NATIONAL ACADEMY of EDUCATION

Standards, Assessments, and Accountability

EDUCATION POLICY WHITE PAPER

Standards-based education reform has a more than 20-year history. A standards-based vision was enacted in federal law under the Clinton administration with the 1994 reauthorization of the Elementary and Secondary Education Act (ESEA) and carried forward under the Bush administration with the No Child Left Behind Act (NCLB) of 2001.[1] In a recent survey of policy makers, standards were acknowledged as the central framework guiding state education policy.[2]

Yet, despite this apparent unanimity about the intuitively appealing idea of standards, there is great confusion about its *operational* meaning: exactly what should the standards be, how should they be set and by whom, and how should they be applied to ensure rigorous and high-quality education for American students are the central questions that challenge policy makers and educators. For example, *content* standards (subject-matter descriptions of what students should know and be able to do) are often confused with *performance* standards (which are more like passing scores on a test), and very different theories of action are used to explain how standards-based reforms are expected to work. Ambitious rhetoric has called for systemic reform and profound changes in curriculum and assessments to enable higher levels of learning. In reality, however, implementation of standards has frequently resulted in a much more familiar policy of test-based accountability, whereby test items often become crude proxies for the standards.

This disconnect between rhetoric and reality is one of the reasons for the failure of prior reforms. For example, early advocates of standards-based reforms were reacting against previous efforts focused on *minimum competencies* (such as balancing a checkbook) that had done little to improve the quality of instruction or student learning. To fight against low expectations and an incoherent, de facto curriculum driven by textbooks and basic skills tests, they called for clear and challenging content standards and a coherent structure of state leadership that would provide long-term support to enable more fundamental changes in instruction.[3] In *Promises to Keep:*

Standards, Assessments, and Accountability Education Policy White Paper

Editors:

Lorrie Shepard, School of Education, University of Colorado at Boulder

Jane Hannaway, Education Policy Center, The Urban Institute

Eva Baker, Graduate School of Education and Information Studies, University of California, Los Angeles

National Academy of Education Working Group on Standards, Assessments, and Accountability:

Eva Baker (Co-Chair), Graduate School of Education and Information Studies, University of California, Los Angeles

Jane Hannaway (Co-Chair), Education Policy Center, The Urban Institute

Patricia Gandara, Graduate School of Education and Information Studies, University of California, Los Angeles

Drew Gitomer, Education Testing Service

Margaret Goertz, Graduate School of Education, University of Pennsylvania

Helen Ladd, School of Public Policy, Duke University

Robert Linn, School of Education, University of Colorado at Boulder

P. David Pearson, Graduate School of Education, University of California, Berkeley

Diane Ravitch, School of Education, New York University

William Schmidt, College of Education, Michigan State University

Alan Schoenfeld, Graduate School of Education, University of California, Berkeley

David Stern, Graduate School of Education, University of California, Berkeley

William Trent, College of Education, University of Illinois at Urbana-Champaign

Mark Wilson, Graduate School of Education, University of California, Berkeley

Creating High Standards for American Students, a panel of policy makers and academics explained that raising performance in the way envisioned would require a systematic and sustained effort from all levels of the education system. Clear and visible standards—identifying what students should know and be able to do—would need to be reinforced by "curricula, teacher training, instructional materials, and assessment practices to enable students to meet them."[4]

A National Academy of Education Panel on Standards-Based Education Reform concurred that findings from cognitive science research make it at least theoretically possible to focus instruction on depth of understanding, and to provide the support for reaching a much more diverse population of students. But, the report cautioned that extrapolating from small-scale, intensive studies to full-system reform was an unprecedented task—one that would require significant investments in teacher professional development and ongoing evaluations to improve the system.

Basic elements of the standards vision were established in the 1994 ESEA. The law required that states set challenging and rigorous content standards for all students and develop assessments, aligned with the standards, to measure student progress. By holding schools accountable for meeting the standards, it was expected that teachers and actors at other levels of the educational system would redirect their efforts and find ways to improve student achievement. In contrast to the hoped-for idea of coherent capacity building envisioned for Goals 2000: Educate America Act of 1994, passed earlier that same year, ESEA set forth primarily an *incentives* theory of change. It assumed that—with sufficient motivation—teachers (and other relevant school personnel) would find the means to improve instruction. Unfortunately, early implementation research showed that many schools lacked an understanding of the changes that were needed, and also lacked the capacity to make them happen.[5] NCLB intensified the commitment to leverage change through test-based accountability, and in at least two significant ways, "upped the ante": (1) its focus on disaggregating data (to make it possible to track the performance of various subgroups such as major racial/ethnic groups or those with limited English proficiency) reflected the widespread belief that the achievement gap was an unacceptable reality in American education and that hard data demonstrating inequality of outcomes would be necessary, if not sufficient, to remedy the situation; and (2) its urgency—as reflected in the requirement that all students reach a certain standard of performance by 2014—spoke to the growing frustration across the nation about the slow pace of progress. The current policy context, therefore, is best understood as a blend of standards rhetoric and test-based accountability practices.

Research findings about the effects of standards and test-based accountability have been both promising and disappointing. Educators have redirected efforts as intended, adopting curricula aligned with state standards[6] and dramatically increasing the amount of instructional time devoted to reading and mathematics.[7] Accountability pressures have resulted in increased use of test data to redirect instructional efforts, extensive test preparation practices, and increasing use of interim and benchmark tests administered periodically to monitor progress toward mastery of standards.[8]

Although it is difficult to tie achievement results to specific policies, some researchers have found positive links between states with stronger accountability policies and relatively more improvement on the National Assessment of Educational Progress (NAEP). For example, from 1996 to 2000 one study found that "high-accountability states" had relatively greater gains on NAEP in mathematics for eighth grade and for African-American and Hispanic fourth graders scoring at the Basic Level.[9] Positive findings for test-based accountability have been partially confirmed by other researchers.[10] Using NAEP data from 1992–2002 one study found higher achievement associated with accountability but no narrowing of the Black–White achievement gap. Positive effects are also challenged by evidence that some states may have excluded large numbers of students from NAEP testing or by evidence that gains were weaker or nonexistent when student cohorts were tracked longitudinally.[11]

A majority of states have reported gains on state tests and a general closing of gaps,[12] but these increases need to be viewed cautiously. For example, because of the pervasive problem of test-score inflation (i.e., score gains that overstate the underlying gains in genuine learning) that can occur when teachers and students become increasingly familiar with the content and format of state tests, researchers prefer to rely on NAEP as a source of more credible information about achievement progress. General increases on NAEP and gap closings have continued since 2002, especially in mathematics, but the improvements are much more modest than on state tests, and there is no difference in the rate of gain before or after NCLB.[13] A fair conclusion from all of these studies might be to say that, since 1992, the era of test-based accountability has been associated with increasing student achievement, but improvements have not been as clear-cut or dramatic as had been hoped and

cannot be attributed solely to accountability policies. Although the trend continues to be positive, the intensification of pressures since NCLB has not produced commensurately greater gains.

Studies showing positive changes in instructional practices because of accountability have also documented significant negative effects. For example, it is again the case that tests have had a stronger impact on teaching than standards.[14] Tested subjects receive much more instructional time than non-tested subjects, driving out art, music, and physical education, but also reducing time for science and social studies, especially for disadvantaged and minority students assigned to increased doses of reading and mathematics.[15] Citizens and policy makers are generally aware of the problem that teachers face strong incentives to emphasize content that is tested, which in some cases can become so strong that they actually "teach the test." Many educators, parents, and policy makers believe it reflects a necessary trade off—arguing that reading and math are the most essential skills and must be mastered even at the expense of other learning. However, research on teaching the test shows that pressure to raise test scores changes not only *what* is taught but *how* it is taught.

If teaching the test means practicing excessively on worksheets that closely resemble the test format, then it is possible for test scores to go up without there being a real increase in student learning. This problem of test score inflation can explain why scores are rising more dramatically on high-stakes state tests than on NAEP, although it is difficult to estimate the exact amount of inflation.[16] More significantly for the students themselves, the emphasis on rote drill and practice often denies students an opportunity to understand context and purpose that would otherwise enhance skill development.[17] It is much more interesting to work on writing skills, for example, after reading a book about Martin Luther King than it is to practice writing to test prompts. Some teachers also report focusing their efforts on *bubble kids*, those who are closest to the proficiency cut score, so that a small improvement can make a big difference in the school's percent proficient number.[18] These kinds of problems—gaming, distortion, and perverse incentives—are well known in the economics literature on incentives[19] and can be expected to occur when performance indicators are imperfect measures of desired outcomes. How these problems can and should be weighed by educators and policy makers is a thorny question for which the available findings on unintended consequences of test-based accountability provide useful but insufficient information.

According to some policy advocates, standards-based reforms have had limited success because underlying incentive structures have not been well enough understood and implemented. For some content experts, especially in mathematics and science, the reforms have failed because standards and assessments still do not reflect what is known from cognitive science about how conceptual learning develops in these fields.[20] For other reformers, the promise that standards would ensure equity has been broken because sanctions have been imposed without sufficient support to make it possible for standards to be met, especially in the poorest schools. Each of these perspectives has validity and, taken together, they can help us understand how it is that standards-based reforms have not yet been implemented in a way that adequately reflects original intentions. These imperfections and costs notwithstanding, policy makers feel an urgent need to use standards as a tool for improved education. Insights about how reforms have fallen short can lead to improvements in the design and implementation of standards and serve to leverage much needed reforms.

Content Standards

The Goals 2000: Educate America Act of 1994 defines *content standards* as "broad descriptions of the knowledge and skills students should acquire in a particular subject area."[21] For many states, the content standards adopted a decade ago represented that state's first effort at trying to develop some kind of curriculum framework. For most, the process was highly political—as well it should be in a democratic society. But without previous experience and access to coherent curricula representing particular curricular perspectives, the political solution of adding in everyone's favorite content area topic created overly-full, encyclopedic standards in some states, or vague, general statements in others. It should thus be no surprise that today's state content standards vary widely in coverage, rigor, specificity, and clarity,[22] despite the admonition in NCLB that all states should adopt "challenging academic content standards" with "coherent and rigorous content."[23]

There is now considerable political support in the United States for new common standards. The National Governors Association and the Council of Chief State School Officers are leading an effort to create shared high school graduation standards and grade-by-grade content standards in math and language arts, and—almost immediately—46 states signed on to the development effort.[24] It is hoped that common standards will be both more rigorous and better focused, but this politically important effort will not necessarily produce a better result unless past problems are better understood. In a recent work-

shop held by the National Research Council that addressed common standards, presenters argued that the development process for standards has often left out more complex, discipline-based expertise about how knowledge, skills, and conceptual understanding can be developed together in a mutually reinforcing way. A recent study of language arts, science, and mathematics content standards in 14 states, for example, found only low to moderate alignment between state standards and corresponding standards defined by national professional organizations (e.g., the National Council of Teachers of Mathematics).[25] There is a strong research base documenting how students develop advanced proficiencies in science and mathematics, and correspondingly what pedagogical practices tend to support such learning.[26] But, in the standards negotiation process, these more complex understandings are likely to be replaced by inclusive but disorganized lists of topics. Research on education reforms has clearly documented the need for *curricular coherence* to make sure that the pieces of reform work together, provide support for teacher learning, and convey consistent messages to students. Curricular coherence can refer both to the ways that policy instruments fit together—standards, assessment, and professional development—and to features of the curriculum itself.[27] Although most states and districts have attended to issues of *alignment* among standards, assessments, and textbooks, these are skeletal match-ups—outlines of similar topics—that do not address deeper issues of conceptual congruence between challenging curricular goals and the underlying structure of prerequisite topics and skills needed to achieve them. It is important to recognize that broad content standards, at least as developed thus far in the United States, do not have the specificity of curricula as typically developed in other countries, where there is greater clarity about the depth of coverage and the appropriate sequencing of topics. Proponents of common standards will need to clarify whether content standards will continue to be curriculum frameworks, intended as rough outlines of what should be taught. Or, will they take on the tougher task of specifying common grade-by-grade curricular goals?

Studies of the top-performing countries in the Third International Mathematics and Science Study (TIMSS) provide examples of coherent curricula. In contrast to U.S. state standards that appear to emphasize "rote memorization of particulars," mathematics and science curricula in the Czech Republic, Japan, Korea, and Singapore reflect a hierarchical sequencing of topics designed to move progressively toward more advanced topics and a deeper understanding of the structure of the discipline.[28] Differences among the top-performing countries indicate that there is more than one way for topics to be organized, but importantly, in each case choices have been made so that each country's curriculum is coherently organized with fewer topics per grade than the overwhelming and repetitive lists typically found in the United States. Although NCLB alignment requirements were intended to correct the "mile wide and inch deep" curriculum problems identified by TIMSS researchers, the most recent research indicates that these problems are largely unabated and also occur with English, language arts, and reading standards. A recent analysis of content standards from 14 states found that they failed to focus on a few big ideas and that they were not differentiated by grade so as to indicate how topics should build from one grade to the next. Of greatest relevance for a common standards effort, when content standards were compared between pairs of states, there was an average of only 20 percent overlap in the topics and level of cognitive complexity intended to be taught at each grade.[29]

Many analysts have pointed to the national control of curriculum in top performing countries as a means to ensure curricular coherence and correspondingly higher achievement. A finer-grained analysis, however, shows that national control is not required for coherence.[30] Rather, *coherence leads to effective outcomes if it is achieved at whatever level of governance has authority over policy instruments*. In a study of the 37 nations in TIMSS, only 19 reported having a single, centralized educational system. Unfortunately, neither states nor districts in the United States have a tradition of curriculum development and associated teacher professional development like that of many other countries, whether at the national or provincial level.

Performance Standards

Complaints that standards differ widely from state to state often confuse content standards with *performance standards*. Whereas content standards refer to the knowledge and skills that students should acquire in a particular subject area, performance standards are "concrete examples and explicit definitions of what students have to know and be able to do" to demonstrate proficiency in the skills and knowledge outlined by the content standards.[31] Performance standards are best represented by showing pieces of student work that illustrate the quality of an essay or the demonstration of mastery that is expected. In practice, however, performance standards are often expressed simply as *cut scores* on a test. For example, students might be required to get 85 percent of the items on a test correct to meet the proficiency standard. Unfortunately, the ideal of "concrete examples and explicit definitions" has been compromised by the

use of cut scores so that the connection between performance standards and content standards is much less obvious and less transparent.

Typically, cut scores are set by panels of judges, usually community leaders as well as educators. Various standard setting procedures are used to help panelists make their judgments; basically the process involves asking each panelist what percentage of items should be answered correctly to demonstrate proficiency, and computing the average of these cut scores across panelists. The procedures for setting cut scores are not scientific, and do not lead to the estimation of some true proficiency standard. Results can vary dramatically depending on whether judges are shown multiple-choice or open-ended items and whether they are asked to set "world class" or grade-level passing standards.[32] Not surprisingly, some states have thus set proficiency standards that only 15 percent of their students can pass and others have set standards that 90 percent of their students can pass. A recent study commissioned by the National Center for Education Statistics verified that these differences were caused by differences in the stringency of the standards, not by real differences in student performance.[33]

Reporting improvements in percent proficient has been the standard metric for tracking progress since the beginning of the standards movement. But, percent proficient does not tell us much if proficiency is defined so differently by states. And, reporting in relation to a proficiency cut score has created the problem of focusing on bubble kids,[34] which would not occur if indices were used that accounted for the status and growth of every student. In addition, statisticians have demonstrated that comparing proficiency percentages can create a very misleading picture of whether gaps are actually shrinking for the majority of students in the reporting groups.[35] For example, if cut scores are set either very high or very low, the gaps between groups appear to be small; conversely, the gaps between groups appear quite large when cut scores are set in the middle of the test score range. Because of limitations of the percent proficient metric, most studies of achievement trends also use some other metric, such as effect sizes, to quantify achievement changes over time.[36]

Policy makers are also aware of the problem that traditional *status measures*—which report the current achievement level for a given group of students—tend to reward schools serving affluent neighborhoods rather than creating incentives to ensure that all students receive the help they need to make significant progress. Progress on status measures is then evaluated using successive cohorts of students, for example by comparing this year's fourth graders to last year's fourth graders. Schools in communities similar to Beverly Hills, Shaker Heights, and Scarsdale do relatively well with status measures, even if the quality of instruction may only be mediocre, because students from advantaged backgrounds enter school with higher levels of achievement and continue to receive additional resources from outside of school.

Recent interest in growth measures and value-added models represent efforts by policy makers and technical experts to try to create reporting metrics that are more likely to capture the educational contribution of specific schools and districts. *Growth measures* follow the same cohort of students across years, and are able to show, for example, how much fifth graders have gained compared to their performance as fourth graders the previous year. The amount of gain can then be evaluated depending upon whether students are gaining at a rate that is faster or slower than the typical rate. *Value-added models* are complex statistical procedures used in conjunction with growth data and are intended to quantify how much each teacher or school has contributed to a student's growth in comparison to the average rate of growth. Although there are serious questions about whether research would support the high-stakes use of value-added models to make decisions about individual teachers,[37] it is clear that some indicator of student growth would add important information to accountability systems beyond that provided by status measures alone.

Assessments

Regardless of intentions, each new wave of educational reform has had to face the problem that high-stakes tests strongly influence what is taught. The authors of *A Nation at Risk* lamented the pernicious effect of minimum competency testing "as the 'minimum' tends to become the 'maximum,' thus lowering educational standards for all."[38] Yet, reform legislation in nearly every state subsequently mandated basic skills testing that perpetuated the problem of dumbed-down instruction. In the early 1990s, advocates for standards used terms like *authentic*, *direct*, and *performance-based* to argue for fundamentally different kinds of assessments that would better represent ambitious learning goals requiring complex analysis and demonstration of skills rather than just recall and recognition of answers. The idea of alignment between assessments and standards was meant to ensure that assessments would, indeed, measure learning goals represented in the content standards. Unfortunately, in practice, alignment has been claimed whenever test items fit somewhere within the standards framework rather than

asking the more important question: Do all of the test items taken together reflect the full reach of what was intended by the content standards?

In the most comprehensive study completed since NCLB, test items were compared to state content standards in each of nine states in mathematics and English, language arts, and reading, and in seven states for science.[39] On average across states, the content and cognitive demand in mathematics matched only 30 percent of the standards' expectations at fourth grade and only 26 percent at eighth grade. The corresponding figures in English, language arts, and reading were 19 percent (fourth grade) and 18 percent (eighth grade), and in science, 24 percent (fourth grade) and 21 percent (eighth grade). Some state assessments agreed with their own state content standards as much as 43 percent and others as little as 9 percent. Consistent with earlier critiques of standardized tests, this mismatch occurred primarily because tests tapped lower levels of cognitive demand than intended by the standards. Especially in mathematics, three-quarters of tested content was at the procedural or recall level of cognitive demand. Thus, standards-based reform rhetoric has not yet produced the envisioned reforms of assessments needed to measure higher order thinking abilities—such as data analysis and generalization, reasoning from evidence and being able to identify faulty arguments, drawing inferences and making predictions, or the capacity to synthesize content and ideas from several sources.

Teacher Professional Development

It was recognized at the beginning of the standards movement that teaching much more challenging curricula to all students was a tall order. It would mean providing all students with rich and engaging instructional activities that previously had been offered only to more academically advanced students. Because this vision would require fundamental changes in instructional practices, capacity building and teacher professional development were seen as key ingredients in support of reforms.[40] Unfortunately, these expectations were rarely translated into policy. Few states invested in training to help teachers teach rigorous subject matter in engaging ways. Even in states like Kentucky, which invested in teacher professional development, training was limited.[41] In most cases, policy makers relied on the state tests to convey changes that were needed. Although this was sometimes straightforward—for example, adding writing tests increased the amount of instructional time devoted to writing—accountability tests did not help teachers learn how to teach for conceptual understanding. Recent surveys still document significant capacity issues at both state and district levels.[42] Teachers lack the training to interpret data about their students and often do not know how to adapt instruction for struggling students.[43] They also may not themselves know enough about the discipline they are teaching and about methods for teaching in that discipline (especially in the case of mathematics and science) to be able to teach in ways that are both engaging and conceptually deep.

Solutions to the capacity problem are likely to be costly. According to recent summaries of evaluation studies, effective professional development programs can be neither brief nor superficial. Effective programs—those that changed teaching practices and improved student outcomes—focused on both content knowledge and particular aspects of content mastery related to student learning; they were coherently linked to curricular expectations, involved the sustained participation of teachers over long periods of time, and allowed teachers the opportunity to try new methods in the context of their own practice.[44] The need to ensure that beginning teachers are adequately prepared to teach challenging curriculum is equally great. Studies of initial teacher preparation find, for example, that program features such as curriculum familiarity and supervised opportunities to gain experience with specific classroom practices account for significant differences in the effectiveness of first-year teachers.[45]

School and System Accountability

Although the vision of standards-based reform called for the redirection of effort at every level of the educational system, accountability requirements have been focused primarily on the individual schools. The school as the locus for improvement has a legitimate basis in research. Research on effective schools, for example, documents that schools with a sense of common purpose and emphasis on academics can produce student achievement well above demographic predictions.[46] But, this research often relied on case studies of exceptional schools.

It has become increasingly clear that poor-performing schools are not able to address the problems that reflect the larger context of which they are a part. Great inequities exist among schools in resources, in the needs of the students they serve, and in the qualifications of the teachers and leaders they are able to attract and retain. Recent findings from the Programme for International Student Assessment (PISA) indicate that socioeconomic background factors have a much bigger impact on student performance in the United States than in most other countries.[47] Research evidence is accumulating to suggest that school inequities in the United States are exag-

gerating existing socioeconomic differences and may be contributing to Black–White and Latino–White gaps in test scores.[48] Schools are much more unequally funded in the United States than in high-achieving nations.[49] Furthermore, large numbers of lower socioeconomic status and minority children are attending increasingly segregated schools, and such schools have difficulty recruiting and retaining high-quality teachers and suffer other resource limitations as well.[50]

Studies focused specifically on the impacts of accountability have documented that school-based accountability mechanisms can indeed be a formula for the rich getting richer. Better-situated schools serving higher socioeconomic neighborhoods with higher quality academic programming are more able to respond coherently to the demands of external accountability.[51] High-performing schools, for example, already have in place the kind of instruction that is needed and thus can redirect the effort of well-qualified teachers to make sure that *all* students are able to meet the standards. In contrast, schools with a high concentration of poor performing students have to try to put in place for the first time the kinds of academic structures that are needed but frequently those schools lack the expertise and resources to do so.[52]

As early as 1999, the National Research Council Committee on Title I Testing and Assessment called attention to the problem of placing too great a weight on schools with limited capacity to respond. As was suggested a decade ago, this imbalance needs to be redressed. To be sure, teachers and school administrators should be held accountable for their part in improving student learning. But equally so, "districts and states should be held accountable for the professional development and support they provide teachers and schools to enable students to reach high standards."[53] In addition, states are responsible for redressing the large inequities among socioeconomic groups that exist before students enter school and that persist throughout. Researchers have estimated, for example, that fully half of the Black–White achievement gap that exists at twelfth grade could be erased by eliminating the differences that exist when children start school.[54]

Recommendations

Standards-based education is still the core idea guiding education policy and education reform. But the foregoing issues need to be addressed if the promises of standards-based education are to be kept. As yet, neither state content standards nor state tests reflect the ambitions of standards-based reform rhetoric, and the link between high expectations for all students and capacity building has been almost forgotten. The intentions of standards-based education—to focus greater attention on student learning, to ensure the participation and success of all students, and to provide guidance for educational improvement—are in the best interest of the country. We know enough to create a new generation of policies, tests, and curricula that will focus greater attention on learning and will reduce the amount of effort spent preparing students for tests that do not adequately reflect the conceptual goals of instruction.

RECOMMENDATION 1: The federal government should encourage the redesign and clear connection of content and performance standards—and the curricula, teacher training, and high-quality assessments to go with them—with the goal of developing clearly articulated statements of the expected progression of learning. Efforts to develop these components may involve partnerships among states, universities, groups of teachers, scholars, and the private sector.

In a well-functioning, standards-based system, all of the components of effective instruction—teaching, curriculum, professional development, assessments—are keyed to the content standards. The standards are coherent and organized around important ideas, skills, and strategies.

Curricula provide teachers and students with a roadmap for how to reach proficiency, acquiring and extending knowledge and skills along the way. Professional development is designed to help teachers move students toward mastery of the standards. Assessments fully represent the standards and ask students to complete tasks that draw on and measure their knowledge of the content, procedural skills, understanding, and ability to apply what they know to new situations. In such a system, teachers do not have to try to decipher the meaning of poorly articulated standards, guess what will be stressed on assessments, or have their students practice narrow, test-like items.

Some states create curriculum frameworks to help teachers in planning units of instruction. To be most useful, these frameworks should be built around established progressions for how students grow in comprehension and skill. Effective learning progressions reflect an understanding of how children learn and what students already know.[55] Although state standards have been in use since the late 1980s, and scholarly work on progressions has made significant strides in recent years, there has been little attention in the United States to incorporating the most up-to-date thinking about cognition and learn-

ing progressions into curriculum materials and assessments.

It is possible for learning to progress in a number of different ways. In this country, for example, fractions are taught before decimals. But in many other countries, decimals are taught before fractions or at the same time. It is possible that either approach may work, but whichever it is, the progression and its underlying rationale and strategy must be carefully articulated and then supported with instructional guidance and appropriate assessments.[56]

In most current practice, content standards are developed in each state and then turned over to test developers to construct state tests that are more or less consistent with the state standards (often less, as noted above). In contrast, the National Research Council report on state science assessments proposed a fundamentally different approach focused on coherence. A successful standards-based science assessment system would be *horizontally coherent* by having curriculum, instruction, and assessment all keyed to the same learning goals; *vertically coherent* by having the classroom, school, school district, and state all working from the same shared vision for science education; and *developmentally coherent* by taking account of how students' understanding of science develops over time.[57] To achieve this kind of coherence, states may need to consider more integrated, concurrent ways of developing standards and assessments. For example, the intention of standards might be conveyed better if they were accompanied from the beginning by prototypes of assessment tasks. More importantly, the development of learning progressions across grades requires empirical testing. Learning progressions as they have been developed in other countries are based on research and professional judgment about the logical sequence of skills and topics, followed by empirical verification as curricular materials are developed, tested, and revised.[58]

President Obama said in March 2009 that he wanted states to adopt "tougher, clearer" standards that rival those in countries where students out-perform their U.S. counterparts. He called on states to join consortia to "develop standards and assessments that don't simply measure whether students can fill in a bubble on a test, but whether they possess 21st-century skills like problem solving and critical thinking and entrepreneurship and creativity."[59]

With the governors and chief state school officers embarked on a joint effort to develop common standards, consortia of states may still be needed to take the next step of developing deeper grade-level curricula, models of teacher professional development, and complementary instructional resources—including illustrations of the kind of tasks that will demonstrate mastery of the standards and providing detailed guidance on assessments. These consortia might involve universities, professional associations, subject-matter experts, think tanks, or other entities, but an important distinction should be drawn between the political process needed to achieve consensus and guide policy decisions versus the scientific expertise needed to develop and rigorously evaluate curricular materials, instructional strategies, and assessments. We recommend that the federal government help a number of these development projects get off the ground, because states do not have the resources or, in many cases, the expertise, to do this on their own.

The question of standards and accountability in high schools is more complex than it is in elementary and middle schools. Students follow different academic paths through high school, with different ends in mind. In most other countries, high school examinations measure students' knowledge of subject matter *or* their readiness for vocations. In the United States, by contrast, most states require students to pass general skill proficiency exams to graduate from high school, and also use the exam results in complying with NCLB accountability provisions. High school exit exams are often unrelated to courses or curriculum and vary across states in the levels of skills measured from eighth to eleventh grade.[60]

In moving away from generic, basic skills tests at the high school level, policy makers are immediately confronted with the issue of what the curriculum should be, because higher-level critical-thinking skills are not content free. Analytical reasoning and problem solving cannot be learned or assessed in the absence of challenging content. So, going deeper requires making choices about which content to cover and what specific content can be fairly assessed. Specifically, then, states or consortia of states will face the question of whether there should be one curriculum or multiple curricula to prepare students for college and the workplace. Common standards do not resolve the question as to whether there should be multiple ways of getting there.

As states seek ways to develop more challenging curricula to engage students who have been ill-served by traditional college preparatory courses, they may wish to consider and carefully evaluate career-related courses or certification programs. Recent studies of some career systems in Europe suggest that explicit career preparation programs can have important benefits, especially those that involve youth apprenticeships or a combination of

part-time work and formal schooling leading to an occupational certificate. In some countries, these programs have high completion rates and support more rapid transitions to employment.[61] Importantly, they could also be a means for providing more authentic learning contexts, which—according to cognitive science research studies—increase learning by helping students draw connections and see why things work the way they do. According to the National Research Council Committee on Increasing High School Students' Engagement and Motivation to Learn, career academies and other occupational-themed programs can improve student motivation and engagement, but only if academic courses are well structured to ensure a wide range of competencies and are integrated well with meaningful work placements.[62] Great care must be taken, however, to avoid old notions of vocational education or dead-end low-ability tracks. In a comparative study, for example, dual system curricula in Austria, Germany, and Switzerland were found to have much greater depth of content in mathematics and science, greater integration of academic and applied content, and higher demand for cross-disciplinary higher-order skills than typical high school curricula in the United States.[63] Furthermore, these features are associated with academic performance in the European countries that matched or exceeded U.S. performance.

Ten states currently use end-of-course exams for accountability purposes, and other states have plans to implement them.[64] These exams differ significantly, however, from the course assessment systems in most high-achieving nations. In contrast to most end-of-course tests in the United States, high school assessments in Australia, Finland, Hong Kong, the Netherlands, Singapore, Sweden, and the United Kingdom—among others—are generally developed by high school and college faculty and comprise largely open-ended questions and prompts that require considerable writing, analysis, and demonstration of reasoning. Most also include intellectually ambitious tasks that students complete during the course, such as science investigations, research papers, and other projects that require planning and management as well as the creation of more extensive products. These tasks are incorporated into the examination score. Finally, the examinations are used to inform course grades and college admissions, rather than to serve as exit exams from high school, which allow them to reflect more ambitious standards.[65]

Advances in assessment may be undertaken by consortia of states—as suggested above—that could work on high school standards, curricula, and related exams. Although we are not yet able to evaluate the quality and rigor of its products, efforts by Achieve to create a common end-of-course exam for algebra for participating states are one example of this kind of approach.[66] A similar effort, perhaps involving employers, could be undertaken to develop certification standards and exams for different careers. Indeed, such certification may provide an incentive for students to stay in school. Because a significant policy question to be addressed is whether students benefit most from a single college-preparatory curriculum or a combination of college preparatory and career preparation options, federal investments should be made in both types of curricular models to allow for rigorous, comparative evaluations of the two systems.

RECOMMENDATION 2: The federal government should support research on accountability system indicators to reflect both the status and growth of students. Performance standards should set ambitious but realistic targets for teaching and learning, and they should communicate to the public, parents, educators, and students themselves what is to be learned. Assessment results should be reported in ways that recognize progress all along the achievement continuum.

In order to have a constructive influence on the behavior of students, educators, and schools, accountability indicators must clearly communicate both what is to be learned and how students are progressing in their learning, as well as illustrating what it means to be proficient. Accountability reporting systems also should indicate how much students are improving on a range of indicators of learning, how they are progressing through school to graduation, and what kinds of resources are available to them. When targets are set to help in evaluating the rate of growth, they should be ambitious but reachable. As basic as these criteria might seem, many state accountability systems do not meet them at present.

The problems with judging schools or teachers using only a percent proficient criterion are now much better understood. Prior to standards-based reform, schools were judged by whether their test scores were above or below average. Comparisons based on averages were unpopular with policy makers because they implied complacency with being mediocre and did not provide any substantive insight into desired levels of performance. Now that the weaknesses of proficiency scores are also understood, alternative metrics should be considered. For example, reporting gains by comparing means is the reporting unit preferred by statisticians because means take account of all of the student scores. And, for those who want additional comparative standards in or-

der to decide whether mean scores are "good enough," the meaning of averages can be augmented substantively by benchmarking to international comparisons along with sample tasks that illustrate performance capabilities at different levels.

The requirement in NCLB that states define Adequate Yearly Progress (AYP) and use it as the basis for holding schools and districts accountable sounds eminently reasonable. Yet, AYP changed the meaning of "adequate" in a misleading way that threatens to undermine the credibility of the accountability system. In NCLB, the term adequate was not defined as normal or even exemplary progress; rather, it was based on a calculation of the rate of progress needed to get to 100 percent proficiency by the year 2014. Even if the deadline was a long way off, 100 percent proficiency was not a reasonable goal. The improvement curve was very steep, especially for second-language learners and special education subgroups who, by definition, need special help to participate fully in regular instruction, but still were required to reach the same target. Critics of this aspect of the law have argued that standards-based reforms could establish much more ambitious goals than have previously been achieved, especially for low-performing and at-risk groups, but should nonetheless set targets that are realistic. The idea of an *existence proof criterion*[67] means that there should be at least one example of a school or a district that achieved an aspirational goal before it can be mandated for everyone. For example, states might set test score targets at the 75th percentile or even the 90th percentile of what similar schools had achieved. This idea uses norms to help decide what is reasonable, but substantially ratchets up expectations rather than assuming that the 50th percentile remains a satisfactory goal.

Increasingly, states are aware that status measures of student performance reward schools that serve the most able students without necessarily reflecting the quality of education in those schools. At the same time, comparing schools based on similar demographics or growth appears to set lower expectations for schools serving poor and minority communities. By examining both indicators, accountability systems can give credit for significant growth, and at the same time attend to the fact that desired performance goals still have not been met. Schools low on both status and growth measures should then receive the greatest scrutiny and assistance.

Because tests and accountability indices based on tests are fallible, accountability reporting systems are less likely to lead to distortions and perverse incentives that focus effort on the wrong outcomes if they attend to multiple sources of evidence. Reporting *both* growth and status is one example of how the reporting system can be improved to support more valid judgments about the quality of schooling. Other ways to improve an indicator system include using multiple measures of reading and math—such as portfolios along with standardized tests, measures of academic goals beyond reading and math, indicators of school climate and non-academic goals, and tracking of progress for significant subgroups. For example, one of the most successful aspects of NCLB reporting has been the disaggregation of test score results and reporting of progress for each subgroup (e.g., major racial/ethnic groups, economically disadvantaged students, students with disabilities, English-language learners, etc.).[68] Although there have been flaws in the specifics—especially the problems of small sample sizes and misleading impacts when the same at-risk students are counted multiple times in overlapping subgroups—the effort to focus attention on historically low-performing and neglected subgroups is generally regarded as one of NCLB's greatest successes.

The federal government should support the development of several different kinds of reporting systems, and once in use they should be studied to see which approaches work the best for different purposes. For example, when Florida wanted to increase attention to learning gains for students scoring in the lowest 25 percent of students, they added this component to their formula for determining school grades, and in 2010 they will add a component for high school students' participation in accelerated coursework. Accountability reporting systems should be evaluated both in terms of the incentives they foster and the information they provide for improving instruction. States or consortia of states should recognize that this is a complex issue requiring careful analysis. Although the basic ingredients can be decided politically—whether graduation rates should be included, for example—the mechanics of how they are reported, whether they are compared to target or comparative criteria and whether they are combined into a single index should be worked out carefully and tested in trial sites. States would be well-advised to assemble teams of individuals with both content and technical expertise to assist them in this process. In general, we recommend that compensatory models be used rather than the current conjunctive model used in NCLB. Compensatory models allow for strengths in one area to offset weaknesses in another area, at least to a certain degree, whereas with conjunctive policies, failing on any one dimension means failing on all of them.[69] Because of the concern that compensatory, composite indices could once again make it possible for schools to use the performance of high-achieving students to hide the failure of low-performing students, reporting systems should require separate reporting by

subgroups or build in specific checks to guard against this type of abuse.

RECOMMENDATION 3: The federal government should support the redesign and ongoing evaluation of accountability systems to ensure that they contribute to school improvement. Less than satisfactory school performance should trigger closer investigation of school operations before remedies or sanctions are applied, and stellar performances should also be verified. Different investigative approaches, including audit assessments, data-driven analyses, or expert constituted inspectorates, should be considered.

It is now a familiar story in nearly every state to read about schools that are "excellent" according to the state accountability system but "in need of improvement" under NCLB. Current NCLB policies identify so many schools in the United States as failing that the number is virtually meaningless. That number will continue to grow as schools and districts fall short of the greater and greater leaps that will be required to reach 100 percent proficiency by 2014. Already, more than a third of U.S. schools are not meeting their NCLB targets. Researchers in California report that all of the state's elementary schools will eventually be considering low achievers.[70] Such schools could be required, under current law, to offer tutoring, allow students to transfer, have their faculty and principal replaced, be turned into a charter school, or be subject to other interventions.

Creating more appropriate accountability indices and using more scientific means to establish defensible targets, as suggested above, will help to address the greatest threats to validity and credibility of accountability systems. In addition, accountability systems should be designed with built-in, self-checking mechanisms and should be evaluated to determine whether the information they provide and subsequent actions are, indeed, improving the education system. Any measure of a school's performance only hints at what is going on inside its classrooms. Test scores alone cannot constitute definitive evidence as to the extent of a school's success or failure. Suppose that test scores in mathematics for students in one elementary school have risen dramatically. How can we know if this is a result of test score inflation or exemplary teaching practices? Should teachers in that school receive bonuses and be visited by other teachers who want to learn about their teaching strategies? Or would it be better first to verify that students from that school are doing well in middle school mathematics and also succeed on an open-ended problem solving assessment administered by the district? What if nearly all of the schools in a particular district show growth rates for English-language learners below the growth rate for similar students elsewhere in the state? How should the locus of the problem be identified and what should the mechanism be for marshalling the needed resources if it is determined that neither schools nor the district have the know-how to provide an adequate remedy? It would be far too costly to try to collect multiple measures of achievement and gather meaningful data on curriculum materials, teacher qualifications, and school climate and safety for every school every year. But, more complete evidence could be collected to verify successes and failures and their likely causes if in-depth investigations were limited to a small sample of schools.

The federal government should support states that want to experiment with the development of two-stage accountability systems, whereby initial test-score results would serve as a trigger, prompting closer examination of performance and educational quality in targeted schools. Persistently low scores could prompt an evaluation designed to identify both educational and external factors that are influencing scores and would help to clarify whether the trends in scores really should be taken as a sign of program effectiveness.[71] The kinds of evidence collected in the second stage could be some form of audit assessment, data-driven analyses of existing data, or visits by expert inspectorates. We recommend, in particular, that the federal government encourage the states to experiment with ways of introducing an element of human judgment into making decisions about which schools merit an aggressive turnaround effort as well as the substantive focus of such efforts.

An important goal of a two-stage approach should be to ensure that attributions of exceptional or poor educational quality are not made from test score data alone. In addition, gathering greater depth of information on a sample of schools will make it possible to evaluate the accountability system and continuously improve its measures and incentive structures.

RECOMMENDATION 4: The federal government should support an intensive program of research and development to create the next generation of performance assessments explicitly linked to well-designed content standards and curricula.

The field of educational measurement, closely tied to research in cognitive science, has already begun to develop major new solutions that will address several of the sub-

stantive issues raised in this paper. With a significant, targeted investment in developing new assessment tools, psychometric and content experts could—within a 5–10-year period—provide a set of cognitive-science based tools to guide students and teachers in the course of learning. Development efforts should include systematic examination of assessments and curricula from high-achieving countries with special attention to assessment tasks that reflect higher-order thinking and performance skills. As new measures are developed, they could then, in turn, be tested for use in accountability systems with particular attention to their ability to withstand distortion in high-stakes settings.

The program of development and research we recommend here would create greater conceptual coherence between what is assessed externally for accountability purposes (e.g., at the state level) and the day-to-day assessments used in classrooms to move learning forward. Enough is known about potential new forms of assessment that an intensive engineering research program, with short cycles of development, field testing, and revision could lead to dramatic improvements within a decade. Such efforts would alter the nature of assessment information, but more importantly would reshape for the better the ways that the character of assessments influences the character of education.

A national program of evidence-based assessment development should be launched as quickly as possible. The needed development and evaluative research should be carried out in multiple laboratories and field tryout sites. However, the program as a whole should be overseen by a single agency so as to promote maximum sharing of ideas, potential solutions, and interim results that might be tried out in operational assessments. A number of promising assessment tools already exist in prototype or experimental form. With appropriate federal government support, these tools might be brought together with curriculum development aimed at higher-level cognitive skills and then be moved into wider-scale tryouts, followed by refinements both to raise reliability and to scale back costs. The goals for such an effort should include the following:

- **Produce learning progressions and assessments that measure both content knowledge and higher-order, problem solving skills.** It is relatively easy to measure content knowledge. But, skills such as adapting one's knowledge to answer new and unfamiliar questions, are difficult to measure easily and reliably. Moreover, for such assessments to be fair and useful, they must be tied to reasonably well-documented learning progressions that demonstrate how students' increasing competence can be supported and advanced. Learning progressions underlying performance-based testing are deeply substantive and go well beyond vertical alignment of traditional, multiple-choice tests. In just the past decade, significant progress has been made in the development of prototype learning progressions in the mathematics and science domains,[72] but because each major concept, inquiry skill, and problem-solving strategy requires analysis and testing to ensure that it can be used effectively in the classroom, a great deal of work remains to be done. Such a program of research cannot be focused on the design of new assessments in isolation; rather, it requires concomitant investments in methodological, cognitive, and subject matter research as well.

- **Accurately and fairly assess English-language learners and students with special needs.** In response to NCLB, the U.S. Department of Education funded state consortia to develop new measures of English language proficiency, which significantly advanced the field. States have adopted accommodation policies to allow less impaired special education students to participate in regular state assessments, and have developed alternative assessments to assess more severely affected students. Basic psychometric issues of reliability have been addressed and progress has been made in recognizing some of the most critical validity issues. For example, there is now much greater understanding of the importance of assessing academic language to enable academic success for second-language learners.[73] Much remains to be done, however, to fairly assess language development and to link academic language demands with progress in mastering specific curriculum content. Second-language learners are vastly different from one another in terms of the first languages they speak and the amount of formal instruction they have received in each of their languages.[74] Each of these differences would require specially tailored assessments to ensure validity. Although it is true, for example, that any student is disadvantaged if the vocabulary demands of a mathematics test do not align well with their textbook, the effects on validity and fairness for second-language learners are much greater. Therefore, more focused research and development is needed to examine the potential validity of both curriculum-linked assessments (which would allow second-language learners to take the same tests as their classmates) and of specially tailored, individualized, classroom-based assessments that could be validly aggregated for large-scale accountability purposes.

Similarly, newly developed alternative assessment measures for students with disabilities have met their primary goal of including hitherto excluded students in the accountability system. But additional work is needed to make sure that accommodations do not artificially inflate test performance and to verify the learning progressions underlying score scales on alternative assessments. Do gains on alternative assessments represent meaningful increments in knowledge and skills for these populations of students that can be used to set instructional goals?

- **Identify, test, and expand the availability of technology to capture important learning goals, enhance validity, and reduce the cost of assessments.** Significant advances have been made in the use of technology to support traditional large-scale assessment programs, and important research and development programs are already under way. For example, the NAEP and the PISA have already done pilot studies of online assessments in mathematics and writing and in science, respectively. Computer scoring of essays is now as reliable as that by expert graders—at least for basic compositions—and is much less costly. Computerized adaptive testing is already used for high-stakes individual assessments on tests such as the Graduate Record Examination, and has great potential to support both large-scale and classroom-level applications of learning progressions because it allows for each student to be tested in greater depth at his or her own particular level of mastery. Of special interest are breakthrough, next-generation assessments where technology is helping to tap complex and dynamic aspects of cognition and performance that previously could not be assessed directly.[75] These cutting-edge developments show how technology can be used to engage students in developing models of scientific phenomena, analyzing data, and reasoning from evidence. This type of work is relatively new and has not been tested in wide-scale applications, but with further development and evaluation, these technology-based assessments could ultimately be used to allow large-scale testing of important analytical and problem-solving skills.

- **Develop valid and useful measures of classroom teaching and learning practices.** Accountability systems drive what both students and teachers do. But changes in teaching practices are presently documented only by test scores, and research on teacher learning warns us that it is much easier to improve scores on traditional tests by teaching the test rather than by making the fundamental changes needed to improve students' long-term learning. There is at present no direct way to measure changes in instruction that would withstand the requirements of high-stakes use. Although research on classroom observational instruments is limited, a great deal is known from cognitive science research about what to look for in classrooms to see if students are being supported to reach more ambitious learning goals, and measures of these features could be reliable enough to be useful for formative program improvement and research purposes. For example, observational ratings could document whether students are aware of learning goals, are actively engaged and take responsibility for their own learning, and whether it is part of classroom norms for students to be regularly called on to explain their thinking.[76] Private foundations (e.g., Bill & Melinda Gates Foundation, William and Flora Hewlett Foundation, and William T. Grant Foundation) are now investing in the development of measures of teaching quality, and with federal assistance, much more substantial progress could be made toward solving these measurement questions.

- **Build greater conceptual coherence between assessments of student performance used for accountability purposes and classroom assessments designed to provide better instructional guidance to teachers.** To be as accurate and useful as possible, accountability tests should ask students to do tasks similar to those they are asked to do in their regular classrooms. But, it would be a mistake to attempt to achieve coherence between large-scale and classroom assessments by locking in testing formats or dominant instructional patterns from the past century. We know from video studies of international and U.S. mathematics classes, for example, that instruction in this country is dominated by practicing procedures and reviewing, in contrast to curriculum and instructional practices in other countries that are focused much more on depth of understanding, reasoning, and the generalization of knowledge.[77] Thus, improving assessments in the United States necessarily requires corresponding improvements in curricula and teaching. An intensive research program aimed at developing next-generation assessment and accountability systems must also undertake, in at least some research sites, reform of curricula, reform of instructional tasks, corresponding changes in large-scale assessments, and significant changes in both teacher preparation and professional development to help teachers teach in profoundly different ways. A next-generation system cannot be built all at once, but with federal research support, one consor-

tium of states might pursue the design of curricula, instructionally-linked learning progressions, and performance-based tasks in middle school science and try them out with accompanying teacher training. Another state or consortium might be funded to undertake similar development in writing and language arts, and so on.

This ambitious research effort should be overseen by teams of the leading cognitive scientists, subject matter experts, and measurement professionals. The recommendation here is not merely to address the shortcomings of the current systems of standards, accountability, and assessment. Rather, the recommendation is to build on developments in technology, assessment, and cognition to create a truly reformed accountability and measurement infrastructure to support the teaching and learning to which we aspire for all of our children.

Notes

[1] P.L. 107-110, 115 Stat. 1425, enacted January 8, 2002

[2] Massell, D. (2008). *The current status and role of standards-based reform in the United States*. Paper prepared for the National Research Council Workshop on Assessing the Role of K-12 Academic Standards in the States. Available at: http://www7.nationalacademies.org/cfe/Massell%20State%20Standards%20Paper.pdf

[3] Smith, M., & O'Day, J. (1991). Systemic school reform. In S. Fuhrman & B. Malen (Eds.), *The politics of curriculum and testing* (pp. 233-268). San Francisco: Jossey-Bass.

[4] Goals 3 and 4 Technical Planning Group on the Review of Education Standards. (1993). *Promises to keep: Creating high standards for American students*. Washington, DC: National Education Goals Panel.

[5] Elmore, R. F., & Rothman, R. (Eds.). (1999). *Testing, teaching, and learning: A guide for states and school districts*. Washington, DC: National Academy Press.

[6] Shields, P., Esch, C., Lash, A., Padilla, C., Woodworth, K., & LaGuardia, K. (2004). *Evaluation of Title I accountability systems and school improvement: First year findings*. Washington, DC: U.S. Department of Education.

[7] Hannaway, J., & Hamilton, L. (2009). Performance-based accountability policies: Implications for school and classroom practices. Washington, DC: The Urban Institute. Available at: http://www.urban.org/UploadedPDF/411779_accountability_policies.pdf

[8] Ibid.

[9] Carnoy, M., & Loeb, S. (2002). Does external accountability affect student outcomes? A cross-state analysis. *Educational Evaluation and Policy Analysis, 24*(4), 305-331.

[10] Raymond, M.E., & Hanushek, E.A. (2003). High-stakes research. *Education Next, 3*(3), 48-55; Braun, H. (2004). Reconsidering the impact of high-stakes testing. *Education Policy Analysis Archives, 12*(1). Available at: http://epaa.asu.edu/epaa/v12n1/; Hanushek, E.A., & Raymond, M.F. (2005). Does school accountability lead to improved student performance?, *Journal of Policy Analysis and Management, 24*(2), 297-327.

[11] Amrein, A.L, & Berliner, D.C. (2002). High-stakes testing, uncertainty, and student learning. *Education Policy Analysis Archives, 10*(18); see also Braun (2004).

[12] Kober, N., Chudowsky, N., & Chudowsky, V. (2008). *Has student achievement increased since 2002?: State test score trends through 2006-07*. Washington, DC: Center on Education Policy.

[13] Lee, J. (2006). *Tracking achievement gaps and assessing the impact of NCLB on the gaps: An in-depth look into national and state reading and math outcome trends*. Cambridge, MA: The Civil Rights Project, Harvard University.

[14] See Hamilton et al. (2007).

[15] Center for Education Policy. (2006). *Year 4 of the No Child Left Behind Act*. Washington, DC: Author; Gross, B., & Goertz, M.E. (Eds.). (2005). *Holding high hopes: How high schools respond to state accountability policies* (CPRE Research Report Series No. RR-056). Philadelphia, PA: Consortium for Policy Research in Education, University of Pennsylvania.

[16] Koretz, D. (2008). Further steps toward the development of an accountability-oriented science of measurement. In K.E. Ryan & L.A. Shepard (Eds.), *The future of test-based educational accountability* (pp. 71-91). New York: Routledge; see also, e.g., Feuer, M.J., Holland, P.W., Green, B.F., Bertenthal, M.W., & Hemphill, F.C. (Eds.). (1999). *Uncommon Measures: Equivalence and linkage among educational tests*. Washington, DC: National Academy Press for discussion of the importance of estimating the degree to which test validity may be compromised and the need to balance this information against potential benefits of test-based information.

[17] Bransford, J.D., Brown, A.L., & Cocking, R.R. (Eds.) (2000). *How people learn: Brain, mind, experience, and school*. Washington, DC: National Academy Press.

[18] See Hamilton et al. (2007).

[19] Baker, G. (2002). Distortion and risk in optimal incentive contracts, *Journal of Human Resources, 37*(4), 728-751; Feuer, M.J. (2008). Future directions for educational accountability: Notes for a political economy of measurement. In K.E. Ryan & L.A. Shepard (Eds.), *The future of test-based educational accountability* (pp. 293-306). New York: Routledge; Koretz, D.M. (2002). Limitations in the use of achievement tests as measures of educators' productivity. *The Journal of Human Resources, 37*(4), 752-777.

[20] National Research Council. (2008). *Common standards for K-12 education: Considering the evidence*. Washington, DC: The National Academies Press.

[21] P.L. 103-227, Sec. 3[4].

[22] Harris, D.N., & Goertz, M. (2008). The potential effects of "high-quality and uniform" standards: Lessons from a synthesis of previous research and proposals for a new research agenda. Paper prepared for the 2008 Workshop Series on State Standards, National Research Council, Washington, DC.

[23] NCLB, 2001, Part A, Subpart 1, Sec. 1111, a [D].

[24] McNeil, M. (2009, April 20). NGA, CCSSO launch common standards drive. *Education Week, 28*(29). Available at: http://www.edweek.org/ew/articles/2009/04/16/29standards.h28.html?qs=CCSSO+launch+common+standards+drive; McNeil, M. (2009, June 10). 46 states agree to common academic standards effort.

Education Week, *28*(33), 16. Available at: http://www.edweek.org/ew/articles/2009/06/01/33standards.h28.html?qs=46+states

[25] Porter, A.C., Polikoff, M.S., & Smithson, J. (2009). Is there a de facto national intended curriculum?: Evidence from state content standards, *Educational Evaluation and Policy Analysis*, *31*(3), 238-268.

[26] See Bransford et al. (2000); Kilpatrick, J., Swafford, J., & Findell, B. (Eds.). (2001). *Adding it up: Helping children learn mathematics*. Washington, DC: National Academy Press; Wilson, M.R., & Bertenthal, M.W. (Eds.). (2006). *Systems for state science assessment*. Washington, DC: The National Academies Press.

[27] Schmidt, W.H., Wang, H.C., & McKnight, C.C. (2005). Curriculum coherence: An examination of U.S. mathematics and science content standards from an international perspective. *Journal of Curriculum Studies*, *37*(5), 525-559.

[28] Ibid.

[29] See Porter et al. (2009).

[30] Schmidt, W.H., & Prawat, R.S. (2006). Curriculum coherence and national control of education: Issues or non-issue? *Journal of Curriculum Studies*, *38*(6), 641-658.

[31] Goals 2000: Educate America Act of 1994, P.L. 103-227, Sec 3[9].

[32] Shepard, L.A. (2008). A brief history of accountability testing, 1965-2007. In K.E. Ryan & L.A. Shepard (Eds.), *The future of test-based educational accountability* (pp. 25-46). New York: Routledge.

[33] U.S. Department of Education. National Center for Education Statistics. (2007). *Mapping 2005 state proficiency standards onto the NAEP scales* (NCES 2007-482). Washington, DC: U.S. Government Printing Office.

[34] Hamilton, L.S., Stecher, B.M., Marsh, J.A., Sloan McCombs. J., Robyn, A., Russell, J.L., et al. (2007). *Standards-based accountability under No Child Left Behind: Experiences of teachers and administrators in three states*. Santa Monica, CA: RAND.

[35] Holland, P.W. (2002). Two measures of change in the gaps between the CDFs of test-score distributions. *Journal of Educational and Behavioral Statistics*, *27*(1), 3-17.

[36] Effect-size statistics were developed to help convey the practical significance of the difference between two group means and to make it possible to compare differences on lots of different tests, all with different score scales. An effect size is calculated by finding the difference between two means and then dividing that difference by the test-scale standard deviation. See Kober et al. (2008).

[37] Harris, D.N. (2008). The policy uses and "policy validity" of value-added and other teacher quality measures. In D.H. Gitomer (Ed.), *Measurement issues and the assessment of teacher quality* (pp. 99-132). Thousand Oaks, CA: SAGE.

[38] National Commission on Excellence in Education. (1983). *A nation at risk: The imperative for educational reform*. Washington, DC: U.S. Department of Education, p. 2.

[39] Polikoff, M., Porter, A.C., & Smithson, J. (2009). The role of state student achievement tests in standards-based reform. *Working Paper*. Philadelphia: Graduate School of Education, University of Pennsylvania.

[40] See Smith & O'Day. (1991); McLaughlin, M.W., & Shepard, L.A. (1995). *Improving education through standards-based reform*. Stanford, CA: National Academy of Education.

[41] McDonnell, L.M. (2004). *Politics, persuasion, and educational testing*. Cambridge, MA: Harvard University Press.

[42] See Massell. (2008).

[43] See National Research Council. (2008).

[44] Blank, R.K., de las Alas, N., & Smith, C. (2008). *Does teacher professional development have effects on teaching and learning?: Analysis of evaluation findings from programs for mathematics and science teachers in 14 states*. Washington, DC: Council of Chief State School Officers; Desimone, L.M., Porter, A.C., Garet, M.S., Yoon, K.S., & Birman, B.F. (2002). Effects of professional development on teachers' instruction: Results from a three-year longitudinal study. *Educational Evaluation and Policy Analysis*, *24*(2), 81-112; Garet, M.S., Porter, A.C., Desimone, L., Birman, B.F., & Yoon, K.S. (2001). What makes professional development effective? Results from a national sample of teachers. *American Educational Research Journal*, *38*(4), 915-945; Putnam, R.T., & Borko, H. (2000). What do new views of knowledge and thinking have to say about research on teacher learning? *Educational Researcher*, *29*(1), 4-15.

[45] Boyd, D., Grossman, P., Lankford, H., Loeb, S., & Wyckoff, J. (2008, August). *Teacher preparation and student achievement*. Teacher Policy Research. Available at: http://www.teacherpolicyresearch.org/portals/1/pdfs/Teacher%20Preparation%20and%20Student%20Achievement%20August2008.pdf

[46] Brookover, W.B., Beady, C., Flood, P., Schweitzer, J., & Wisenbaker, J. (1979). *School social systems and student achievement: Schools can make a difference*. New York: Praeger; Edmonds, R. (1979). Effective schools for the urban poor. *Educational Leadership*, *37*, 15-24; Purkey, S.C., & Smith, M.S. (1983). Effective schools: A review. *The Elementary School Journal*, *83*(4), 426-452; Rutter, M., Maughan, B., Mortimore, P., Ouston, J., & Smith, A. (1979). *Fifteen thousand hours: Secondary schools and their effects on children*. Cambridge, MA: Harvard University Press.

[47] Programme for International Student Achievement (PISA). (2007). *PISA 2006: Science competencies for tomorrow's world, Volume 1: Analysis*. Paris, France: Organisation for Economic Co-operation and Development. Available at: http://www.oecd.org/dataoecd/30/17/39703267.pdf

[48] Fryer, R.G., & Levitt, S.D. (2006). The Black-White test score gap through third grade. *American Law & Economics Review*, *8*(2), 249-281.

[49] See PISA. (2007); Organisation for Economic Co-operation and Development (OECD) (2008). Education at a glance: OECD indicators, 2007. Paris: OECD; OECD (2007, December), OECD briefing note for the United States. Available at: http://www.oecd.org/dataoecd/16/28/39722597.pdf

[50] Clotfelter, C.T., Ladd, H.F., Vigdor, J., & Wheeler, J. (2007). High poverty schools and the distribution of teachers and principals. *University of North Carolina Law Review*, *85*(5), 1345-1380.

[51] Elmore, R. (2003). Accountability and capacity. In M. Carnoy, R. Elmore, and S. Siskin (Eds.), *The new accountability: High Schools and high stakes testing* (pp 195-210). New York: Routledge Falmer; see also Bryk, A.S., Sebring, P.B., Allensworth, E., Luppescu, S., & Easton, J.Q. (2009). *Organizing schools for improvement: Lessons from Chicago*. Chicago: University of Chicago Press.

[52] Siskin, L.S. (2004). The challenge of the high school. In S.H. Fuhrman & R.F. Elmore (Eds.), *Redesigning accountability systems for education* (pp. 167-188). New York: Teachers College Press.

[53] See Elmore & Rothman. (1999). P. 6.

[54] Phillips, M., Crouse, J., & Ralph, J. (1998). Does the Black-White test score gap widen after children enter school? In C. Jencks & M. Phillips (Eds.), *The Black-White test score gap* (pp. 229-272). Washington, DC: Brookings Institution Press.

[55] Pellegrino, J.W., Chudowsky, N., & Glaser, R. (Eds). (2001). *Knowing what students know: The science and design of educational assessment.* Washington, DC: National Academy Press.

[56] See Wilson & Bertenthal. (2006).

[57] Ibid.

[58] Masters, G., & Forster, M. (1996). *Progress maps: Assessment resource kit*. Melbourne, Australia: The Australian Council for Educational Research.

[59] Transcript: President Obama's Remarks to the Hispanic Chamber of Commerce. (2009, March 10). *New York Times.* Available at: http://www.nytimes.com/2009/03/10/us/politics/10text-obama.html?_r=1

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

[61] Quintini, G., Martin, J.P., & Martin, S. (2007). The changing nature of the school-to-work transition process in OECD countries, Discussion Paper Series, IZA DP No. 2582. Bonn: Institute for the Study of Labor; Ryan, P. (2001). The school-to-work transition: A cross-national perspective. *Journal of Economic Literature*, *39*(1), 34-92.

[62] National Research Council, (2004). *Engaging schools: Fostering high school students' motivation to learn*. Washington, DC: The National Academies Press.

[63] King, J. (2003). Academic quality and vocational education: Evaluating dual system vocational curricula. A report prepared by the United States-European Network for Education and Training for the National Assessment of Vocational Education. As cited in U.S. Department of Education, Office of the Under Secretary, Policy and Program Studies Service. (2004). National Assessment of Vocational Education: Final Report to Congress. Washington, DC: Author. Available at: http://www.ed.gov/rschstat/eval/sectech/nave/navefinal.pdf

[64] Vranek, J., de Barros, J., Brown, R., Gross, B., & Mazzeo, C. (2007). *Independent study of state end-of-course assessments.* Seattle: Education First Consulting.

[65] Darling-Hammond, L., & McCloskey, L. (2008). Assessment for learning around the world: What would it mean to be internationally competitive? *Phi Delta Kappan*, *90*(4), 263-272.

[66] Achieve is an independent non-profit education organization created in 1996 by governors and state leaders to help states raise academic standards and graduation requirements, improve assessments, and strengthen accountability. A description of the Algebra I and II end-of-course exams created by a consortium of states in partnership with Achieve is available at: http://www.achieve.org/ADPAssessmentConsortium.

[67] Linn, R.L. (2003). Performance standards: Utility for different uses of assessments. *Education Policy Analysis Archives*, *11*(31). Available at: http://epaa.asu.edu/epaa/v11n31/

[68] The National Academies Committee on Incentive and Test-Based Accountability in Public. (2007, November 16). Workshop on Multiple Measures (transcript). Available at: http://www7.nationalacademies.org/bota/Introductions.pdf

[69] Ibid.

[70] Bryant, M.J., Hammond, K.A., Bocian, K.M., Rettig, M.F., Miller, C.A., & Cardullo, R.A. (2008, September 26). Assessment: School Performance will fail to meet legislated benchmarks. *Science*, *321*, 1781-1782.

[71] See Koretz. (2002).

[72] See Pellegrino et al. (2001); Wilson & Bertenthal. (2006).

[73] Abedi, J. (2007). *English language proficiency assessment in the nation: Current status and future practice*. Davis, CA: University of California, Davis, School of Education.

[74] Solano-Flores, G. (2008). Who is given tests in what language by whom, when, and where?: The need for probabilistic views of language in the testing of English language learners. *Educational Researcher*, *37*(4), 189-199.

[75] Quellmalz, E.S., & Pellegrino, J.W. (2009, January 2). Technology and testing. *Science*, *323*, 75-79.

[76] Gollub, J.P., Bertenthal, M.W., Labov, J.B., & Curtis, P.C. (Eds.). (2002). *Learning and understanding: Improving advanced study of mathematics and science in U.S. high schools*. Washington, DC: National Academy Press.

[77] Given, K., Jacobs, J., & Hollingsworth, H. (2006). What does teaching look like around the world? *ON-Math*, *4*(1). Available at: http://my.nctm.org/eresources/view_article.asp?article_id=7396&page=1

This manuscript is a product of the **Education Policy White Papers Project**, an initiative of the **National Academy of Education (NAEd)**. The goal of this project is to connect policymakers in the Administration and Congress with the best available evidence on selected education policy issues: equity and excellence in American education; reading and literacy; science and mathematics education; standards, assessments, and accountability; teacher quality; and time for learning.