the effects of SBA 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 SBA, and addressing the diversity in the assessment programmes and accountability policies in different states and districts. Even when the necessary data have been collected by states or other entities, it is often difficult for researchers to obtain these data because those responsible for the data refuse to grant access, either because of concerns about confidentiality or because they are not interested in having their programmes scrutinised by researchers. 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 SBA and school and classroom practices, and studies of links between SBA and student achievement.

Quality of Content Standards. 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 (Archbald, 1998), with some sets of standards receiving positive marks from one organisation and while being criticised by another (Valencia & Wixson, 2001). These discrepancies may reflect the fact that the language describing the features that characterise 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 assigned by different organisations. Differences in the importance attached to the various dimensions of quality in the evaluation process reflect, to some degree, a diversity of understanding about the purposes and roles of standards and diversity of values regarding teaching and learning. Regardless of which criteria are used to evaluate quality, however, the existing reviews of state standards suggest that there is room for improvement: Relatively few states receive uniformly high marks under any set of criteria (Finn, Julian, & Petrilli, 2006; Glidden, 2008).

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.

School and Classroom Practices. A second collection of research addresses the important question of how SBA 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 (Lauer et al., 2005). In fact, the bulk of research relevant to SBA has focused on the links between high-stakes tests and educators’ practices rather than standards and practices. The preponderance of research on the impact of testing rather than the impact of standards reflects the emerging realisation that “standards-based accountability” has largely given way to “test-based accountability”, 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 (Center on Education Policy, 2006; Goertz, 2007; Hamilton, 2003; Lane, Parke, & Stone, 2002; Stecher, 2002, Stecher et al., 2008). Some of the changes would generally be considered beneficial. These changes include the provision of additional instruction to low-performing students; improved alignment of the school curriculum across grades, and efforts to improve teachers’ pedagogical skills. However, teachers and administrators have reported other reactions that arise raise concerns about possible negative effects on the breadth and quality of instruction, including the reallocation of 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-emphasising material that is not tested (Koretz & Hamilton, 2006). Although proponents of SBA may view the reallocation from non-tested to tested material as a positive outcome, particularly in the context of SBA 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).

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
and 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). Accountability policies that focus on the percentage of students who perform above or below a specific cut score, as is the case with NCLB’s emphasis on proficiency, have been shown to promote an additional type of response: Teachers and other school staff have 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).

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. At the same time, research on performance-based assessment makes it clear that such tests are not always resistant to score inflation, and in fact their use can sometimes promote score inflation, particularly if the test is characterised by a high degree of predictability and a narrow range of content (Hambleton et al., 1995; Koretz & Barron, 1998). The greatest concern in the context of accountability is that these changes in emphasis lead to inflated test scores, making students appear more proficient and schools appear more successful than they really are (Koretz, 2008). There is ample evidence that score increases in high-stakes tests do not always generalise to other measures of the same content, for example, because they primarily reflect narrow test-preparation activities geared toward a specific test.

Together, these studies suggest that SBA exerted a significant effect on the actions of public school teachers and administrators. One of the likely reasons for the widespread nature of these effects is that virtually all parties involved in the SBA system have incentives to raise scores. Teachers are motivated to engage in practices that increase scores, and principals and other school leaders are likely to encourage these practices because they also benefit from improved performance. And, as noted earlier, even the commercial publishers of curriculum materials are likely to benefit when increases in scores accompany the adoption of their products. Most SBA systems lack any mechanism to encourage participants in the system to avoid engaging in practices that might raise scores without necessarily improving the quality of the education students receive.

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 SBA policies (Hamilton et al., 2008). 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 programmes have failed to induce deeper pedagogical change or fundamentally alter the ways teachers deliver instruction (Diamond, 2007; Firestone, Mayrowetz & Fairman, 1998). SBA has led to some beneficial changes in school practices at both the organisational and classroom levels, but it has not always produced the kinds of instructional improvement that advocates hoped for. These findings raise concerns about the generalisability of achievement gains (as we discuss in the next section), the equity of educational opportunities, and the ethical conduct of educators.

**Effects of SBA on Student Achievement.** Improved student achievement is widely viewed as the primary goal of SBA, but for a variety of reasons it is difficult to measure the impact of SBA on 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 render it difficult to establish a direct causal relationship between SBA 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 SBA. Finally, the question that may be of greatest interest to policymakers and the public is whether SBA or the test-based reform that has predominated in recent years has improved student learning. Recent gains at 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 (Center on Education Policy, 2007, 2008; Jacob, 2007). 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 score inflation, discussed above.

Past research that examines score gains in low-stakes (non-accountability) tests, such as the National Assessment of Educational Progress (an assessment that is given on a period basis to a sample of the nation’s students, with no stakes attached), suggests there has been some increase in achievement associated with state accountability policies (Carnoy & Loeb, 2002; Hanushek & Raymond, 2005; Jacob, 2005), though the gains at low-stakes tests are not as large as the gains on the high-stakes state tests (Hamilton, 2003; Jacob, 2007), so it is difficult to estimate the size of the actual improvement.

In a recent report, the National Research Council reviewed 15 rigorous studies (using experimental or strong quasi-experimental designs) of the impact of test-based incentives on student achievement (National Research Council, 2011). They concluded that large-scale accountability systems that targeted incentives at the school level (such as NCLB) had a positive impact on achievement in fourth-grade mathematics, with an average effect size of 0.08, although effects were not observed in other subjects and grade levels. Moreover, the report notes that “the effects are small compared to the improvements the nation hopes to achieve” (NRC, 2011, p. 4). The report eliminated from consideration a large number of studies that were unable to disentangle the
effects of SBA and high-stakes testing from other initiatives taking place during the same period or had other design limitations. 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 SBA’s effects remains elusive, the available research does help us identify several lessons that point to challenges faced by those who develop SBA 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. The U.S. Department of Education is currently funding multi-state consortia to develop a new generation of tests aligned with the CCSS that will use a wider range of formats to assessment challenging skills (Educational Testing Service, 2011).

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 SBA tend to focus more on tested material and tested formats 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. These kinds of distortions are evident in a variety of high-stakes accountability systems outside the field of education (Stecher et al., 2010) and are consistent with a phenomenon known as Campbell’s Law: “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” (Campbell, 1979, p. 87). The problem stems from the incomplete sampling of measures included in an accountability system. In education, the academic subjects in which tests are typically given (primarily mathematics and reading) comprise only a portion of the broad set of goals schools are typically expected to achieve and, within those subjects, the content that can be included on a standardized test represents only a small portion of all of the skills and knowledge students are expected to attain in those subjects. The relatively small fraction of outcomes that can be measured by accountability tests creates incentives for educators to focus on those measured outcomes, often at the expense of other important but unmeasured outcomes.

**SBA allocates responsibility in ways that can conflict with traditional educational governance.** Some early proponents of SBA 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.

**Where We Might Go from Here**

There are legitimate questions about the appropriate role for governments in SBA, especially given the mixed findings in the U.S. to date relating to the implementation and impact of NCLB and the long tradition of a limited federal role in education. The existing evidence does not provide definitive guidance regarding the SBA system features that would be most likely to promote desirable outcomes. However, research findings on the quality and consequences of SBA policies can be used to generate guidance to inform development of the next generation of SBA systems. Here we provide several suggestions that governments might want to consider as they work to develop or overhaul their standards, assessments, and accountability provisions in the interest of improving student learning.

- Develop accountability indices that create more effective incentives by addressing the shortcomings that research has identified in current accountability metrics. For example, indices that 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 is more consistent with public goals for education. Indices based on growth may also provide a better basis for accountability, but there are many technical and practical concerns that would have to be addressed (see Baker et al., 2010, for a discussion of some of these concerns).

- Adopt an audit mechanism to facilitate the measurement of effects of SBA on student learning. The widespread presence of score inflation limits the utility of the high-stakes accountability test as a measure of student achievement. Tests such as the NAEP (National Assessment of Educational Progress) in the U.S. not only provide a means of comparing student performance across states, but have also allowed us to monitor achievement in subjects not typically included in state accountability systems. In addition, such tests 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 SBA efforts. If a government takes steps to require or promote SBA, it should also set aside resources for evaluating the effects of these policies. Such evaluations should measure more fully the impact of SBA 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 SBA, including graduation rates, course-taking patterns, and student learning in subjects not included in the high-stakes testing system.

- Expand alignment efforts to reflect a more systemic view of the education 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 SBA 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 the 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, resources are needed to help teachers use data for instructional decision-making; such resources would eschew the test-focused use of data and foster student-focused individualisation to address students’ unique needs and interests.
• Take advantage of advances in psychometrics and technology to develop new assessment methods that could tap a broader range of skills and knowledge than existing 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 SBA. 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. The CCSS Assessment Consortia in the U.S. have appealing plans that incorporate many of these ideas, particularly by adopting a variety of item formats and using technology to support administration and scoring of open-ended items. If they are successful, they can provide models for others, though it will be important to evaluate the tests and the scores in them to determine the extent to which they are able to reduce score inflation.

• Finally, consider adopting a comprehensive indicator system to replace the test-focused 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, reduce the pressure to focus exclusively on tested material, and could allow schools to experiment with innovative curricula and programmes 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. It could also minimise the predictability that is inherent in today’s narrower test-based systems and therefore reduce the likelihood that educators will constrict 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 (see Schwartz et al., 2011 for a more extensive discussion of expanded measures in accountability).
SBA 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 realised, but some of the fears of early critics have materialised. Ongoing efforts to improve our knowledge of the effects of SBA and disseminate this knowledge to decision-makers at all levels will be critical for improving SBA systems that promote high-quality teaching and learning.

Laura Hamilton is a Senior Behavioral Scientist and Research Quality Assurance Manager at the RAND Corporation and an Adjunct Associate Professor in the University of Pittsburgh’s Learning Sciences and Policy program. Her research focuses on educational assessment, accountability, and the measurement and evaluation of instruction and school leadership.

Brian Stecher is a Senior Social Scientist and the Associate Director of RAND Education. Dr. Stecher’s research focuses on measuring educational quality and evaluating education reforms, with a particular emphasis on assessment and accountability systems.

Kun Yuan is an Associate Behavioral Scientist at the RAND Corporation. Her research focuses on educational measurement, program evaluation, multilevel mixed-effects modeling, and the use of technology in education.
Standards-Based Accountability in the United States: Lessons Learned and Future Directions

References


Koretz, D., and Barron, S. I. (1998) *The Validity of Gains on the Kentucky Instructional Results Information System (KIRIS)*. MR-1014-EDU, Santa Monica: RAND.


and Assessment, Center for Education, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.


Endnotes

1 Material in this paper has been adapted from a paper commissioned by the Center on Education Policy: Hamilton, L.S., Stecher, B.M., & Yuan, K. (2009) Standards-based Reform in the United States: History, Research, and Future Directions. Washington, DC: Center on Education Policy. Portions of this work were 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.

2 A complete review of research on SBA is beyond the scope of this article. Readers interested in additional detail should refer to the review articles cited throughout this section.
Editorial

THEMATIC SECTION
Christina Segerholm The Quality Turn: Political and Methodological Challenges in Contemporary Educational Evaluation and Assessment
Laura S. Hamilton, Brian M. Stecher & Kun Yuan Standards-Based Accountability in the United States: Lessons Learned and Future Directions
Peter Dahler-Larsen Constitutive effects as a social accomplishment: A qualitative study of the political in testing
Barbara Crossouard Classroom assessment and education: Challenging the assumptions of socialisation and instrumentality
D. Royce Sadler Assessment, evaluation and quality assurance: Implications for integrity in reporting academic achievement in higher education
Thomas A. Schwandt Quality, Standards and Accountability: An Uneasy Alliance

OPEN SECTION
Maria Hedlin Admission policy of Swedish teacher education favouring men: Discussion in Parliament in 1962
Stig-Börje Asplund Being a skilled reader: Reception patterns in vehicle engineering students’ literature discussion
Anna-Carin Jonsson & Dennis Beach Predicting the use of praise among pre-service teachers: The influence of implicit theories of intelligence, social comparison and stereotype acceptance
Kerstin Bergqvist “Own work” in primary school – A teaching and learning practice in the context of administration and control