STANDARDS, ASSESSMENTS, AND EDUCATIONAL POLICY: IN PURSUIT OF GENUINE ACCOUNTABILITY

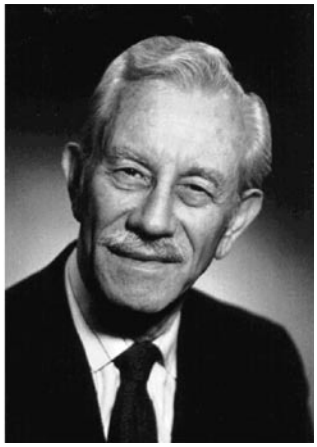
By Linda Darling-Hammond

WILLIAM H. ANGOFF MEMORIAL LECTURE Series

Policy Evaluation & Research Center

Policy Information Center

William H. Angoff
1919-1993

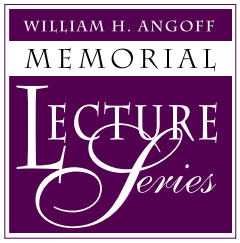


William H. Angoff was a distinguished research scientist at ETS for more than forty years. During that time, he made many major contributions to educational measurement and authored some of the classic publications on psychometrics, including the definitive text "Scales, Norms, and Equivalent Scores," which appeared in Robert L. Thorndike's Educational Measurement. Angoff was noted not only for his commitment to the highest technical standards but also for his rare ability to make complex issues widely accessible.

The Memorial Lecture Series established in his name in 1994 honors his legacy by encouraging and supporting the discussion of public interest issues related to educational measurement. The annual lectures are jointly sponsored by ETS and an endowment fund that was established in Angoff's memory.

The William H. Angoff Lecture Series reports are published by the Policy Information Center, which was established by the ETS Board of Trustees in 1987 and charged with serving as an influential and balanced voice in American education.

Standards, Assessments, and Educational Policy:
In Pursuit of Genuine Accountability



The eighth annual William H. Angoff Memorial Lecture was presented at Educational Testing Service, Princeton, New Jersey, on October 30, 2002. This publication represents a modest revision and update of that lecture.

Linda Darling-Hammond

Stanford University

Educational Testing Service

Policy Evaluation & Research Center

Policy Information Center

Princeton, NJ 08541-0001

Preface

In the eighth annual William H. Angoff Memorial Lecture, Dr. Linda Darling-Hammond discusses the circumstances in which standards and assessments in American education undermine or enhance students' opportunities to learn and teachers' capacities to teach. In particular, she targets the current practice of using tests as substitutes for accountability systems and discusses how good accountability in education is predicated on well-qualified teachers, coherent curriculum, and work aimed at a higher order of thinking and performance.

Dr. Darling-Hammond is the Charles E. Ducommun Professor of Education at Stanford University. Her research, teaching, and policy work has focused on issues of school restructuring, teacher education, and educational equity. From 1994 to 2001, she served as executive director of the National Commission on Teaching and America's Future. This blue ribbon panel's 1996 report, What Matters Most: Teaching for America's Future, was acclaimed as a blueprint for transforming education to guarantee all children access to high quality teaching. The commission's work led to sweeping policy changes affecting teaching and schooling at all levels of government and to ongoing reform in the preparation of teachers.

Prior to her appointment at Stanford, Dr. Darling-Hammond was William F. Russell Professor in the Foundations of Education at Teachers College, Columbia University, where she also was the codirector of the National Center for Restructuring Education, Schools, and Teaching. She is a past president of the American Educational Research Association, a member of the National Board for Professional Teaching Standards, and a member of the National Academy of Education.

Dr. Darling-Hammond has been intimately involved in the development of assessment systems for teachers and has been at the center of the debate about the importance of assuring that teacher development demonstrates skills and the understandings critical to effective teaching. Both a supporter and critic of standardized assessments and at times of ETS, she has always kept at the forefront the impact assessments have on the quality of teaching and learning in our schools. A deep understanding of research, schools, and policy has enabled her to play a leadership role in influencing—not just writing about—educational practice in this country.

It is not an exaggeration to say that virtually every major reform effort in teacher education for the last two decades has been shaped by Dr. Darling-Hammond, and in all of these initiatives, issues of assessment are at the forefront.

The William H. Angoff Memorial Lecture Series was established in 1994 to honor the life and work of Bill Angoff, who died in January 1993. For more than 50 years, Bill made major contributions to educational and psychological measurement and was deservedly recognized by the major societies in the field. In line with Bill's interests, this lecture series is devoted to relatively nontechnical discussions of important public interest issues related to educational measurement.

Ida Lawrence
Senior Vice President
ETS Research & Development
January 2006

Acknowledgments

In addition to the lecturer's scholarship and commitment in the presentation of the annual William H. Angoff Memorial Lecture and the preparation of this publication, ETS Research & Development would like to acknowledge the ETS Policy Information Center for publishing this lecture, Kim Fryer for editorial support, William Monaghan and Madeline Moritz for administrative support and arrangements, Christina Guzikowski for desktop publishing and layout, and most importantly, Eleanor Angoff for her continued support of the lecture series.

Abstract

Standards-based reforms in U.S. education have created demand for increased testing of students and teachers as the basis for a broad range of policy-making decisions. Proponents claim that standards and assessments can enhance learning and render educational systems more accountable for improvements. Opponents claim that inequalities are exacerbated by many current uses of these tools. Unfortunately, such debates often treat both tests and their policy uses as black boxes for improving education. Adding to the controversy are the varying ways that states are using assessments and educational standards in schools. To develop genuine accountability for student learning, the United States needs education policies that use assessments to guide improvements in schools, rather than reduce the amount and quality of education students receive.

Introduction

I am pleased to be here to deliver the Angoff Lecture and to discuss with so many respected colleagues our progress in the quest for educative assessment and equitable education that has been central to much of the work of the Educational Testing Service (ETS). This organization has demonstrated how, when used thoughtfully, assessments can contribute to the service of learning and to the quality and equity of education. For instance, the National Assessment of Educational Progress (NAEP), so ably managed for many years by Archie Lapointe and his team at ETS, has improved the caliber of educational assessments in this country as it has reported both on student achievement and on the conditions of teaching and learning in schools. Pathbreaking teacher-performance assessments developed for the National Board for Professional Teaching Standards by Mari Pearlman and colleagues at ETS have reshaped how the educational community thinks about assessing teachers. The Board's portfolio processes help to improve and measure teaching. And a series of reports over many decades have emanated from this institution examining issues of educational equity and insights for the improvement of teaching (see, for example, Barton, Coley, & Goertz, 1991; Wenglinsky, 2000).

All of these initiatives have shown that educational assessment efforts can contribute to the ongoing process of improving education — a process in which use of standards and high-quality assessments is critical. Building on that framework, I want to talk about the issues associated with creating and using assessments to improve teaching and learning in the context of contemporary accountability policies, given the current conditions of U.S. schools. In an era where assessments are being used more and more to drive policy and school practice, there are possibilities for leveraging productive change and dangers of exacerbating inequalities and reducing educational opportunities for the most vulnerable students. I will examine different approaches to standards-based reforms in states across the nation, describe some of the outcomes of different test uses in these policy systems, and suggest what would be needed to develop a system of genuine accountability on behalf of students and families.

Part 1: The Standards-Based Reform Movement

The standards-based reform movement that dominates today's educational policy scene was launched in the early 1990s. As outlined by Jennifer O'Day and Mike Smith (1993), the fundamental idea of this reform is that if governments can clearly specify what students should know and be able to do, then these standards can shape curriculum, assessment, instruction, teacher education, professional development, the allocation of resources, and all of the other elements of the educational system. Advocates of standards-based reform have argued that bringing this kind of clarity and coherence to our cacophonous, decentralized system could result in a host of improvements, including alignment between student goals, teacher preparation, and other policies; a more focused set of efforts to improve education; stronger incentives for change; and more-targeted applications of resources to needs that are identified by assessments.

Over the past decade, at least 48 states have pursued this general approach — setting standards and developing new assessments for students. In some places, comprehensive reforms such as those anticipated by the initiators of the movement have, in fact, unfolded. But in other places, these elements of standards-based reform have not emerged. Instead, the notions of standards and accountability have often become synonymous with mandates for student testing that are unconnected to policies addressing the quality of teaching, the allocation of resources, or the nature of schooling.

A reliance on testing as *the* reform, rather than a component of a comprehensive improvement agenda, is particularly problematic when these reforms are layered on a starkly unequal educational system. There are wide variations in the availability of school resources to students in different communities, and these are strongly correlated with race and class. Within states, it is not unusual for the top-spending districts to spend three or four times as much as the bottom-spending districts (Darling-Hammond, 1997; Barton et al., 1991). By contrast, most other countries that are peers or competitors to the United States fund their schools centrally and equally: Salaries for teachers are the same across districts, as are expenditures, with the exception of additional resources for those with higher needs. Although U.S. students receive very different educational resources, they are still expected to meet the same standards.

Furthermore, where a comprehensive reform is not pursued, assessments are often unaligned to the standards. And in many states, the standards and curriculum are aimed more at superficial coverage of topics than mastery of content. For example, international assessments such as the Trends in International Mathematics and Science Study (TIMSS) have shown that students in the United States are typically asked to cover many more topics in fields such as mathematics as students in other countries that have higher levels of achievement (Darling-Hammond, 1996; Lapointe, Mead, & Phillips, 1989; McKnight et al., 1987). Education systems in these other countries tend to focus more on big ideas and delve more deeply into them. For example, in Japan's national curriculum, teachers are asked to tackle only four or five major concepts in a school year and to do so with great intensity. The concepts — fundamental notions such as *ratio and proportion* or *estimation* — are the linchpins of mathematical thinking. By contrast, in U.S. classrooms, teachers may be asked to cover 30 chapters of a textbook, spending only a week on any given topic, making solid learning unlikely for many students. Topics such as fractions,

decimals, and long division need to be retaught year after year, because they are never properly taught or deeply learned. Unfortunately, many state and local standards documents reinforce this unproductive approach by listing dozens or even hundreds of topics to be covered, necessarily briefly, during the course of a year.

The design and operation of our schools is another problem. The factory-model approach developed by scientific managers such as Frederick Taylor (1911) and Franklin Bobbitt (1918) in the early part of the 20th century produced schools organized to process large batches of students in assembly-line fashion rather than to ensure that students are well-known by their teachers and treated as serious learners. Urban high schools, for example, typically hold at least 2,000 to 3,000 students, who may see six or eight teachers each year for 45 minutes apiece. In cities such as Los Angeles, teachers daily see 180 to 200 students, who cycle through the classroom to be stamped with a lesson as if they were on a conveyer belt. Teachers are asked to individualize curriculum for the needs of every learner when they have no way of coming to know their students well. Furthermore, the group of learners is much more diverse than at any other time in our history. Thus, the conditions for achieving high standards are lacking in many schools in the United States. These conditions are critical targets for change in a comprehensive accountability system.

Part 2: What is Accountability?

In public conversation, testing and accountability are often conflated or used as synonyms. And that's a problem for test developers, the public, and policy makers, as well as educators and students. Accountability occurs when policies and practices work to provide good education and to correct problems as they occur. Accountable systems increase the probability of high-quality practice, leading to positive outcomes. They reduce the probability of malpractice or educational harm, and they call attention to problems and needs. Furthermore, accountability must be two-way: If students are accountable for learning to certain standards, schools, districts, and states must be accountable for providing them with the necessary resources for learning.

Conceptualized this way, it should be clear that testing does not equal accountability. Test scores are information for an accountability system; they are not the system itself. The information tests provide has been improving as many states have built more productive assessments that use constructed response items, essays, and other performances to evaluate learning; that better measure higher-order thinking and performance skills; and that use criterion-based scoring to assess learning in terms of standards. The ability to continue to develop and preserve the gains that have been made in developing useful assessments depends on helping policymakers understand that tests are only indicators that offer information for accountability systems. Otherwise, tests are asked to take on burdens of decision making and of instructional improvement, which they are not designed to carry and are not capable of accomplishing.

True accountability occurs only when policymakers and educators can act on the information provided by an accountability system in ways that create better opportunities and outcomes for both individual students and groups of students. The consequences of tests depend as much on the system of funding, school organization, and professional development that surrounds them — and on the ways in which they are used — as on the quality of the instruments themselves.

What Matters for Student Learning

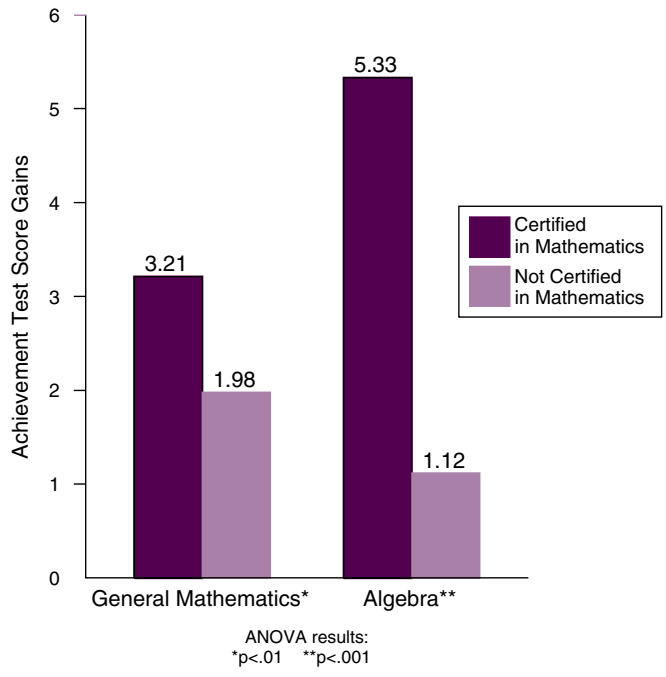
While assessments play a part in improving education, there is evidence about the critical aspects of education that matter most for student learning. Key factors include well-prepared teachers, well-designed and coherent curriculum, skillful instruction that is adapted to students' needs, and personalized learning environments in which students are well-known by their teachers (Darling-Hammond, 1996). Providing these key features of a sound education is a major foundation of an accountability system.

Well-qualified teachers are at the center, as they are the vehicles for developing a sound curriculum, implementing successful pedagogies, and designing more personalized schools. Research has found that effective teachers have strong content knowledge and pedagogical training in the field in which they teach, as well as an understanding of students and how they learn (for a review, see Darling-Hammond, 2000). Harold Wenglinsky (2000) found, for example, that achievement gains in science and math were greater for students whose teachers had a major or a minor in the field taught, who had had teacher education or professional development in teaching for higher-order thinking and for working with diverse learners (including special education students and English language learners), and who used hands-on approaches to learning

(e.g., manipulatives in mathematics and laboratory methods in science).

A number of studies have demonstrated that teacher effects can be quite substantial. Figure 1 provides one example from a matched comparison group study of middle school mathematics teachers with and without certification in mathematics (Hawk, Coble, & Swanson, 1985). Student achievement gains in general mathematics were significantly larger for students taught by certified mathematics teachers — and the gap in learning gains was even larger by the time students reached algebra. As students are expected to reach higher and higher standards, the extent to which they are provided with teachers who can help them do so is a fundamental question about the accountability of states and school systems to the students they serve.

Figure 1. Effects on Student Achievement of Teacher Certification in Mathematics



Note: Data from Hawk, Coble, & Swanson, 1985.

With respect to curriculum, U.S. studies as well as international assessments show that higher levels of student achievement are associated with access to a core curriculum that is coherent, in which subjects are related to one another, and topics build on one another (Darling-Hammond, 1996; McKnight et al., 1987; Lee, Bryk, & Smith, 1995). Without such coherence, students must by themselves make sense of a fragmented patchwork of instructional activities that seem unrelated to each other or to the real world. The curriculum should also provide opportunities for authentic learning, where students actively engage in tasks that ask them to organize knowledge and create products, rather than merely responding passively to questions on worksheets or tests. Research by Fred Newmann, Valerie Lee, and colleagues has found that what they call *authentic instruction* — that is, instruction that supports higher-order thinking and activities that are more like real-world performances — boosts students' standardized test scores as well as other measures of learning (Lee, Smith, & Croninger, 1995; Newmann & Associates, 1996; Newmann, Marks, & Gamoran, 1996).

There is also evidence that students achieve at higher levels and are more attached to school when they learn in settings where they are well-known, where they have longer-term relationships with fewer adults (Lee, Smith, & Croninger, 1995), when they are in smaller schools that range from 300 to 600 students (Green & Stevens, 1988; Howley, 1989), and when they are in smaller classes (Ferguson, 1991; Greenwald, Hedges, & Laine, 1996).

School systems that are accountable to children will ensure that they have well-qualified teachers in adequately resourced schools that are designed to support

teachers in providing good instruction. Assessments of learning and other indicators of school conditions can help evaluate the extent to which educational goals are accomplished. But the focus of accountability must be kept on what is needed actually to improve achievement as well as on how progress is to be measured. As any dairy farmer knows, good milk is not produced just by weighing the cow regularly; one also must feed it and care for it.

The Teaching Gap: How Unequal Opportunities to Learn Compromise Accountability

Despite the critical importance of expert teachers who can skillfully teach a well-crafted curriculum to students whom they know well, many students in the United States do not have access to even minimally qualified teachers. The issue of what states and districts are accountable for was poignantly raised by an episode of *The Merrow Report* titled "Teacher Shortage: False Alarm?" (Merrow, 1999). The video revealed the reality of schooling for a group of students in Oakland, CA., but it could have as easily been about schools in Philadelphia, Los Angeles, Chicago, or New York City, given similar situations in these and other cities. A transcript from part of the episode follows:

Voice-over: But there are classrooms without qualified math and science teachers. School systems say they just cannot find teachers. For example, inside this portable classroom at Bret Harte Middle School in Oakland, CA., is an eighth-grade math class that's been without a regular math teacher for most of the year.

John Merrow: How many math teachers have you had this year?

Boy: Let's see, there is Mr. Berry, Miss Gaines, Mr. Lee, Mr. Dijon, Mr. Franklin…Coach Brown was one of our substitutes one day.

Girl: We had Miss Nakasako, we had Miss Gaines, we had Miss Elmore, we had this other man named…he had like curly hair. His name was Mr. umm…

Merrow: So you've had so many teachers you can't remember all their names?

Girl: Can't remember…yeah.

Voice-over: A few miles away at Oakland High School, this ninth-grade science class has had nothing but substitutes all year long, the entire year without a certified science teacher.

Merrow: What has that been like having, what, 16 teachers or 7 or 9 during the year?

Boy: It's just weird. It's like we have to get used to a new teacher every couple of weeks or so.

Boy: I'm feeling short-handed cause this is the third year. Ever since I got in junior high school, I haven't had a science teacher.

Merrow: So you've had substitutes?

Boy: All three years.

Girl: All we learn is like the same thing all over again. When a new teacher comes, sometimes we've got to skip chapters and start all over again; and it's difficult.

Merrow: Have you learned much science this year?

Boy: Nope.

Boy: Not really. Haven't had the chance to.

Nancy Caruso: It breaks my heart.

Voice-over: Nancy Caruso teaches science at Oakland High School.

Caruso: People have come from those classes over there, and they come down and they beg me, "Can I get into your class. Please, I want to learn. I need a science class." And they're not getting it.

In segregated schools serving African American and Latino students in California school districts such as Oakland, Compton, Pasadena, or Los Angeles, it is not uncommon for most of the teachers to be inexperienced and untrained and for classrooms to be filled by a steady parade of substitutes. This episode of *The Merrow Report* went on to interview several credentialed science teachers who had applied to teach in Oakland and had not gotten a call back from the personnel office. As in some other urban districts serving low-income and minority students, the hiring of unprepared teachers and temporary staff was less a function of shortages than of dysfunctional administration and efforts to save money in a state willing to sacrifice students rather than enforce standards and create incentives for urban teaching.

This teaching gap creates much of the achievement gap. A number of studies in states across the country have shown inequality in students' access to qualified teachers (National Commission on Teaching and America's Future [NCTAF], 1996; Darling-Hammond, 1997). Studies demonstrate that students taught by underqualified teachers have significantly lower achievement on state reading and mathematics tests after controlling for factors such as poverty levels and the language background of students (Betts, Rueben, & Danenberg, 2000; Darling-Hammond, 2000; Fetler, 1999; Goe, 2002; Strauss & Sawyer, 1986). This is made more problematic by the fact that these tests are increasingly used for a variety of purposes, including grade retention and denial of the diploma.

These teacher effects are largest for low-income students who most rely on schools for their education. Consider, for example, that in a selective private school where the students are already high-achieving and well-supported and have parents at home who can help with homework, the fact that a teacher may not have a wide range of teaching skills will be much less problematic than if the teacher were in a setting where many students do not speak English, do not have reading skills developed at home, and have a need for very skillful teaching. Too often, where well-qualified teachers are needed the most, they are the least likely to be found.

Dramatically different access to educational opportunity in the United States is tightly tied to race and class (Ferguson, 1991; NCTAF, 1996; Oakes, 1990; Urban Teacher Collaborative, 2000), and this situation has grown worse over the past 20 years. Since the late 1990s, lawsuits contesting unequal and inadequate resources in low-wealth schools, typically serving low-income and minority students, have been filed in at least 20 states. For obvious reasons, discontent is brewing with the quality of the system. When high school exit examinations were introduced in California, minority communities linked the tests and the high stakes attached to them to the harm that will come to students who, having been deprived of a minimally adequate education, will be additionally deprived of a diploma. A coalition of organizations has worked to delay or derail the exam and has run full-page newspaper ads that read: "Exit exam = racism." Their concerns have merit.

Now, from the perspective of those who create assessments, the examination is not causing the situation in Oakland schools that the video captured. However, there is a need for testing companies and policymakers to acknowledge what is going on in Oakland and other underresourced urban communities across the country, so that tests are not blamed for a policy system that does not provide students with a basic level of

education. There is also a need for those who make and mandate tests to commit to the professional testing standards that argue against the use of tests alone to make important decisions about a student's future (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999, p. 142). While tests used in making high-stakes decisions can exacerbate existing injustices, there are also ways by which standards and assessments can contribute to the correction of these problems, and that is what we need to be working toward.

Standards-Based Reform: Two Approaches

States have approached standards-based reform and the use of assessments differently, with very different results. In some states, tests have been almost the entire reform. Rather than embedding tests as informational tools in a system of broader investments, these states have attached rewards and sanctions for students, teachers, and schools to the test results and to changes in those test results. This approach requires tests, by themselves, to carry the burden for mechanical decisions about promotion and graduation for students, merit pay, and dismissal for teachers and additional funding or reconstitution and other sanctions for schools.

In other places, assessments have been used as part of systemic reform where standards for teachers have been developed in concert with standards for students and then linked to investments in standards-based teacher education and professional development. Standards have been used to guide curriculum reforms and criterion-referenced assessments, and the results of assessments have been used by the system primarily to obtain information to stimulate ongoing improvements in curriculum and professional development, rather than to allocate rewards and sanctions to students, teachers, or schools. Assessment is used as a vehicle for continual improvement of the educational system, instead of a tool for administering threats and punishments to adults and children already deprived of resources for learning.

The difference between these two approaches to the use of assessments is illustrated in an analysis I conducted examining state-level predictors of student achievement on NAEP (Darling-Hammond, 2000). After controlling for poverty and language background of students, I found that most of the differences among states in their reading and mathematics scores between 1990 and 1996 were accounted for by measures of teacher quality, which were, in this analysis like many others, much more influential than measures such as class size. The strongest predictor of student achievement was the state's proportion of well-qualified teachers — teachers who had full certification and a major or minor in their fields. In addition, there appeared to be no relationship between high-stakes testing and student achievement: While most of the states had assessment systems, none of the highest-scoring states during these years had high stakes for students attached to their assessments. Among these high-scoring states, most of those with state testing systems had developed performance-oriented assessments and used them to inform improvements in curriculum and teaching. By contrast, most of the lowest-scoring states had high-stakes testing systems that used primarily multiple-choice tests for many purposes, including grade retention and graduation, merit pay, and school intervention. The two states with the longest-standing and most aggressive high-stakes systems had had no change in student

performance over the years examined. High-achieving states did not appear to need high-stakes tests to enable learning, and low-achieving states did not appear to benefit from having introduced high-stakes tests.

Using Standards and Assessments for Systemic Change: The Case of Connecticut

In this analysis, Connecticut stood out as one of the states that, over the course of the 1990s, had among the steepest improvements in NAEP scores. By the end of the 1990s, it was one of the top-ranked states in the nation in reading, writing, science and mathematics. The National Education Goals Panel commissioned a study of Connecticut to try to understand what it was doing to cause these steep increases, and several other researchers also did follow-up studies (see, for example, Baron, 1999; Wilson, Darling-Hammond, & Berry, 2001).

These researchers concluded, first of all, that the gains in NAEP scores in Connecticut were not a result of changes in the student population. Connecticut is often thought of as a wealthy state, but, in fact, the public school population is not reflective of the wealthy subsection of the state. More than one third of students are African American or Hispanic or have recently immigrated, and a growing number are low-income and language-minority students. Furthermore, over that decade, class size and instructional time did not change.

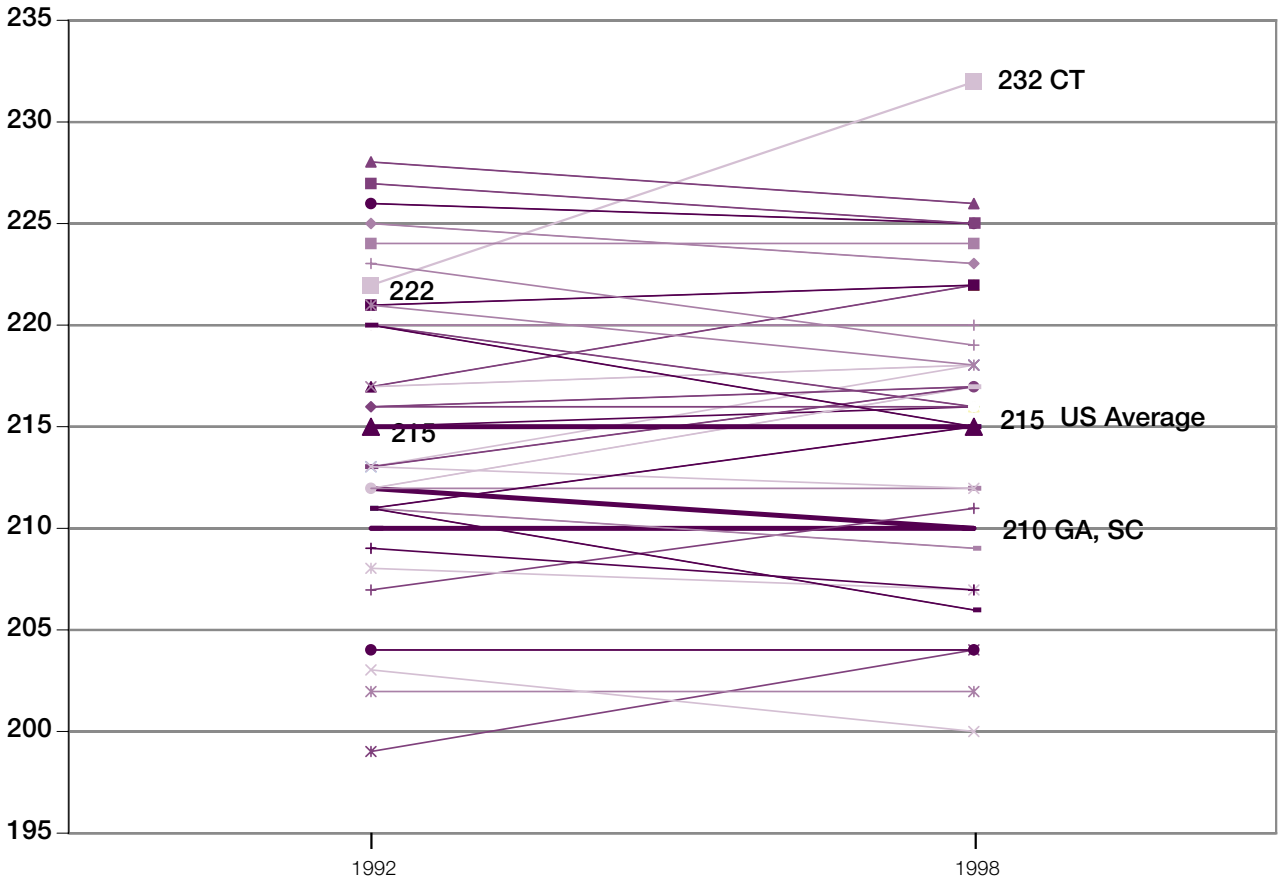
What did make a difference was a purposeful reform of teaching put in place in 1986 under Connecticut's Excellence in Education Act. The act raised teachers' salaries to the highest in the nation and raised standards for teachers at the same time. It eliminated emergency credentialing and created scholarships to entice people to teaching in fields with shortages. It also raised standards for teacher education, requiring all teachers to complete a major in their field and to be extensively prepared in pedagogy, including preparation for teachers of all subjects to teach reading as well as to teach English language learners and special education students. The state put in place a beginning-teacher support program guided by a portfolio assessment modeled on the portfolio assessments of the National Board for Professional Teaching Standards, required mentors for all beginning teachers, and instituted intensive professional teacher development across the state. In the most disadvantaged urban districts, the state provided additional categorical aid to support preschool education, professional development for teachers, and curriculum reforms.

The state also used its performance-based assessment system in very productive ways. By law, the assessments are not allowed to be used for making unilateral decisions about students, including denying diplomas or holding students back. They are used as information to guide curriculum improvement and investments in stronger programs. Like many states, Connecticut has assessments at several grade levels, and it reports the scores both to the public and the schools. But the state department goes further than most states by disaggregating the data in many different ways and working with districts to analyze what different groups of students are learning and how progress is being made over time on particular skills.

Connecticut's approach to standards-based reform provided students with much better prepared teachers and much stronger teaching, and it addressed inequalities, eliminating teacher shortages in the cities within three years. As one example of how dramatic the

changes were, when I lived in New Haven many years ago, I could teach in the city schools without having even a bachelor's degree. A few years ago, my daughter, who also lived in New Haven at that time, wanted to be a teacher's aide there, but all of those positions were filled by credentialed teachers. With a purposeful approach to reform, Connecticut completely changed the labor market for teachers in a short period of time. The state's approach also pushed achievement scores up significantly for both majority and minority students to levels well above those of their counterparts elsewhere. See Figure 2.

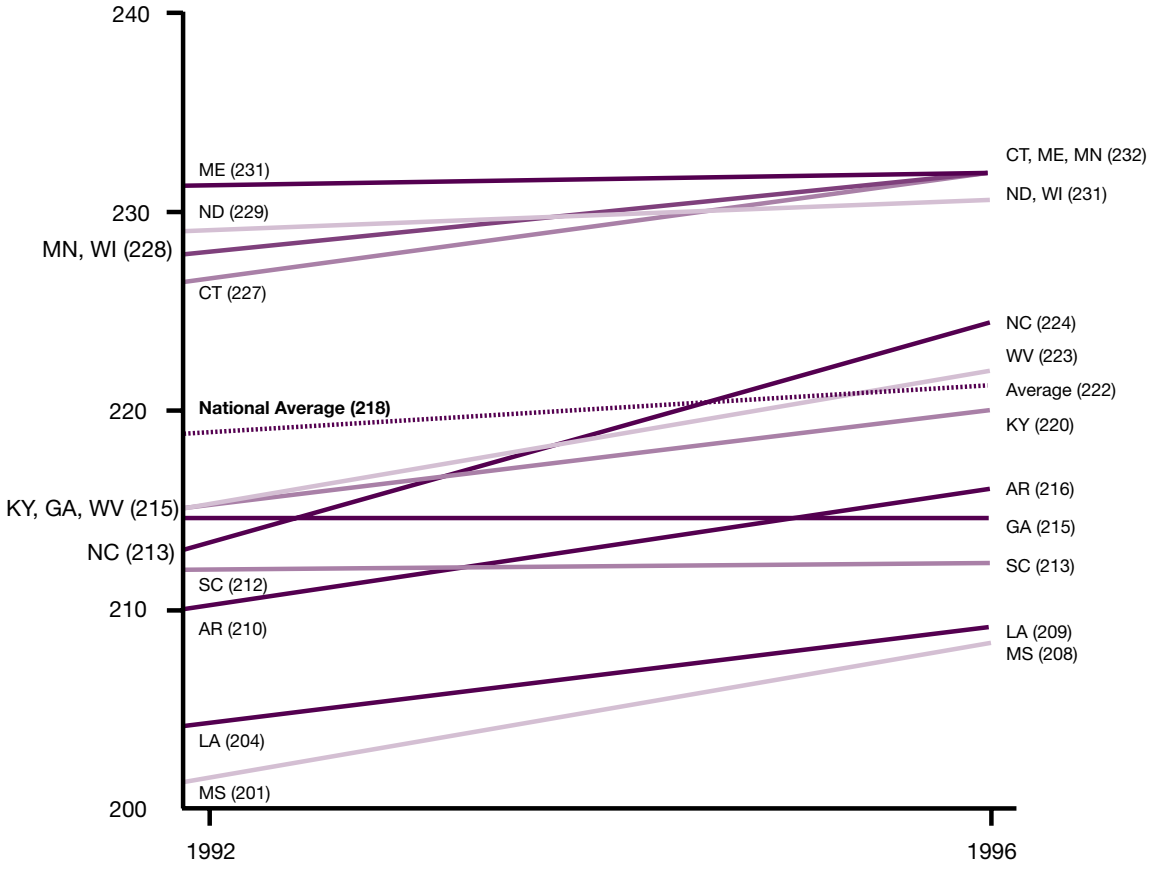
Figure 2. State Trends in Reading Achievement, Fourth Grade



Note. Data from National Center for Education Statistics [NCES] (1999), Table 5.1, p. 113.

In this same study, I described how the policy strategies of several other states that had high performance or strong improvements on the NAEP (for example, Arkansas, Kentucky, Maine, Minnesota, North Carolina, Vermont, and Wisconsin) were in many ways quite similar to those of Connecticut — including substantial investments in teacher quality and professional learning, standards alignment for teachers and students, and equalization of school resources (Darling-Hammond, 2000). See Figure 3. Most of these states also developed systems that used criterion-referenced, standards-based assessments that included performance-based measures, often involving teachers in their development and scoring. As I describe below, the consequential effects of assessments for teaching and learning — and for system improvements — can be positive, especially when performance measures are used, teachers are involved in developing and scoring these assessments, states use the assessments to inform the system, and progress is tracked on criterion-referenced measures.

Figure 3. State Trends in Mathematics Achievement, Fourth Grade



Note. Data from NCES (1997), Table 2.2, p. 28. Adapted from Darling-Hammond, 2000.

The importance of performance assessment. Like Connecticut, a number of states have incorporated extended writing, performance tasks, and classroom work into their assessment systems. These have included portfolio systems in Vermont and Kentucky; performance tasks in states such as Connecticut, Maine, Maryland, and Washington; and classroom-based systems of learning profiles and tests, such as those developed by Minnesota and Nebraska. Studies have found improvements in teaching as well as increases in student performance on higher-order thinking and performance tasks in states that have integrated more performance-based assessments into schools' work (Appalachian Education Laboratory, 1996; Firestone, Mayrowetz, & Fairman, 1998; Herman, Klein, Heath, & Wakai, 1995; Koretz, Mitchell, Barron, & Keith, 1996; Lane, Stone, Parke, Hansen, & Cerrillo, 2000; Murnane & Levy, 1996; Newmann, Marks, & Gamoran, 1995; Stecher, Barron, Kaganoff, & Goodwin, 1998; Stecher, Barron, Chun, & Ross, 2000).

The value of teacher involvement. The use of assessments to improve teachers' understanding of learning and to reflect on curriculum and teaching strategies is also beginning to return in some places in the United States. I say "return" because educational assessment in the United States used to be much more like assessment in Europe, where teachers are routinely involved in developing and scoring assessments. Typically assessments in European countries are essay and oral examinations, and classroom-based assignments such as research papers, problem-solving tasks in mathematics, or experiments in science. Teachers help to develop, administer, and score the assessments within their school and sometimes in other schools. Moderated scoring sessions and auditing procedures help to calibrate how different teachers evaluate the same tasks.

In the United States, teacher involvement in developing and scoring examinations has continued to be a feature of New York's Regents examination system and the Advanced Placement Program® testing system. Vermont and Kentucky introduced performance assessments and portfolios and have also involved teachers in scoring student work. Studies have found that teachers who are involved in scoring performance assessments with other colleagues and discussing their students' work feel the experience helps them change their practice to become more problem-oriented and more diagnostic (for example, Darling-Hammond, Ancess & Falk, 1995; Falk & Ort, 1997; Goldberg & Rosewell, 2000; Murnane & Levy, 1996). When teachers look at student work and evaluate it together, they have the opportunity to think about what counts as good work and how learning is demonstrated. They also have the chance to think together about how they can stimulate that learning in their students. Having the opportunity to share both very concrete analyses of students' work and their thoughts about teaching strategies that can support specific kinds of learning is a powerful form of professional development.

Learning through informative use of assessments. It is also important that systems learn from assessments so that schools improve. Providing assessment data to schools in informative disaggregated forms, as Connecticut does, supports diagnosis of what is working and what is needed and stimulates ongoing reform. Some districts and states are able to analyze the longitudinal gains of individual children who have remained in a school from one year to the next as a true value-added measure of school performance. (Most states currently examine only average changes in group performance from year to year, which makes it impossible to know whether changes in scores reflect

changes in the student population or actual improvement in learning.) This is very important both to give an accurate picture of how the school has influenced learning and to send the right signals regarding incentives, a point I return to later. Test developers typically leave decisions about how data will be aggregated and reported to policymakers. It would be helpful, however, if policy systems could gain expert assistance in aggregating and reporting information about student performance in more useful ways so that it provides a better understanding of school influences on learning and does not create incentives for pushing out low-achieving students.

As part of this mission to improve the uses of assessments and the information they produce, test developers could contribute to the more appropriate uses of tests by educating policymakers about the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999). These standards not only clarify the role of tests in decision making — suggesting that no single test should be used alone as the basis for a high-stakes decision without other indicators of student work — they also outline the conditions needed for appropriate assessment of students, including those with disabilities and English language learners. The standards, unfortunately, are substantially ignored across the country: Very few people in the education and policy-making communities are aware of them, and test makers only infrequently insist upon their use. This inattention contributes to the harm that results from unaccountable systems of education and educational assessment.

The importance of criterion-referenced assessment. Where standards-based reform has been working well to improve school performance, there has been recognition among policymakers of the importance of criterion-referenced assessments, which reveal the status of student learning against a standard, rather than merely ranking students against each other. Much of this awareness has been stimulated by NAEP, which has changed the way people think about assessment and benchmarking levels of performance. Advanced Placement Program® tests, which are becoming more widespread, also set a criterion standard, and they allow states to track whether more students are learning certain kinds of content over time.

But the norm-referenced view of the world — the view that tests should be constructed to array individuals along a normal curve in order to rank them — is deeply rooted in the popular culture. Policies that adopt such tests as the basis for decisions about students and schools create a great deal of mischief. It is very difficult, for example, to explain to some state legislators why all students cannot score above the norm on norm-referenced tests, which are designed, of course, to ensure that 50% of students will score below the norm. In fact, legislation was once introduced in Pennsylvania requiring that all students should score above the norm on a state-adopted norm-referenced test. This idea may seem ludicrous to test makers but it seems like common sense to laypeople, who view scoring at the norm or at grade level as the appropriate goal for all students, and who do not realize that, no matter how well-educated students are, this cannot be the result on a norm-referenced test. It is also difficult to explain to these legislators why there will always be 25% of students scoring in the bottom quartile on such tests and why punishing these students or their schools, as some policies do, will not change that. Even if the policies hold back or push out some students below the bar — or reconstitute schools that fall in the bottom

segment — some other students or schools will have to replace them in a norm-referenced world.

It is critical that test makers help the entire education field and the policy-making community to understand that, to implement standards-based reform, we must use assessment mechanisms that evaluate standards over time and that can reveal progress, rather than use tests that are continually renormed to reproduce the normal curve and that eliminate items when too many students have come to know the answers to them. This is not just a technical issue: It is important for the long-term health and survival of public schools, because there is a growing political movement that argues that public education has failed and must be replaced by private alternatives. If public schools are evaluated against a norm-referenced benchmark, they will never be able to demonstrate the progress they are, in fact, making. An approach to assessment that continually shows 50% of students scoring below the norm feeds the idea that there is no improvement in student learning; that things are not getting better and, perhaps, that they cannot get better. The answer increasingly offered is to shift to for-profit schools, vouchers, and other private alternatives — options that are frequently not subjected to the same accountability mechanisms as public schools. The technical issues of assessment, in fact, have very substantial public policy implications that are infrequently raised and even more rarely understood.

Using Tests as the Reform: Negative Consequences of Inappropriate Assessment

Some of the misunderstandings I have described regarding what tests can accomplish are at the heart of state policies that have been much less helpful to educational improvement and, in some cases, decidedly harmful to students. This has often been the case in states where little investment has been made in improving the quality of education but many uses have been found for using test scores to make decisions about students, teachers and schools. With substantial, continued inequality of resources, critical questions arise about many students' opportunities to learn. In some cases where the tests are quite narrow, the curriculum is becoming even more rote-oriented (Haney, 1999; Herman & Golan, 1993; Jones & Egley, 2004).

Georgia and South Carolina are examples of states that initiated high-stakes testing early in the accountability era without developing a comprehensive systemic reform of education and teaching. In 1984-1985, these two states established the first statewide test-based accountability policies for their educational systems, using the tests to allocate a wide range of rewards and punishments for students, teachers, and schools. These included holding students back in grade and denying diplomas based on test scores, giving teachers merit pay based on rising school scores, and putting schools on probation when scores did not improve. Yet, despite all of these incentives, there was still little overall improvement in student achievement in either state by 1998 as measured by NAEP. See Figures 2 and 3.

Although these states created many kinds of carrots and sticks to try to raise scores, rewards and punishments are not enough to inspire high quality learning if the tests do not require higher-order thinking and performance, if educators are not supported in learning new instructional strategies, and if fundamental resources are lacking. Tests alone do

not improve the quality of student performance when they are unaccompanied by other measures that could actually improve schools.

Indeed, labeling schools as low-achieving can actually make them worse. Although such strategies are intended to stimulate improvement, they can also discourage qualified teachers from teaching in schools that are subject to test-based sanctions or stigma, thus reducing teacher quality. This outcome was reported as an early outcome of Florida's use of average test scores for school rewards and sanctions. Press reports noted that qualified teachers were leaving the schools rated D or F "in droves" (DeVise, 1999), to be replaced by teachers without experience and often without training. As one principal queried, "Is anybody going to want to dedicate their lives to a school that has already been labeled a failure?" A more systematic study of the effects of the North Carolina accountability system found that sanctions negatively affected schools serving low-performing students by impairing the schools' ability to retain teachers (Clotfelter, Ladd, Vigdor, & Diaz, 2004). Thus, rather than improving education for the students who are already most underserved, an accountability system can create a dynamic in which the quality of education is driven even lower.

Another side effect of high-stakes testing in these and other states is higher rates of grade retention and lower rates of graduation. Even though states might have rising test scores on their own tests, this can be a function of either teaching to the test or manipulating the population of students counted in test results, or both. According to studies in many states, increases in test-based grade retention have been associated with increasing dropout and pushout rates (Darling-Hammond & Falk, 1997; Haney, 2000; Orfield & Ashkinaze, 1991). A recent study by the Chicago School Reform Consortium tells a common story (Roderick, Bryk, Jacob, Easton, & Allensworth, 1999; Roderick, Jacob, & Bryk, 2002). Chicago, like several other urban districts and some states, instituted a policy of grade retention for students who did not meet a certain target in third, sixth, and eighth grades. More than 20,000 students were kept back in the first two years of the program, producing many dysfunctional side effects: Many students were retained by mistake because scores were often reported inaccurately; the thousands sent to summer school overwhelmed the city's ability to serve them or keep track of them. There were not enough teachers or seats in summer school for all the children required to attend.

Furthermore, the retained students did not do better than previously socially promoted students. In fact, these students' gains were smaller than the gains for students with similar scores who had been promoted prior to the policy. Also troubling was that one-year dropout rates among eighth graders were higher under this policy. In short, as the evaluators found,

> …Chicago has not solved the problem of poor performance among those who do not meet the minimum test cutoffs and are retained. Both the history of prior attempts to redress poor performance with retention and previous research would clearly have predicted this finding. Few studies of retention have found positive impacts, and most suggest that retained students do no better than socially promoted students…The [Chicago public school] policy now highlights a group of students who are facing significant barriers to learning and are falling farther and farther behind (Roderick et al., 1999, pp. 55-56).

As an accountability policy for improving how

children are taught and what they learn, Chicago's policy was not successful. Chicago was not the only city with this experience. In the 1980s, New York City's Promotional Gates program created performance standards for promotion in the fourth and seventh grades. Promotional Gates had the same results as the policies in Chicago. Thousands of students were retained, often remaining for several years in the same grade, and scores did not go up for the students who were retained. Students who were in underresourced schools continued to receive poor teaching; those with special education needs were still inadequately served; many dropped out in greater numbers, and those that stayed did not improve their achievement. Chancellor Joe Fernandez terminated the program in 1990 as a failed experiment. A decade later, with no institutional memory, another chancellor brought back the same policy that had failed only a few years earlier ("Ending Social Promotion in the New York City Public Schools," 2000).

These experiences are consistent with the findings of studies suggesting that when schools are rewarded or punished for students' average scores, there are substantial incentives for low-scoring students to be retained in grade to make their test scores look better, which increases their odds of dropping out or even of being pushed out so that the school's average scores will increase. Schools that have leeway in admissions may also seek to prevent weaker students from enrolling. Smith and colleagues explained the widespread engineering of student populations that he found in his study of New York City's implementation of test-based accountability as a basis for school-level sanctions:

> [S]tudent selection provides the greatest leverage in the short-term accountability game....The easiest way to improve one's chances of winning is (1) to add some highly likely students and (2) to drop some unlikely students, while simply hanging on to those in the middle. School admissions is a central thread in the accountability fabric (Smith, 1986, pp. 30-31).

Policies that reward or punish schools for their average test scores can create a distorted version of accountability, one in which beating the numbers by manipulating student placements or enrollments overtakes efforts to serve students' educational needs well. Below is an example of how the incentives operate in an accountability policy that rewards schools if their average scores go up from year to year and sanctions them if the scores go down — an approach that many states have adopted, and that is encouraged under the No Child Left Behind Act. In this school, the average test score went up from 70% to 72% between 1998 and 1999, and the proportion of students meeting the standard or target score increased from 66% to 80%. See Figure 4. Under the state policy, this school would get rewarded.

Figure 4. Average State Assessment Scores for a School Over Two Years

Ranking Schools: Rewards or Sanctions?

	1998	1999
Laura	100	90
James	90	80
Felipe	80	70
Kisha	70	65
Jose	60	55
Raul	20	
	Ave. Score = 70 % meeting standard = 66%	Ave. Score = 72 % meeting standard = 80%

However, if you look closely at the data, every student's score actually went down between 1998 and 1999. The average went up because the lowest-scoring

student, Raul, had left the school. Analyzing the same data using a value-added approach that examines individual student scores over time shows that the average for students who were in the school over two years declined by eight points between 1998 and 1999, from 80 to 72. Thus, by allowing Raul to leave, the school went from a threat of sanctions to a promise of rewards.

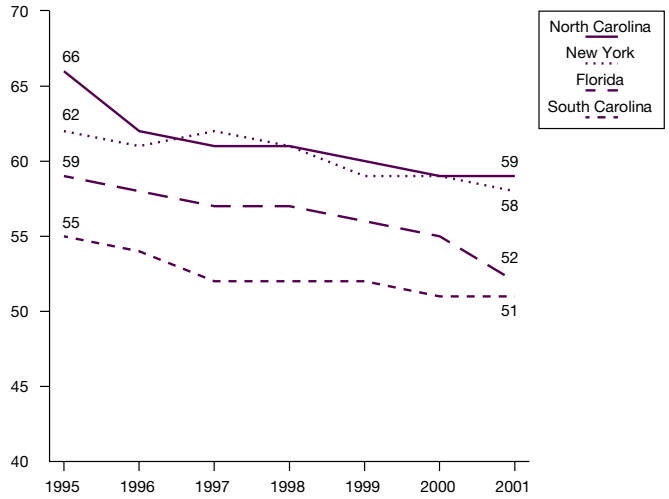
Most states do not use value-added measures that would evaluate over time what individual students have learned as the basis for their accountability systems, and few take school exclusions into account, or even measure dropouts accurately. Nearly all of the states' accountability plans, because they only track average scores, inadvertently create incentives for not admitting or getting rid of students who are problematic or who perform poorly on tests. A number of studies in the United States and abroad have found that testing systems with high-stakes outcomes, particularly when they are tied to school-level sanctions, can result in exclusions of low-achieving students, who drop out or are pushed out of school (Jacob, 2001; Lilliard & DeCicca, 2001; Orfield & Ashkinaze, 1991; Rustique-Forrester, 2005). For example, Schiller and Muller (2000) found that teachers' identifications of at-risk students were more likely to result in students leaving school when schools were subject to test-based sanctions and more likely to result in greater help for students otherwise. And Wheelock (2003) found that high schools receiving state awards for gains in pass rates on the Massachusetts 10th-grade test showed substantial increases in ninth-grade retention rates from the prior year and in the percentage of "missing" 10th-graders. And in this case, as others, the students retained, dropped out, or missed are disproportionately African American or Hispanic.

Attaching rewards and sanctions to scores disaggregated by race, which many have advocated in order to call attention to inequalities, can, paradoxically, lead to greater hardships for students of color. To get African American or Latino students' scores up, for example, schools have an incentive to push out low-achieving members of these groups. Recent evidence suggests that disproportionate numbers of such students are leaving school in high-stakes testing states and that graduation rates for students of color have dropped the most precipitously. In addition, these students are leaving earlier and are less likely to return (Haney, 2000; Orfield, Losen, Wald, & Swanson, 2004; Wheelock, 2003). Data from the National Center for Education Statistics show declining graduation rates in a number of states that used exit exams during the 1990s as the primary measure for graduation. See Figure 5.

In all of these states, the four-year graduation rates dropped below 60%, and the rates for minority students were significantly lower. These are shocking statistics in a world where most European and Asian countries are now routinely graduating more than 90% of their students.

The process of pushout can happen in many ways: Students may become discouraged when they fail tests on multiple occasions or are repeatedly retained in a grade, counselors or administrators may encourage them to transfer to a GED program or continuation school, or zero-tolerance disciplinary procedures, extensive use of suspensions and expulsion, and attendance penalties can be used to remove students from school. And in schools of choice, such as magnet schools or charter schools, there are many requirements that can be used to avoid admitting difficult students as well as to get rid of them.

Figure 5. Declining Graduation Rates in States With Exit Exams



Graduation rates equal the number of graduates in a given class divided by the number of ninth-graders 3½ years earlier. Data are from NCES (2001).

This poses a devilish dilemma for society. Schools that work to hold onto the students who are struggling the most — students who have multiple needs and, often, unstable family situations — know they are likely to depress their average test scores while expending energy on students who may not ultimately succeed at the expense of some who might. The school that lets these students go will be rewarded with better-looking test scores and more resources to spend on easier students. What looks like an accountability success is actually, however, a failure for the society as a whole.

In an economy where the odds of a high school dropout getting a job are less than one in two — and less than one in five if he or she is African American (NCES, 1998, p. 100) — what happens when there are large numbers of young people with only an eighth- or ninth-grade education out on the streets unable to become gainfully employed? The answer is that they tend to go on welfare or into the criminal justice system. For example, prison "enrollments" more than doubled in the 1980s (U.S. Census Bureau, 1996, p. 219). And expenditures on criminal justice went up by over 600% (Miller, 1997), while public education spending went up by about 26% (NCES, 1994). Rather than paying taxes and enriching the society, those who have been failed by schools pose enormous costs to society in both monetary and human terms. Ill-conceived accountability schemes divorced from educational investments in the most vulnerable communities are creating a bigger and bigger cadre of people who are not adequately taught in school, who are in a system that spits them out of school as early as possible, and who have few options and little support for finding a productive future. The "school-to-prison pipeline," as it has been dubbed, is not what we want or need the result of an accountability educational system to be.

Part 3: What Can Be Done?

To develop genuine accountability for students — accountability that enhances their educational opportunities — we need to develop policies that use assessments to guide educational improvements, rather than to further reduce the amount and quality of education students receive. Frustration with the current system is producing test boycotts, lawsuits against testing, and legislative alternatives to testing in a growing number of states. This backlash against testing is both predictable and unfortunate, because, properly used, testing can be a valuable tool. Professional communities of educational research, practice and policy, along with organizations that foster assessment development, need to construct better policy alternatives to accomplish the goals of standards-based reform.

The process of raising standards cannot be separated from issues of teaching, assessment, school organization, professional development, and funding. Efforts aimed at genuine accountability must include changes that address the overall fabric of education. Several principles for the more productive use of assessments are key:

- *Use standards and authentic assessments of student achievement as indicators of progress to improve teaching and provide needed supports, not as arbiters of rewards and sanctions for students and schools.* Using tests alone to make decisions they are not designed to make produces unintended negative consequences for the quality of education students receive, especially those who are least well-served by the current system. Evidence suggests that low-stakes uses of tests for information and improvement are as effective as high-stakes approaches in raising student achievement and improving instructional quality while keeping students in school.
- *Expand performance components that provide "tests worth teaching to" (Resnick, 1987) — assessments that encourage the kinds of higher order thinking and performance skills students will need to use in the world outside of school.* If the goal is stronger education, then investments in more productive assessments are not just testing costs, they are part of the core costs of instruction and professional development. And where teachers are involved in developing and scoring these assessments, their learning is part of the capacity-building that is essential if tests are to improve rather than restrict student learning opportunities.
- *Eliminate artificial testing barriers to students demonstrating what they know to the fullest extent possible.* In addition to providing performance components, some states are eliminating time limits on assessments and expanding alternatives for students with language differences and students with disabilities, so that students can demonstrate what they know in valid and appropriate ways. Test developers should insist on policy systems that take into account these professional testing standards.
- *Develop systems that include multiple measures.* As suggested by the testing standards, some states require that any decision that includes information from tests also includes information from other sources, including coursework and curriculum-embedded performance assessments. Districts can select these other assessments from those already developed or create them locally. These other indicators help to give a full picture of what a student has learned, so that individual tests are not being asked to take on more than they are designed to accomplish.

- *Require and fund diagnostics for students who are not succeeding.* If students are not doing well and not meeting the standards, it makes little sense to hold them back in an already dysfunctional system and give them the least trained and experienced teacher the next year, which is what typically happens to the lowest-achieving students. Students with identified and unidentified learning disabilities, who are the ones most likely to fail to meet the standards, need something more than what is being offered to them in most schools. A diagnostic process is needed to help uncover the problems students are having, so they can be addressed.
- *Create systems that report value-added scores by school and provide data about school conditions.* In addition to examining student progress over time, these systems should include a variety of data about school practices and outcomes — not just test scores, but also promotion and graduation rates, retention, school resources and teacher-quality indexes. When test scores are posted, policy makers and the public should also be informed about the conditions in which students attend school. These data are essential both to properly measure student progress and to prompt states to implement the kind of changes that are needed to help students actually meet the standards.
- *Provide data to schools in a way that they can learn to use it.* Schools are desperate to learn how to use assessment data for more productive curriculum decision making, but they rarely have the resources to analyze data in sophisticated ways. This is an area where schools need supports that can be provided by state agencies and by assessment developers.
- *Use accountability to upgrade teaching and provide the kinds of professional development opportunities, curriculum reforms and resource allocations that standards-based reform anticipates.* When people talk about accountability and then mention only testing, they demonstrate a misunderstanding of what accountability means. Tests do not equal accountability; they are only a useful lever to help achieve accountability in schools by providing information that, if well-understood and acted upon, could create improvements that would make schools more accountable to the children and families they serve.

In *The School and Society*, John Dewey noted that "what the best and wisest parent wants for his own child, that the community must want for all of its children. Any other ideal for our schools is narrow and unlovely. Acted upon, it destroys our democracy" (Dewey, 1968, p. 3). If we can remember that our goal is, ultimately, to provide the best and wisest form of education for all children, we may eventually be able to create genuine accountability.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Appalachia Educational Laboratory. (1996, February). Five years of reform in rural Kentucky. *Notes from the field: Educational reform in rural Kentucky*, (*5*)1.

Baron, J. B. (1999). *Exploring high and improving reading achievement in Connecticut*. Washington, DC: National Educational Goals Panel.

Barton, P., Coley, R., & Goertz, M. (1991). *The state of inequality*. Princeton, NJ: Policy Information Center, ETS.

Betts, J. R., Rueben, K. S., & Danenberg, A. (2000). *Equal resources, equal outcomes? The distribution of school resources and student achievement in California*. San Francisco: Public Policy Institute of California.

Bobbitt, F. (1918). *The curriculum*. Boston: Houghton Mufflin.

Clotfelter, C. T., Ladd, H. F., Vigdor, J. L., & Diaz, R. A. (2004). Do school accountability systems make it more difficult for low performing schools to attract and retain high quality teachers? *Journal of Policy and Management*, *23*(2), 251-272.

Darling-Hammond, L. (1996). *The right to learn*. San Francisco: Jossey-Bass.

Darling-Hammond, L. (1997). *Doing what matters most: Investing in quality teaching*. New York: National Commission on Teaching and America's Future.

Darling-Hammond, L. (2000). Teacher quality and student achievement: A review of state policy evidence. *Education Policy Analysis Archives*, *8*(1). Retrieved August 30, 2005, from http://epaa.asu.edu/epaa/v8n1/

Darling-Hammond, L., Ancess, J., & Falk, B. (1995). *Authentic assessment in action*. New York: Teachers College Press.

DeVise, D. (1999, November 5). A+ plan prompts teacher exodus in Broward County. *Miami Herald*.

Dewey, J. (1968). *The school and society*. Chicago: University of Chicago Press.

Ending social promotion in the New York City public schools. (2000, Winter/Spring). *NYC Schoolwatch*, *4*(2).

Falk, B., & Ort, S. (1997, April). *Sitting down to score: Teacher learning through assessment*. Presentation at the annual meeting of the American Educational Research Association, Chicago, IL.

Fetler, M. (1999). High school staff characteristics and mathematics test results. *Education Policy Analysis Archives*, *7*(9). Retrieved March 1, 2005, from http://epaa.asu.edu/epaa/v7n9.html

Ferguson, R. F. (1991, Summer). Paying for public education: New evidence on how and why money matters. *Harvard Journal on Legislation*, *28*(2), 465-498.

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

Goe, L. (2002, April). Legislating equity: The distribution of emergency permit teachers in California. *Educational Policy Analysis Archives*, *10*(42). Retrieved March 30, 2005, from http://epaa.asu.edu/epaa/v10n42

Goldberg, G. L., & Rosewell, B. S. (2000). From perception to practice: The impact of teachers' scoring experience on the performance based instruction and classroom practice. *Educational Assessment*, *6*, 257-290.

Green, G., & Stevens, W. (1988). What research says about small schools. *Rural Educators*, *10*(1), 9-14.

Greenwald, R., Hedges, L. V., & Laine, R. D. (1996). The effect of school resources on student achievement. *Review of Educational Research*, *66*, 361-396.

Haney, W. (1999). *Supplementary report on Texas Assessment of Academic Skills Exit Test* (TAAS-X). Boston: Center for the Study of Testing, Evaluation, and Educational Policy.

Haney, W. (2000). The myth of the Texas miracle in education. *Education Policy Analysis Archives*, *8*(41). Retrieved August 30, 2005, from http://epaa.asu.edu/epaa/v8n41/

Hawk, P. P., Coble, C. R., & Swanson, M. (1985, May-June). Certification: It does matter. *Journal of Teacher Education*, *36*(3), 13-15.

Herman, J. L., & Golan, S. (1993). Effects of standardized testing on teaching and schools. *Educational Measurement: Issues and Practice*, *12*(4), 20-25, 41-42.

Herman, J. L., Klein, D. C. D., Heath, T. M., & Wakai, S. T. (1994). *A first look: Are claims for alternative assessment holding up?* (CSE Technical Rep. No. 391). Los Angeles: University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing.

Howley, C. B. (1989, Fall). Synthesis of effects of school and district size: What research says about achievement in small schools and school districts. *Journal of Rural and Small Schools*, *4,* 2-12.

Jacob, B. A. (2001). Getting tough? The impact of high school graduation exams. *Education and Evaluation and Policy Analysis*, *23*(2), 99-122.

Jones, B. D., & Egley, R. J. (2004). Voices from the frontlines: Teachers' perceptions of high-stakes testing. *Education Policy Analysis Archives*, *12*(39). Retrieved August 10, 2004, from http://epaa.asu.edu/epaa/v12n39/

Koretz, D., Mitchell, K. J., Barron, S. I., & Keith, S. (1996). *Final report: Perceived effects of the Maryland school performance assessment program* (CSE Technical Rep. No. 409). Los Angeles: University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing.

Lane, S., Stone, C.A., Parke, C. S., Hansen, M. A., & Cerrillo, T. L. (2000, April). *Consequential evidence for MSPAP from the teacher, principal and student perspective*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Lapointe, A., Mead, N. A., & Phillips, G. (1989). *A world of differences: An international assessment of mathematics and science*. Princeton, NJ: ETS.

Lee, V. E., Bryk, A. S., & Smith, J. (1995). The organization of effective secondary schools. In L. Darling-Hammond, J. Ancess, & B. Falk (Eds.), *Authentic assessment in action: Studies of schools and students at work*. New York: Teachers College Press.

Lee, V. E., Smith, J., & Croninger, R. (1995, Fall). Another look at high school restructuring. *Issues in Restructuring Schools* (Issues Rep. No. 9). Madison, WI: Center on Organization and Restructuring Schools.

Lilliard, D., & DeCicca, P. (2001). Higher standards, more dropouts? Evidence within and across time. *Economics of Education Review*, *20*(5), 459-73.

McKnight, C. C., Crosswhite, F. J., Dossey, J. A., Kifer, E., Swafford, J. O., Travers, K. J., et al. (1987). *The underachieving curriculum: Assessing U.S. school mathematics from an international perspective*. Champaign, IL: Stipes Publishing Company.

Merrow, J. (Host). (1999, September). Teacher shortage: False alarm? [Television series episode]. In J. Merrow (Executive producer), *The Merrow report*. New York and Washington, DC: Public Broadcasting Service.

Miller, J. G. (1997, June). African American males in the criminal justice system. *Phi Delta Kappan*, K1-K12.

Murnane, R. J., & Levy, F. (1996). *Teaching the new basic skills: Principles for educating children to thrive in a changing economy*. New York: The Free Press.

National Center for Education Statistics. (1994). *Digest of education statistics*, 1994. Washington, DC: U.S. Department of Education.

National Center for Education Statistics. (1995). *The condition of education, 1995*. Washington, DC: U.S. Department of Education.

National Center for Education Statistics. (1997). *NAEP 1996 mathematics report card for the nation and states*. Washington, DC: U.S. Department of Education.

National Center for Education Statistics. (1998). *The condition of education, 1998*. Washington, DC: U.S. Department of Education.

National Center for Education Statistics. (1999). *The NAEP 1998 reading report card for the nation and the states*. Washington, DC: U.S. Department of Education.

National Center for Education Statistics. (2001). *Common core of data, 1995-2001* [Data file]. Available from http://nces.ed.gov/ccd/

National Center for Education Statistics. (2003). *Digest of education statistics, 2003*. Washington, DC: U.S. Department of Education.

National Commission on Teaching and America's Future. (1996). *What matters most: Teaching for America's future*. New York: Author.

Newmann, F. M., & Associates. (1996). *Authentic achievement: Restructuring schools for intellectual quality*. New York: Jossey-Bass.

Newmann, F. M., Marks, H. M., & Gamoran, A. (1996, August). Authentic pedagogy and student performance. *American Journal of Education*, *104*, 280-312.

Oakes, J. (1990). *Multiplying inequalities: The effects of race, social class, and tracking on opportunities to learn mathematics and science*. Santa Monica, CA: The RAND Corporation.

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

Orfield, G., & Ashkinaze C. (1991). *The closing door: Conservative policy and black opportunity*. Chicago: University of Chicago Press.

Orfield, G., Losen, D., Wald, J., & Swanson, C.B. (2004). *Losing our future: How minority youth are being left behind by the graduation rate crisis*. Retrieved August 30, 2005, from http://www.urban.org/url.cfm?ID=410936

Resnick, L. (1987). *Education and learning to think*. Washington, DC: National Academy Press.

Roderick, M., Bryk, A. S., Jacob, B., Easton, J. Q., & Allensworth, E. (1999). *Ending social promotion in Chicago: Results from the first two years*. Chicago: Consortium on Chicago School Research.

Roderick, M., Jacob, B., & Bryk, A. S. (2002). High stakes testing in Chicago: Effects on achievement in promotional gate grades. *Educational Evaluation and Policy Analysis*, *24*, 333-358.

Rustique-Forrester, E. (2005). Accountability and the pressures to exclude: A cautionary tale from England. *Education Policy Analysis Archives*, *13*(26). Retrieved August 30, 2005, from http://epaa.asu.edu/epaa/v13n26/

Schiller, K., & Muller, C. (2000). External examinations and accountability, educational expectations, and high school graduation. *American Journal of Education*, *108*(2), 73-102.

Smith, F. (1986). *High school admission and the improvement of schooling*. New York: New York City Board of Education.

Stecher, B., Barron, S., Chun, T., & Ross, K. (2000). *The effects of the Washington state education reform on schools and classrooms* (CSE Technical Rep. No. 525). Los Angeles: University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing.

Stecher, B. M., Barron, S., Kaganoff, T., & Goodwin, J. (1998). *The effects of standards-based assessment on classroom practices: Results of the 1996-97 RAND survey*

of Kentucky teachers of mathematics and writing (CSE Technical Rep. No. 482). Los Angeles: University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing.

Strauss, R. P., & Sawyer, E. A. (1986). Some new evidence on teacher and student competencies. *Economics of Education Review*, *5*(1), 41-48.

Taylor, F. (1911). *Principles of scientific management*. New York: Harper Bros.

Urban Teacher Collaborative. (2000). *The urban teacher challenge: Teacher demand and supply in the great city schools*. Washington, DC: Council of Great City Schools.

U.S. Census Bureau. (2004). *Statistical abstract of the United States: 2004-2005* (124th ed.) Washington, DC: Author.

Wenglinsky, H. (2000). *How teaching matters: Bringing the classroom back into discussions of teacher quality* (Policy Information Center Report). Princeton, NJ: ETS.

Wheelock, A. (2003). *School awards programs and accountability in Massachusetts: Misusing MCAS scores to assess school quality*. Retrieved August 30, 2005, from http://www.fairtest.org/arn/Alert%20June02/Alert%20Full%20Report.html

Wilson, S., Darling-Hammond, L., & Berry, B. (2001). *A case of successful teaching policy: Connecticut's long-term efforts to improve teaching and learning*. Seattle: Center for the Study of Teaching and Policy, University of Washington.

Visit us on the Web at www.ets.org/research

Listening.
Learning.
Leading.

1043 • Printed in U.S.A.