Test-Based Accountability: The Promise and the Perils

TOM LOVELESS

Reform movements in American education are based on theories of social change. The standards and accountability movement is based on the theory that a sequence of three activities will improve education: first, defining what students should learn (setting standards); second, testing to see what students have learned (measuring achievement); third, making the results count (holding educators and students accountable). Most analysts date the standards and accountability movement to the early 1990s, when states began establishing standards in academic subjects. States then instituted testing programs and implemented incentives for schools and students based on pupil test scores. The systems are mature enough to have produced some preliminary results.

What is known so far about the effects of accountability systems on student achievement? Do they work? Are there any unintended consequences? In general, evaluations of accountability systems have been quite positive. In raising student achievement, states that have implemented such systems are outperforming states that have not done so. Although the potential for serious unintended consequences cannot be ruled out, the harms documented to date appear temporary and malleable.

Promising results, however, do not guarantee the longevity of an education reform.[1] Various threats to accountability exist, in particular, the political perils that state systems face when policies are implemented. What do these threats portend for the future of test-based accountability in the United States? That question is especially relevant today as the No Child Left Behind Act, the landmark legislation that federalized what had been primarily a state and local

I am grateful to Alice Henriques for research assistance.

policy domain, comes under increasingly heavy fire from critics. "An Education Rebellion Stirring," headlined the *Christian Science Monitor* in February 2004. In March of the same year, an article in the *New York Times* reported that legislation or resolutions criticizing the law had passed at least one house in twelve state legislatures (both chambers in seven states) and described the coalition of anti–No Child Left Behind Act forces in Oklahoma as "liberal Democrats and states' rights Republicans, angry over what they see as a cumbersome federal intrusion on local schools."[2] A reasonable idea of the future of the No Child Left Behind Act can be gleaned from analyzing the politics of accountability in the states.

What Does the Evidence Show?

Because many states did not have systems in place before 2000, studies of accountability are relatively recent. The preliminary results are positive, but researchers are not unanimous. Audrey L. Amrein and David C. Berliner of Arizona State University are the authors of two 2002 studies sharply critical of high-stakes testing.[3] The researchers examine twenty-eight states with a mixture of student and school accountability policies. States are categorized as having high stakes for students if test scores are linked to promotion in grade level, graduation from high school, monetary awards or scholarships, or the freedom to transfer out of poorly performing schools. States are considered as having high stakes for schools if the state offers monetary awards to educators in high-performing schools or can take over low-scoring schools, replacing principals or teachers, based on test scores. One of these studies investigates the effects of high-stakes testing on various indicators of academic achievement. The other looks at the effects of a high school exit exam on dropout rates and enrollments in General Educational Development (GED) programs.

In the first study, Amrein and Berliner find that states showed no clear academic gains after adopting high-stakes testing. The findings from the second study are decidedly negative. In the sixteen states with exit exams, dropout rates increased, graduation rates declined, and GED rates went up relative to national changes on the same measures. Amrein and Berliner conclude that there is no solid evidence that high-stakes tests produce achievement gains and considerable evidence that high-stakes tests produce negative consequences.

The Amrein and Berliner studies drew withering criticism, primarily on methodological grounds. Comparing data from high-stakes states with national averages does not offer the clear contrast that comparing high-stakes states with

non-high-stakes states would provide.[4] The studies also did not control for demographic changes or other factors influencing achievement scores, nor did the researchers run significance tests to measure the probability that the results appeared by chance.

Studies with more sophisticated methods have produced evidence that accountability systems positively affect student achievement. In a 2003 study examining the effects of accountability on state National Assessment of Educational Progress (NAEP) scores, Martin Carnoy and Susanna Loeb rate the strength of each state's system on a five-level scale; both student and school accountability contribute to the rating.[5] Regression analyses controlled for per pupil revenues, student enrollment, and the percentage of African American and Hispanic students in each state. Carnoy and Loeb find that between 1996 and 2000, the stronger the accountability system, the greater the gains states made in raising the percentage of eighth graders functioning at or above the basic level in mathematics. A two-rank increase in the accountability index was associated with about a one-half standard deviation gain, which is statistically significant. The results were significantly positive for black, white, and Hispanic students and held up after controlling for how many students each state excludes from NAEP testing. The exclusion factor is important in addressing the suspicion that some states artificially inflate NAEP scores by overidentifying students in special education or limited-English programs, thereby exempting such students from NAEP testing.

Margaret E. Raymond and Eric A. Hanushek also investigate the effects of accountability systems on NAEP scores. The researchers compute the gains made by a pseudocohort of students, using scores of fourth graders in 1996 and then calculating how much higher eighth graders scored in 2000. They look only at states with school, not student, accountability, and they compare the test scores of accountability states with those of nonaccountability states. Education spending and parental education were statistically controlled. Raymond and Hanushek find that states with accountability systems increased their NAEP scores by 1.6 percent, and scores in states with no accountability system rose 0.7 percent. They also find that states that required only the release of school report cards rose 1.2 percent—statistically indistinguishable from the 1.6 percent gain in accountability states, suggesting that the mere public reporting of school performance may lead to improvement.[6]

John Bishop of Cornell University has examined systems targeting both students and educators. He analyzes the 1996 and 1998 NAEP scores of eighth graders in states with different accountability regimes—for students, meeting basic course requirements, passing minimum competency exams, and passing

curriculum-based external exit exams; and for schools, receiving rewards or sanctions based on test scores. Students in states requiring curriculum-based external exit exams (New York and North Carolina) exhibited the highest levels of achievement, with an advantage of 0.45 grade levels in math and science, followed by states that reward and sanction schools, with gains of 0.20 grade levels. Minimum competency tests had a positive but insignificant effect. Requiring particular courses for high school graduation had no effect.[7]

The most interesting evidence on student accountability comes from a series of studies by Bishop, including an analysis of international evidence. He discovered that students in countries with curriculum-based external exit exams scored higher on international math and science assessments than those in countries with less stringent promotion requirements.[8] The same pattern held true for Canadian provinces. Those employing such tests exhibited higher test scores than the provinces that did not.[9] In the United States, Bishop finds positive achievement effects for the New York Regents program and the Michigan Merit Award Program. Michigan's program offers one-year $2,500 scholarships to students who meet or exceed standards in reading, math, science, and writing. The tests include demanding material, and students who fall short do not face negative consequences. Thus the program is not high stakes for failure, nor is it predicated on students' demonstrating minimum competency in basic subjects, two aspects of accountability systems that have drawn fire from critics.[10] Bishop concludes that systems combining student and school accountability hold great promise for raising academic achievement, especially when performance on end-of-course exams or other curriculum-based tests is the outcome that states measure and reward.

Unintended Consequences

Critics argue that even if test-based accountability does raise achievement, it also creates harmful effects. The most serious indictment concerns persistence in schooling—that is, whether students progress through school with their age mates by completing one grade of education each year or drop out of school after becoming disillusioned by high-stakes tests. States with strong student accountability typically require students to pass a test before being promoted from one grade to the next or being allowed to graduate from high school. The practice of social promotion—passing students on to the next grade based on age rather than academic accomplishments—comes to a halt.

Carnoy and Loeb examine whether strong accountability systems affect either retention rates (that is, students repeating a grade) or dropout rates from the eighth to twelfth grades. In a study of high school exit exams from 1988 to 1992, Brian Jacob finds that the tests had an adverse affect on low-achieving students, increasing the likelihood that they dropped out of school.[11] Bishop finds the same effect but concludes that states could more than offset the increase by building dropout rates into school accountability.[12] Consistent with that idea, Carnoy and Loeb find no evidence that accountability systems in the 1990s led to increased student retention or higher dropout rates.[13]

Research on whether retention is an effective strategy for boosting achievement is generally negative. Students who are retained score lower on achievement tests than similar students who are socially promoted and continue on to succeeding grades.[14] Retained students are also more likely to drop out of school. These findings may not apply to contemporary retentions, as the literature is dominated by studies conducted several decades ago. Many have inadequate statistical controls for selection effects; that is, they do not take into account that students with a greater likelihood of dropping out frequently exhibit characteristics making them candidates for retention. A study in Texas finds that retained students later made significant academic gains, but a critic of the study has observed that the retained students' test scores were initially so low that regression to the mean was the likely cause of the improvement.[15]

Retained students may be treated differently today. In the past, they simply repeated the same grade without any special intervention to fix academic problems. Today's systems, which either require students to attend summer school and Saturday classes or offer other assistance for academic difficulties, may not produce the same negative effects. Two recent studies of the Chicago approach, which has ended social promotion and mandated summer school for failing students, come to opposite conclusions—one finding that the system benefited retained students, the other finding that it did not. Since 1996, Chicago has retained between seven thousand and ten thousand students a year in the third, sixth, and eighth grades. The studies focused on retentions in the third and sixth grades. Jacob and Lars Lefgren find that the gains in test scores of retained third graders exceeded the average annual gain of promoted students by 41 percent in reading and 33 percent in math. The researchers also find that these students benefited from the mandatory summer school attended before retention. Sixth-grade retentions showed no such benefit, and in some of the analyses a slight loss was indicated. Jenny Nagaoka and Melissa Roderick, on the other hand, have found a small positive effect for third graders, which faded two years after

retention, and a negative effect for sixth graders, who scored about 6 percent lower on reading tests two years after being retained.[16]

Why do the studies arrive at different conclusions? The studies use statistical models that differ in several ways, most importantly in how they control for selection effects. In addition, Nagaoka and Roderick analyze gains in reading scores, whereas Jacob and Lefgren examine both math and reading gains. Suffice it to say that contemporary research on retaining students is not as decisively negative as earlier studies, but studies with positive findings remain in the minority, and a negative impact of retention cannot be ruled out.

A related issue involves schools' practice of removing low achievers from the pool of tested students, either by retaining them in a lower grade, in order to inflate test scores in a later grade that is monitored by the state, or by placing students in special education or limited-English programs, to exempt them from testing. These are efforts to circumvent accountability systems. Walter Haney has charged that after Texas implemented a high school exit exam in 1991, schools began retaining low-achieving students in the ninth grade to keep them out of the tenth grade, the first year students take the exam and results are publicly reported.[17] Linda McNeil also uncovered evidence of this in case studies of schools in Texas, which, being the home of so many education officials in the Bush administration, is a state that has received intense scrutiny.[18] Scant empirical evidence has been produced to document the problem nationally. The difficulty in researching these phenomena is that they predate accountability systems and—certainly in the case of placement in special education and limited-English programs—may occur both legitimately and illegitimately. Changes in the identification rates of special programs can be driven by many factors that are difficult to collect data on or to pin down with school records. After examining these issues, a research team from the Rand Corporation has concluded that "the extent to which these negative effects occurred and the factors that may influence their occurrence remain uncertain, but there is a clear need for further monitoring of these effects and research on them."[19]

A more public flaw of accountability systems is the misidentification of failing schools and students, in which case states invoke unwarranted sanctions. Incentives that are applied haphazardly cannot be expected to work properly. Several embarrassing mistakes have been reported in the press. Officials in New York City incorrectly required nine thousand students to attend summer school in 1999 before discovering that the students' test score data were flawed.[20] In June 2000 Minnesota denied diplomas to fifty-four high school seniors whose high school exit exams had been incorrectly scored.[21]

Maryland postponed release of its 2001 test results when some schools registered implausibly large swings in test scores.[22] These mistakes are rare and can be attributed to human error, but with a student's future or a school's reputation on the line, they undermine the public's perception of accountability systems as fair and accurate.

Thomas J. Kane and Douglas O. Staiger pinpoint volatility in test scores as a technical flaw inherent in test-based accountability systems.[23] School test scores, like any other measurement, will always exhibit a degree of naturally occurring error. Scores may fluctuate even though a school's underlying, real level of achievement remains stable. Kane and Staiger analyze data from California and North Carolina, both of which reward schools for a single year's improvement in scores, to illustrate the arbitrary nature of incentives that ignore this basic statistical property. Small schools, by having fewer students, exhibit more volatility in test scores than do large schools. In statistical parlance, they are more likely to appear in both tails of a state's distribution of school test scores. In comparison with large schools with identical levels of achievement, then, small schools are more likely to be singled out as having significantly improved, and thereby deserving rewards, or having significantly declined, thereby deserving penalties. Kane and Staiger recommend that states either apply statistical filters to smooth out the volatility of small schools' test data or base rewards and sanctions on several years' scores.[24]

Critics assert that accountability systems serve to narrow the curriculum to topics that are tested, a charge based primarily on case study evidence. Few studies have surveyed large numbers of schools on how accountability shapes the school curriculum. Fewer yet have attempted to untangle the effects of testing in general from the effects of tests linked to accountability. It is known that teachers shift instruction toward topics that appear on standardized tests, but this effect was present in the 1980s, when test results triggered few consequences for teachers. Several studies from Rand have examined classroom practices in the 1990s as states adopted test-based accountability schemes. Brian Stecher and Sheila Barron, for example, compare the amount of time fourth- and fifth-grade teachers in Kentucky devoted to different subjects. Fourth-grade teachers spent about four more hours a week on reading, writing, and science, subjects tested in the fourth grade. Fifth-grade teachers, on the other hand, spent almost six more hours a week on subjects tested in the fifth grade.[25]

A 2004 study in Florida finds that schools that had been labeled as failing because of low Florida Comprehensive Assessment Test (FCAT) scores spent an inordinate amount of time teaching writing, the section of the FCAT that educators believe is most amenable to rapid improvement. Data support the

educators' belief. Florida schools receive an F grade if they fail three subjects—reading, writing, and math—and are removed from the list of F schools if they raise any one of the subjects to a passing level. Receipt of two consecutive Fs qualifies a school's students for vouchers to move to other schools. Based on the 1999 FCAT scores, seventy-eight schools went into the 2000 testing period with their first F. All seventy-eight improved to passing on the writing portion and avoided voucher eligibility (57 percent improved in more than one subject). Four schools received an F in 2000, and all four improved enough in writing to receive a passing grade in 2001.[26]

Although research suggests that narrowing of the curriculum does take place in response to accountability, the crucial question, of course, is whether such narrowing is good or bad. One person's "narrowing of the curriculum" is another person's "focusing on what is important." An emphasis on writing may be a sound educational strategy for Florida if it is the subject on which that state's students need the most help. Moreover, long-standing philosophical disputes concerning what schools should teach are often reflected in analysts' judgment of whether concentrating more on some subjects at the expense of others is educationally sound. Educational traditionalists may applaud paring the curriculum back to an emphasis on basic skills. Ideology plays a significant role in the politics of accountability.

In sum, the research on the effect of state accountability systems on student achievement is generally positive. The evidence on unintended consequences is mixed. The literature is negative on achievement effects of retaining students, but recent evaluations of contemporary retentions are less critical of the practice, and at least two studies have found retained students benefiting academically. Some studies of high school exit exams in the early 1990s find that these high-stakes tests elevated dropout rates; others find no evidence of such an effect.

If states discover that high-stakes testing exacerbates dropout rates, incentives for schools to address the problem can be built into accountability systems, as Bishop recommends. Other unintended consequences appear fixable, as well. Both students and schools potentially can be mislabeled as failing when in fact they are not, but establishing tighter quality controls to catch data-reporting errors and using statistical techniques to reduce score volatility would help solve the problem. Test-based accountability appears to narrow the school curriculum. Teachers spend more time on subjects and topics that are tested. Those who believe the emphasized content constitutes valuable learning will view the change positively; those who do not will see the narrowing as harmful.

What Now for Research?

Studies that offer the best guidance on accountability policies are relatively recent. More research is needed, as longer-term effects may prove to be different. Accountability would not be the first educational reform to flash early promising signs only to see gains fade. In addition, the studies of national achievement analyze math scores and are focused on the fourth and eighth grades (grades with state NAEP scores). If educators know more about how to improve learning in math than in other subjects—or with elementary and middle school students than with high school students—the same incentives may not produce the same gains in other subjects or with other age groups. On both the long-term-trend NAEP, given since 1969, and the main NAEP, given since 1990, younger students have made significantly greater gains than students in high school. Effective ways of boosting high school achievement have proved elusive. Mathematics scores have risen significantly, but reading scores have not. Whether the beneficial effects of accountability systems extend to subjects other than math or to students beyond the eighth grade is currently unknown.

With the exception of John Bishop, researchers also have not yet teased apart the impact of different types of accountability.[27] The No Child Left Behind Act stresses school accountability and is silent on student accountability, so the debate, understandably, has shifted toward schools. Bishop estimates that student accountability produces gains (0.45 grade level) more than twice as large as school accountability (0.20 grade level), and Raymond and Hanushek find that formal systems of rewards and sanctions for schools have a positive effect that is statistically indistinguishable from a system that simply publicizes school test scores. Considered together, these two findings raise questions about designing efficient accountability systems. Using tests to assess the mastery of content in individual courses, as curriculum-based external exit exams do, rather than giving tests that measure everything students have learned up to a particular grade, as most state exit exams currently do, may send stronger signals to students on the material that they must learn. If publicizing test scores is enough to motivate schools, formal systems of rewards and sanctions may be unnecessary. Along with saving money from reduced administrative costs, the endless political bickering that rewards and sanctions instigate—on how to identify successful and failing schools, the fairness of various rewards and penalties, the appropriate interventions to help failing schools, and so on—might be avoided. Research has only scratched the surface on assessing the independent effects of student and school accountability, and within those two categories, little work has been done to weigh the

costs and benefits of different strategies or find the most efficient mix of policies to spur better schooling.

Research establishing that accountability systems boost student learning should be followed by research that explains how the systems work. That is, the causal mechanisms linking policy to achievement need to be specified. Some might argue that, in the face of an ambiguous production function, it makes more sense to push authority over processes down to the classroom and then provide incentives for attaining the desired results than to try to pinpoint accountability's causality. This issue goes to the heart of the theory underlying accountability systems and its interaction with political realities. If policymakers rely on incentives for improving either a school or a student, then the question arises, incentives to do what? What exactly should educators in failing schools do tomorrow—that they do not do today—to produce more learning? What should a failing student do tomorrow that he or she is not doing today? For both parties, perhaps it is as simple as trying harder, a behavioral change ripe for incentives to influence. If the solution is not that simple, however, trying harder will lead to marginal gains. Greater gains will materialize only for those who know what to do. There will be students and teachers who try hard and fail—and they will be penalized for their failures. The specter of that entails political risks.

A peculiar phenomenon is unfolding. Evidence of accountability's positive effect on achievement is emerging at the same time that political opposition to accountability systems is picking up steam and becoming increasingly vehement. The primary dissenters are teachers (along with principals and other career educators), in the case of school accountability, and students and their parents, in the case of student accountability. Obviously, some opposition should be expected; no one likes to be held accountable. But defenders of accountability would be foolish to dismiss the opposition cavalierly. Opponents of accountability have had an impact. Several states have modified accountability systems to meet the demands of protesters. The potential for the opponents of accountability to grow in political power is very real and is enhanced by factors that are endemic to schooling.

Political Opposition to Accountability

The greatest threats to accountability's future spring from political aspects of education that are deeply ingrained in American schools. The two summaries of press accounts that follow—one of teacher and student protests

against test-based accountability, the other of states modifying their accountability systems—offer a glimpse into how the political opposition to accountability has played out in different states. The subsequent analysis focuses on structural factors supporting the opposition to accountability.

Teacher and Student Protests

Over the past five years, teachers and students have demostrated their opposition to test-based accountability in several states.

California. In 2002 approximately fifty thousand California students opted out of testing.[28] Two years earlier, a group of teachers who were given rewards for producing high test results turned over the money to nearby schools, where they thought it could be put to the best use.[29]

In May 2001 *Education Week* reported, "Two high schools in California's Tamalpais Union High School District, in wealthy Marin County just north of San Francisco, will be ineligible for state rewards this year because of the high proportions of students who refused to take the Stanford-9. . . . About 600 of the district's 2,700 students got permission from their parents to opt out of taking the test earlier this month, as they are allowed to under state law."[30] In 2002 *Education Week* reported that "Stacey Miller was put on leave with pay . . . from J. W. Fair Junior High School in the heart of Silicon Valley while the district conduct[ed] its probe into allegations that she [had been] encouraging her students to opt out of the state's testing program."[31]

That fall, "citing frustration with the educational system and administration, David Munoz quit his job as an English teacher . . . and incited about 150 students to walk out in protest of state-mandated tests, which Munoz called a hindrance to learning. On Thursday, the 23-year-old again led a crowd of [about 200] students to protest benchmark tests—practice tests that prepare students for the real thing. Munoz said the optional tests are filled with grammatical errors and unclear questions. He said they waste instructional time and should be thrown out or minimized from seven tests to just one."[32]

Georgia. In 2000 "James Hope, a fourth-grade teacher at Centerville Elementary School, made the Gateway exam questions public . . . as a protest against the test."[33]

Illinois. Writing in *Phi Delta Kappan*, Gerald Bracey reports, "In 1999 the Chicago Public Schools (CPS) suspended then fired teacher George Schmidt. The district also sued him for $1 million, upped that to $1.3 million and later to $1.4 million. Schmidt had been so appalled at the quality of CPS tests that he had published them in *Substance,* the monthly newspaper that he puts out

for Chicago teachers. The CPS suit claimed that it needed $1.4 million to replace the published items that were now useless. In 2003 CPS dropped the amount to $500, and this is to be paid only if Schmidt loses all of his First Amendment appeals, which [were] still in progress. When a group dubbed the Curie Twelve—a dozen teachers at Chicago's Marie Curie High School—called the tests "flawed and invalid" and announced that they would boycott the tests this year, CPS dropped the tests altogether.[34]

Massachusetts. In 2000 twenty-six students at Arlington High School were suspended after boycotting the Massachusetts Comprehensive Assessment System exam. More than 100 students at Cambridge Rindge and Latin School also boycotted the test.[35] The following spring, *Education Week* reported that "only about 20 students in Massachusetts, primarily 8th graders, sat out the first round of testing last month, according to state education department officials. Last year, hundreds of high school students boycotted the state's assessment test in protest of what they viewed as a wave of testing, but this year, 10th graders must pass the math and English portions of the test in order to graduate."[36] In late 2000, "Massachusetts' largest teachers' union, in a highly unusual move, . . . launched a biting, $600,000 advertising campaign that attack[ed] the state's high-stakes accountability tests. . . . The Massachusetts Teachers Association began running a television ad Nov. 8 that show[ed] a clock ticking as nervous students struggle[d] through a MCAS exam, which three years from now students must pass in English and mathematics to graduate."[37]

Minnesota. "About half the junior class of Southwest High School in Minneapolis didn't show up for a state math test Wednesday, marking the first significant boycott in the five years of statewide testing."[38]

New York. In Scarsdale, New York, in March of 2001, more than 60 percent (195 out of 290) of eighth graders stayed home on test day. The test in question was the state science exam. In addition, less than half of the students took the English exam. However, the parents who led the boycott say that the math and language arts exams are all right. The boycott was not repeated after the state scolded and warned the district that disciplinary action would be taken in the event of another test boycott.[39]

Newsday reported in March 2002 that "about 120 eighth-graders from two Manhattan middle schools boycotted the first day of the state standardized tests yesterday."[40] The schools were two alternative schools. There also were protests in Westchester County but no boycotts. A few months later, a parent group directed a protest of its own: "'It's a shame that our tax dollars are going to Kaplan [a test preparation company],' said Jane Hirschmann, chairwoman

of New York Parents' Coalition to End High-Stakes Testing, which led a boycott of the tests this week at six city middle schools."[41]

Ohio. In 2003 "Cleveland teachers said they would refuse to give the assessment [K–3 diagnostic tests]." As a consequence, the state postponed the tests until teachers were given further training in how to administer them. Apparently, the refusal was supported or organized (or both) by the Ohio Federation of Teachers.[42]

Washington. "About 1,000 10th graders balked at the WASL [Washington Assessment of Student Learning] tests last year [2001]. That's more than twice the number as in 2000, but it's still only slightly more than 1 percent of the roughly 80,000 sophomores statewide." In elementary and middle school, less than 0.4 percent of students did not take the test in 2001. "In Longview, R. A. Long High School principal Rollie Johnson says he's not seeing the same refusal rate he did a year ago—when 70 sophomores balked—because he's now reaching out to families about the importance of the test."[43]

The States Respond

States have responded to complaints about accountability by modifying accountability systems, at times backsliding and at others navigating midcourse corrections. In the case of student accountability, three common responses are to make allowances for students with disabilities, to delay the effective date of exit exams, and to provide alternative ways for students to pass.

Alabama: allowance for disabilities. "Under the rule change, a student with dyslexia, for example, who has not been able to pass the reading section of the exam because of that disability but who has met other requirements for a diploma will now qualify to graduate."[44]

Alaska: delay. "In 2002 the [state] Legislature delayed the exam's effective date to 2004." Exemptions were granted to new Alaska residents, any student whose parent died in the last semester of senior year, and those who suffered a serious illness or injury.[45]

Arizona: delay. "Arizona launched AIMS [Arizona's Instrument to Measure Standards] in 2000 but [has] twice postpone[d] the graduation requirement after debates about low scores and initial problems with content and scoring."[46]

California: delay. In the summer of 2003, the high-stakes exit exam was delayed from the class of 2004 to the class of 2006 when studies showed that 20 percent of students would fail the math section and therefore not graduate.[47]

Delaware: delay. On April 6, 2004, the state House of Representatives "approved a bill delaying the implementation of a three-tiered high school diploma system for two years and passed a resolution establishing a task force to study the proposed diploma system and the state's student testing program." The original plan would tie diploma level (distinguished, standard, and basic) to state test results.[48]

Florida: alternative passing. In 2003 the state legislature allowed students to replace failing FCAT scores with other test scores (ACT [American College Test] or SAT [Scholastic Assessment Test]). Currently, a bill is before both houses of the legislature to apply the same substitution this year.[49] "About 12,000 seniors failed the FCAT in 2003. 'If a student failed to get a score of 284 on the language portion of the FCAT, a 14 on the ACT or 370 on the SAT would suffice. If the required 295 math score on the FCAT is not achieved, a 14 on the ACT or a 350 on the SAT would be satisfactory for graduation,' Education Department officials said."[50] The bill (HB185) also would allow a student to take the math portion of the FCAT in his or her native language.

Massachusetts: allowance for disabilities. The state requires testing but recently passed the first change, to allow special-needs students who have fulfilled their district requirements the right to file an appeal to demonstrate that their class work is the equivalent of a passing score. Most appeals are being accepted.[51]

New Jersey: alternative assessment. In 2003 nearly one-quarter of New Jersey graduating seniors had either failed the exit exam or were exempted from having the scores count. They all received diplomas. Students were then given the alternate assessment. When the exit exam was launched in 1995, officials said no one failing the test would graduate. Sixteen percent took the alternate test in 2003, up from 9.5 percent in 2002. In addition, more than 7 percent of students, all in special education, were exempt from the exit exam.[52]

New York: delay. In the fall of 2003, the state delayed raising the threshold passing score for the exit exams for at least two years.[53]

North Carolina: delay. "A new 11th grade exit exam was supposed to begin with the class of 2003. But development of that exam was suspended two years ago [in 2002] after resistance from educators and an outright block by the state legislature."[54]

Tennessee: delay. The class of 2005 is supposed to be the first class to face new exit exams, but the state may be backing away from upholding the standards. Concern is growing about what to do with those who do not pass the tests and the tests' impact on limited-English, special needs, and disadvantaged students. A committee from the state board of education has suggested

a matrix that would outline alternative criteria according to which a student would be allowed to graduate.[55]

Washington: alternative assessment. The high-stakes test will first affect the class of 2008. In 2004 the state legislature added a provision that students be allowed up to four retries to pass the test. The final bill also has provisions for alternative assessment, such as an assessment of classroom work.[56]

Factors Supporting Opposition Politics

The political structure of education works against the sustainability of accountability systems. Ideologies play a central role by creating core constituencies for dissent. The contrast between progressive and traditionalist educational philosophies echoes throughout the accountability debate, especially in defining academic standards. Sides are chosen early. Traditionalists tend to support measurable learning standards, describing in clear language the knowledge and skills that students will learn. They support basic skills in mathematics over National Council of Teachers of Mathematics–oriented math, phonics over whole language in reading, the study of history over studies in the social sciences, and the mastery of scientific knowledge over process-oriented instruction or learning how to "think like a scientist." They are not offended by standardized tests with multiple-choice items and are suspicious of portfolio-based assessment systems, test items that allow for subjective scoring and partial credit for wrong answers, and other alternative approaches to assessment.

Progressives, on the other hand, view educational curriculum more holistically, valuing the acquisition of inquiry and problem-solving skills as much as factual knowledge. They believe that the primary-grade mathematics curriculum should extend beyond arithmetic, with first graders and even kindergartners using calculators as they explore concepts in algebra, geometry, and statistics. Whole-language approaches to reading instruction, although no longer favored by federal and state reading policies, remain popular with progressives, who support student-centered learning—instruction that students themselves have had a hand in selecting—and are dubious of learning mandated by external authorities. Many progressives favor "real world" learning—that is, experiential as opposed to book learning—and "authentic assessment" as opposed to standardized tests with multiple-choice items.[57]

Most state standards strike a compromise between traditionalist and progressive notions of curriculum, including elements to satisfy both sides. In

states in which clear winners and losers have emerged—California, Massachusetts, and Virginia being the most famous examples—the losers quickly organized political opposition. Such opposition has proved potent, as all three states threw out progressive curricular standards and adopted more traditionalist ones. Now those policies are opposed by organizations with a progressive bent.[58] As the spotlight has shifted from developing standards to implementing testing regimes, progressives' complaints about accountability have escalated, portraying test-based accountability as a conservative, regressive policy response to education's problems. Alfie Kohn, in the April 2004 issue of *Phi Delta Kappan*, asserts that "the devastation is already underway" and charges that the ultimate objective of accountability advocates is to privatize schooling.[59]

Frederick M. Hess has pointed out that policies governing collective goods are vulnerable to attack even if they enjoy broad public support.[60] Most of the public supports standards and accountability, but not passionately so. When policies governing collective goods exact costs from particular groups, the aggrieved parties have a strong incentive to organize against the policies. The institutions governing education grant highly committed factions a power exceeding their numbers, education's loosely coupled organizational structure offering multiple entry points to derail policy adoption or implementation (for example, state legislatures, local school boards, the superintendent's office, the classroom). Progressives and traditionalists supply ideas that transform personal complaints into philosophical positions. Progressives who oppose test-based accountability appeal to wary students, parents, and educators with the argument that accountability not only threatens them personally but also promotes the wrong kind of education. A national group, Students Against Testing, describes itself as "a nationwide network of young people who resist high-stakes standardized testing and support real-life learning."[61] These are not students who simply dislike tests; they are students with a philosophy of education.

The fragmented structure of educational governance also works against accountability. Authority over schools is shared by federal, state, and local officials. The arrangement is not hierarchical in the sense of greater powers lodged at higher levels of governance. Rather, the distribution of power follows constitutional language granting states primary authority in educational matters, creating a patchwork of programs and practices that conform to individual state political cultures and local traditions that have evolved largely by historical accident. The federal government arguably has the least power. Consider the No Child Left Behind Act. Federal enforcement of the act is limited to

withholding Title I program monies from noncooperative states. The Title I program, which targets poor children, did not exist before its founding in 1965 as part of the Elementary and Secondary Education Act. Total federal funding of K–12 education constitutes only 7 percent of national revenues, and thus the leverage that is generated from threatening to cut off a single program's funds is severely limited. What leverage there is comes from state officials who view federal education money as an entitlement.

In the view of state officials, any threat of coercive action by the federal government unjustly tramples on federalism. Ironically, at the same time state officials invoke federalism in complaining about the No Child Left Behind Act, many of them have been charged by school boards and district educators in their own states with using accountability to usurp local control. Accountability exacerbates intergovernmental tensions. By its very nature, it requires a superior authority to call a subordinate to task if a prescribed outcome is not accomplished. The central lesson of intergovernmental relations in education is that most governmental officials would rather be the regulator than the regulated, but being the regulator does not mean wielding decisive power.

States have constitutional authority over schools, but their powers are constrained. Terry Moe has written pessimistically about the future of accountability, describing both a control problem and a political problem. The control problem is that accountability encounters all of the classic hurdles of principal-agent relationships in large organizations.[62] States, acting as principals, want students to learn certain skills and knowledge. They build elaborate bureaucracies that at the bottom feature teachers, acting as agents. Agents actually produce an organization's outputs. Principals are responsible for an organization's overall production but have difficulty controlling agents. States do not hire teachers, nor can they closely monitor teachers' work. Insisting that certain things are taught and gathering evidence to confirm that learning has actually taken place is fraught with difficulty. Agents have critical information—knowing, for example, what really is going on in classrooms—that principals do not have. As Moe explains,

> These information asymmetries put the principal at a disadvantage. It is not just that he does not know certain things about the agent's type or behavior, which is bad enough. It is also that the agent does know these things and can use this private information to his advantage—allowing him (if he wants) to slack off in pursuit of his principal's goals and substitute his own interests in the performance of his job, all the while giving the appearance of being a good agent.[63]

That agents have their own interests and goals leads to a political problem, which, according to Moe, coalesces in teachers unions. As organizations

representing teachers' political interests, unions are charged to protect teachers from unfair or onerous work demands. Agents' interests may diverge from those of principals. Thus unions seek to weaken accountability systems not because they are selfish or evil but simply because they represent agents—and agents do not want to be held accountable. Teachers unions use their power in the political arena to influence policies dealing with accountability. The result, Moe concludes, are accountability systems in name only:

> As a result, a system of accountability may look like an exercise in top-down control, but it is really a system that has been shaped, perhaps profoundly so, by the self-interest of the very people it is supposed to be controlling. There is every reason to believe, therefore, that it will do a poor job of achieving genuine accountability. Indeed, to the extent that agents can succeed in wielding their political power, it will be *designed* to do a poor job.[64]

Moe's analysis is illuminating, but it remains to be seen whether its pessimism is borne out empirically. Indeed, the national unions oppose test-based accountability systems in their current form. The National Education Association's position is that "using high-stakes tests as a measure of 'accountability' is a fundamentally flawed concept."[65] The American Federation of Teachers (AFT) took a more moderate position in a 2002 resolution: "[Although] the AFT strongly supports standards-based reform, including appropriate testing, it is especially outrageous that this critical reform is threatened by uninformed implementation. The public and teachers are understandably deeply troubled that standardized tests are all too often being used inappropriately, are usurping too much instructional time, and are crowding out recognition of other important subject areas."[66]

At the state level, it is unclear whether the power of unions is negatively correlated with accountability systems. The experience of southern states supports the argument; that is, they tend to have weak teachers unions and strong accountability systems. But the southern states are unique on a range of topics in educational governance (for example, dominance of state funding, large parish- or county-based school districts, state textbook adoption). In addition, several states in the South (Alabama, Florida, North Carolina, South Carolina, and Tennessee, for example) have tempered the toughest components of their accountability systems.

Outside the South, the evidence is murkier. New York and Pennsylvania have strong teachers unions, but the states' accountability systems receive high rankings in an analysis conducted by the Thomas B. Fordham Foundation (4.1 and 4.0, respectively, on a scale of 1 to 5; the national average is 3.7).[67] Massachusetts also receives a relatively high ranking, despite a strong union and a

legislature dominated by the Democratic Party, the political party most closely allied with teachers unions. The Massachusetts Teachers Association is clearly unhappy with the state's accountability system and, as noted earlier, spent $600,000 on a media campaign against it. California was not ranked by the Fordham study, but unions in the state are clearly not pleased with the state's accountability plan, which was passed by a Democratically controlled legislature and signed by former Democratic governor Gray Davis. In addition to the California Teachers Association, local chapters in Oakland, San Francisco, and Los Angeles have posted resolutions sharply critical of the state's accountability program. Yes, teachers unions are powerful, and, yes, they seek to weaken accountability systems. But their current discontent suggests that not everything is going their way on the matter.

Looking Ahead

Formidable political factors are arrayed against accountability. Progressive education supplies an ideology hostile to test-based accountability. Politically powerful teachers unions represent actors with interests divergent from the aims of accountability systems. Governance in education is both open and porous, providing multiple venues for challenging implementation. Parents, students, and individual teachers (that is, teachers acting alone, not as union members) are the opponents of accountability most prominently featured in the press.

Accountability targets marginal producers. Teachers and students who fall into that category—or who fear falling into that category—are a ready-made group of opponents. The political headwinds encountered by accountability will climax when teachers and students who try hard and nonetheless fail are sanctioned. The public is more likely to see accountability as fair if bad teachers and poor students have been given a chance to improve. But how can that be done? James Q. Wilson cites the school as a prime example of a coping agency, an organization with mostly unobservable work and ill-defined outcomes.[68] Critics of educational research and teacher-training institutions argue that the scientific foundations of teaching and learning are exceedingly weak. Education lacks a clear, reliable technology, a body of knowledge linking specific teaching strategies to specific student outcomes (including learning), one that has been codified and rigorously tested and that all teachers are trained to master. This shortcoming is politically threatening to accountability systems in the short term, but in the long term it may offer reasons for viewing accountability's future optimistically.

Accountability operates through incentives. At the classroom level, even teachers who have been motivated to change by accountability must know what to do differently to convert struggling learners into accomplished ones. That knowledge is undoubtedly specialized, contingent on such things as the age of students and the content being taught. To improve at the elementary level, where typically all subjects are taught by the same teacher, teachers must know how to do "something different" in several subject areas. Good high school math teachers and history teachers may share some pedagogical expertise, but their solutions to instructional problems also surely differ in crucial ways. Because students change every year, the solutions may change, as well.

These are all prime topics for research. Accountability may give a new urgency to educational research. It is difficult to sanction someone for an unacceptable outcome—and, in democratically governed institutions, to justify the sanctioning as fair—when no one can describe, with reliability and precision, how to produce an acceptable outcome. This is not to let teachers off the hook. I am not saying that good teachers do not exist or that it is impossible to distinguish good ones from bad ones. Teachers should be held accountable, as should any public employee. Good teachers do exist, and we can tell them apart from bad teachers. We just cannot describe in great detail what makes good teachers good, nor a series of specific steps a bad teacher can take to become good. This conundrum explains why accountability in education has historically been addressed through informal systems of evaluation and based on a supervisor's observations rather than on the hard data of formal accountability systems.

When it comes to good and bad students, the situation is similar. The rules governing principal-agent relationships also govern the relationship of adults with students. Parents and teachers attempt to elicit work from children who have their own interests and may be indifferent or even hostile to academic accomplishments. Much of learning—important insights, the "click" of connecting two pieces of knowledge, the "aha" of seeing something for the first time—is unobservable. The tasks of learning—review, practice, memorization—can be tedious. Older students, in particular, hold so many cards in the learning process that great high school teachers are often known not for their instructional acumen but for how well they motivate students. We still do not know what to do with the hardcore group of students, perhaps 10 to 20 percent, for whom learning is, sadly, primarily a daily experience with failure.

That accountability raises the stakes for teachers, students, and parents enhances the value of educational research. Indeed, accountability's long-term viability probably depends upon advances in research. With teachers' jobs and

students' diplomas on the line, we need to know much more about how good teaching works and how learning takes place. If research uncovers additional evidence, like that analyzed in the first half of this paper, that accountability raises achievement and does not have serious negative consequences, then state systems will gain additional public support. If research is also able to make new discoveries about teaching and learning—and thereby give clear guidance on how students and schools can become better—then the political obstacles analyzed here are bound to recede.

Comment by Robert M. Costrell

Tom Loveless poses the striking conundrum that just as the social science research is beginning to indicate the promise of test-based accountability, a political backlash is threatening to stop the movement in its tracks or even reverse it. The first part of his paper, on the social science research, provides the promise, and the second part, on political prospects, documents the perils. The paper provides a good overview of where we are and much insightful analysis, if not, in the end, a lot of "hopeful signs."

My comments largely draw on my experience over the past five years as an academic in state government in Massachusetts, one of the few states that has successfully instituted high-stakes testing. Since June 2003, students have been required to pass the English and mathematics tests of the Massachusetts Comprehensive Assessment System (MCAS) exam in order to get a diploma. The passing score is low, but the tests are rigorous. Consequently, a nontrivial number of students have been denied diplomas. The vast majority, however, have received diplomas that now mean something. The Cambridge and Brookline boycotts and the high-priced ads of the Massachusetts Teachers Association, of which Loveless writes, did not carry the day.

Effects on Achievement

Loveless provides a reasonable read of the literature on the effects of standards on achievement. Overall, standards-based reform seems promising, but as it is still early, data are too thin to be definitive. Only a few states have content-based graduation exams (as opposed to the old minimum-skill competency exams). Others have delayed or backed off in one way or another. So it is difficult as yet to gauge the effects of the most rigorous high-stakes testing using

the usual standards of cross-sectional statistical research. If only a few states have high-stakes testing, it is hard to be sure that it is the testing regime, rather than other features of those states, that drives improved performance.

Still, in Massachusetts we are encouraged by what is at least a happy coincidence—although we have reason to believe it is more than that—between our high-stakes testing regime and the strong performance of our students, both in levels and improvement, on a variety of external tests, including the National Assessment of Educational Progress and Scholastic Assessment Test.[69] We have also generated great improvement on the MCAS test itself, a set of exams that is widely respected (for example, by the evaluations of Achieve) and not easily gamed.

These results are consistent with a huge amount of qualitative intelligence on how the MCAS has changed practices on the ground, especially in the urban schools. There has been renewed focus on academic achievement in many concrete ways, at least in English and math, including double-block scheduling, increased writing assignments, greater emphasis on problem solving in math, and improved use of data to identify student weaknesses.[70] Based on both the sense on the ground and the data, there is broad agreement among reformers in Massachusetts that although we are still far from achieving the goal of proficiency for all, high-stakes testing has been a key element in raising achievement. This is the view not only of those who had pressed for high stakes all along but also among some who were getting cold feet as the moment of truth approached. There is little doubt among the urban superintendents (some of the strongest proponents of standards-based reform) that the mobilization for such improvement could not have occurred with lesser forms of accountability, such as school report cards. Loveless reports that Margaret Raymond and Eric Hanushek find no statistical difference between states with school report card systems and those with stronger forms of school accountability, but the study did not examine systems of student accountability.[71]

Unintended Consequences

Much of the evidence that Loveless reviews on the effect of standards on dropout rates refers to an earlier generation of standards: minimum competency exams (in Brian Jacob's paper) or course-taking requirements (in John Bishop's) rather than content-based exit exams.[72] Nonetheless, as Julian Betts and I have argued, theory suggests that with student accountability alone (absent adult accountability), higher standards should be expected to increase the number of discouraged students on the margin.[73]

Betts and I argue that even if the number of dropouts does increase, much of the change may reflect sorting effects rather than incentive effects. Some students who fail to meet the graduation requirement and drop out are those who would not have acquired additional skills had they remained in school. In Massachusetts, for example, we know that failure on the MCAS test is highly correlated with absenteeism. Students who do not attend school regularly have already de facto dropped out, even if they are not yet in the statistics, and many of them will completely drop out before the end of grade 12 quite independent of the MCAS. But even among those who previously would have graduated, at least some are simply relabeled as nongraduates by the exit exam—a pure sorting effect—so these students are no worse prepared for later life than they would have been had they remained in school under the prior regime.

Graduates with few skills may fare better than dropouts in the short run, as employers have difficulty immediately distinguishing skilled graduates from unskilled ones. But economic studies indicate that dropouts and graduates with similar skill levels eventually fare about equally well in the labor market. From a policy viewpoint, then, the question is the extent to which any effect of testing on the number of dropouts is merely a sorting effect as opposed to a genuine adverse effect on the incentive to acquire skill. That said, the political effect (to jump ahead to the second part of Loveless's paper) of increased dropout rates is probably independent of whether it is a sorting effect or an incentive effect, as such distinctions get lost in the political arena.

In any case, that there has been no significant change in the dropout rate would seem to indicate that intensive adult efforts to bring kids over the bar—and the broad social support for doing so—can offset the adverse incentive effect of having to work harder to graduate. There is no doubt in Massachusetts that exit exams have brought greater school attention than ever before to those students at risk of failing, many of whom would have dropped out anyway under the previous system, even without the exams.[74]

Politics of Accountability

In the second part of his paper, Loveless accurately identifies the political problems in establishing and maintaining accountability. Chief among these is the intensity of opposition among organized interest groups, as opposed to the more diffuse support among broader constituencies and the public at large. The Massachusetts experience suggests some key factors that make a difference in establishing and maintaining high-stakes regimes.

THE GRAND BARGAIN. In 1993 the Massachusetts Education Reform Act set out a grand bargain, committing the state to a two-phase process. The first focused on funding, including a massive increase in state aid, especially to struggling urban districts, the second on accountability, particularly the graduation requirement. A foundation budget was established, for the first time, for all school districts, and state aid grew about 12 percent a year for seven years, to bring all districts up to that level.

Once that had been accomplished, it became hard for political leaders who had been party to the first phase of the grand bargain to back out of the second. They were, of course, under intense pressure from certain constituencies, and especially from the teachers unions. The unions and other elements of the education establishment, such as the association of school committees, had been happy to take the huge infusion of money but now insisted that it was unfair to hold students accountable for results. This was a hard case to make, however, even to those political leaders who had close ties with and financial support from the teachers unions. Political leaders stuck together across party lines, in large measure because they believed in the plan but also bolstered by fear of being charged with having nothing to show for the billions of dollars spent.

ORGANIZED GROUPS AND OPINION MAKERS. As Loveless points out, the opponents of test-based accountability are backed by some effective standing organizations, such as the unions, along with energetic new organizations created for this specific issue. Such groups will often prevail over more diffuse public support for standards unless the proponents have their own organized groups with the sophistication and moral standing to influence policy.

In the case of Massachusetts, a group of business leaders (the Massachusetts Business Alliance for Education) was among the key drivers of the reforms from the beginning. They were motivated by firsthand knowledge of how poorly educated many recent graduates were as they took their first jobs. The abysmal results of employment tests provided powerful evidence of the need for reform. These business groups established credibility with the public and certain segments of the education community by advocating for a large funding commitment linked to subsequent accountability. In this respect, the Massachusetts business community defied the facile caricature often leveled at such groups by the more ideological opponents of standards-based reform. Their influence was based on the persuasive moral and practical case they were able to make as hardheaded business people, convincing legislators that both funding and standards were essential to the social and economic well-being of the Commonwealth. Unlike the teachers unions, on whom many

legislators depended for campaign funds, the business groups' financial resources were devoted to research for policy development and (in the case of business-backed Mass Insight Education) provided nitty-gritty support to the schools, such as training elementary math teachers, helping in curriculum alignment, training in the use of data, running statewide competitions for innovative schools, and the like. Generally, these groups also lobby and track public opinion, but relatively little in the way of financial resources was devoted to shaping public opinion, unlike the massive antitesting ad campaigns of the Massachusetts Teachers Association.[75]

The urban superintendents of Massachusetts schools proved to be a powerful group for reform, based on both ends of the grand bargain. The education reform money went primarily to the urban districts, giving them the resources to do some things they had never before been able to do. They also knew that the MCAS was the most dramatic mobilizing force they had seen, providing the focus of a meaningful diploma—more nearly comparable to suburban diplomas—around which to organize district and school efforts. For both reasons—fear that the money would dry up were MCAS to be suspended and fear that their educational focus would be dissipated—the urban superintendents supplied crucial support for staying the course. Interestingly enough, their support remained firm, even as some of their school committees wavered, in the face of high prospective failure rates. Together with key urban legislators, the superintendents cast an urban character to the drive for high standards. This made it difficult for suburban liberals to carry the moral high ground.

Editorial support was strong across the state for standards-based reform, including both Boston newspapers, the *Globe* and the *Herald*. The strong editorial stance of the *Boston Globe* was particularly significant because it carried the imprimatur of the state's liberal establishment.

PUBLIC HIGHER EDUCATION: MISSING IN ACTION. One might have expected public higher education to press for higher standards from kindergarten through twelfth grade, for the same reason that the business groups did: firsthand knowledge of the shortcomings of K–12 graduates. This never happened. Indeed, to the extent that any segments of higher education did weigh in, they typically did so in opposition to standards, notably from the schools of education. Faculty bodies that might have formed a counterweight, such as the faculty senates, are often dominated by the faculty union, affiliated with the K–12 teachers union. In short, the factors that Loveless cites in general—ideology and organized interest groups—were probably sufficient to keep higher education from pushing to raise the standards of their own feeder schools.

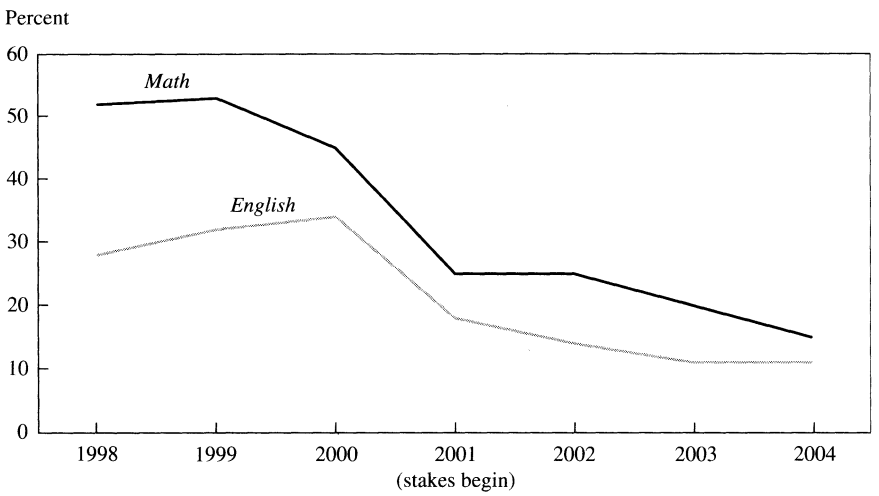
SETTING THE BAR. As a political reality, test-based accountability will not be established or maintained if it denies diplomas to too many students. The means to achieve an acceptable pass rate must themselves be seen as legitimate. This includes such obvious measures as setting an initial hurdle that is below full proficiency, with a credible ramp-up plan to proficiency, consistent with No Child Left Behind's schedule. This is easier said than done. To credibly establish high standards, the tests should be challenging from the outset rather than ramping up in difficulty; instead, it is the standard of passing performance on the test that should be raised over time. In the case of Massachusetts, in the fall of 1999 the decision was made at the highest levels of government to set the cut rate for the class of 2003 at "needs improvement." This corresponds to raw scores of approximately 52–57 percent in English language arts and 35–40 percent in math. As low as these cutoffs are, the test was sufficiently challenging that tenth-grade failure rates were as high as 53 percent in math and 34 percent in English.

GETTING A FIRST RUN WITH HIGH STAKES. The leap from trial runs of testing to the establishment of a high-stakes regime in Massachusetts was somewhat Kierkegaardian: it took a certain faith in our students' potential and our teachers' skills, as well as in the logic of human behavior, to believe that the failure rate would drop as dramatically as it did once the test mattered. The sudden reduction in the failure rates by about 20 percentage points in 2001, shown in figure 1, took quite a bit of steam out of the opposition. The failure rate among tenth graders in the first class subject to high-stakes testing was still high (25 percent in math, 18 percent in English), and the gaps among ethnic groups remained wide, but the demonstration of what a concentrated effort could achieve kept the policy alive.

RETESTS AND REMEDIATION. Students were offered multiple retest opportunities, the main ones being in the fall and spring of the eleventh and twelfth grades. In addition, up to $50 million of state funds for remediation was made available, resulting in a plethora of after-school, summer, and in-school programs directed toward helping the class of 2003 (and subsequent classes) over the bar.[76] A host of community groups provided a variety of support activities, such as volunteer tutoring efforts from some local businesses, as well as the concrete efforts by the education reform groups mentioned above. At the very least, all these activities helped create a social environment for sharp and sustained focus on passing.

As figure 2 shows, these efforts helped raise the pass rate (passing both math and English tests, to qualify for graduation) from 68 percent in the spring of tenth grade to 95 percent by the end of senior year.[77] In addition, as figure 3

Figure 1. MCAS Failure Rates, Grade 10

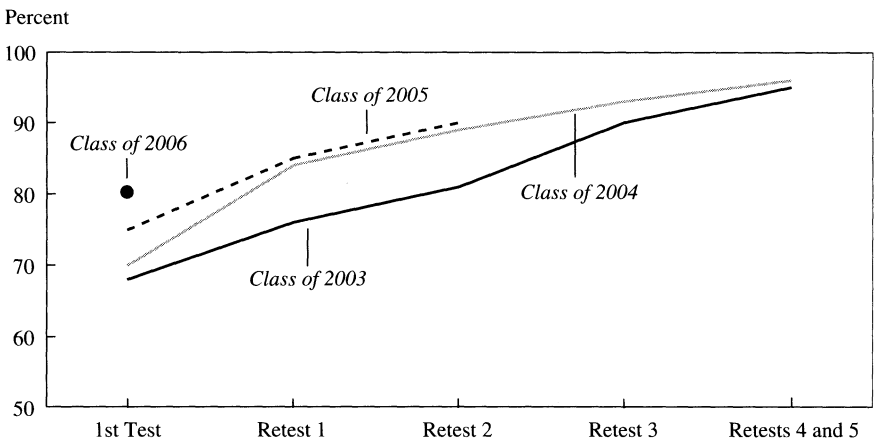


Source: Data from Massachusetts Department of Education.

shows, the retests and remedial efforts helped to sharply narrow the racial gaps in pass rates. Similarly, the pass rate of urban students rose dramatically with retests and remediation, illustrating what the urban superintendents had described as the powerful educational impact of high-stakes testing in mobilizing urban schools. Of course, the political import of the improvement in Boston's pass rate (not shown) from a low of 40 percent to 83 percent after retakes was crucial, given that Boston is the home of the state government and the state's major media. The pass rates also shot up for students with limited English proficiency and for students with special educational needs. Since the issue of special education was quite salient among suburban opponents, the improvement on that dimension had additional political significance.

If retests and remediation are to be accepted as legitimate means of raising the pass rate, there must be reassurance on how real the gains are. Do they truly signify cognitive improvement or merely better test-taking skills? No doubt there is an element of the latter, but the remedial programs certainly include much cognitive emphasis; in particular, the math exam, which is the biggest hurdle, is hard to slide through on test-taking skills alone. Moreover, the noncognitive lessons of retakes—persistence, self-discipline, and organization—have great value in the student's future, as the work of the economist James Heckman and others have shown.[78]

Figure 2. Students Passing Both Math and English, Classes of 2003 to 2006



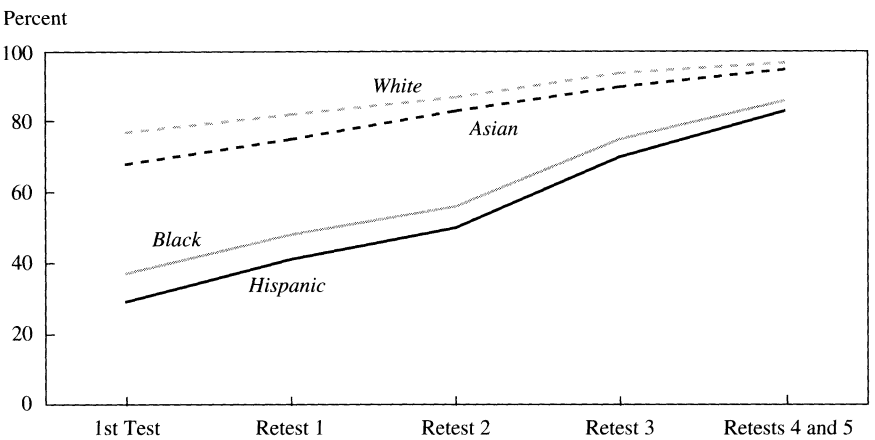
Source: See figure 1.

Finally, as figure 2 shows, pass rates on the first test have been rising among subsequent cohorts, thereby reducing reliance on retests. The rise in initial pass rates is particularly strong for low-performing subgroups (not shown), narrowing the gaps between them and other groups even without the retests. The initial pass rate among Hispanics has risen steadily from 29 percent in the class of 2003 to 52 percent in the class of 2006; for African Americans, the rate has risen from 37 percent to 60 percent. The initial pass rate for students in special education has risen from 30 percent to 50 percent.

ALTERNATIVE ASSESSMENTS AND THE APPEALS PROCESS. Alternative assessments and an appeals process can provide a necessary safety valve, but as Loveless points out, they can also be a means of watering down standards. In Massachusetts these measures helped defuse opposition in the legislature, while the criteria established were designed to maintain standards.[79] However, questions have been raised regarding some of the accommodations for students in special education on the regular MCAS exam.[80]

BIPARTISAN LEADERSHIP AND SHEER POLITICAL LUCK. Massachusetts was fortunate to have strong bipartisan political leadership on the establishment of test-based accountability, as well as an effective commissioner, chair, and members of the board of education. But it could easily have gone differently. If the Speaker of the Massachusetts House of Representatives, a conservative

Figure 3. Students Passing Both Math and English, by Ethnicity, Class of 2003



Source: See figure 1.

urban Democrat, had been less powerful, suburban legislators might have been able to force a difficult vote when failure rates were still high. The Senate also held firm under its president, a liberal Democrat who had been a coauthor of the grand bargain some years earlier. Finally, Massachusetts elected a string of Republican governors who stayed the course, narrowly defeating candidates strongly backed by the state teachers unions in 1998 and 2002; those candidates, greatly indebted to the unions, might well have "postponed" the graduation requirement.

Legal Perils and Educational Challenges

None of the foregoing is to suggest that the political perils Loveless discusses are not real, only that they can sometimes be overcome. However, when interest groups fail to achieve their goals through the political process, they can always try to impose their will through the courts. Although recent suits in state court (including Texas and Massachusetts) have not succeeded in invalidating state accountability systems, such efforts persist. Challenges that fail on one legal theory are often followed by different challenges based on a different theory. Statutory challenges are followed by constitutional challenges. Suits in state court are followed by suits in federal court. Judge shopping is not

unknown. The legal peril to accountability systems never seems to go away entirely, no matter how many cases are successfully defended.[81]

School funding lawsuits are also related to accountability systems in some interesting ways. In years past, successful finance lawsuits have sometimes led to the establishment of accountability systems, as legislatures decided that if more money were to be spent, test-based accountability for results would need to follow. More recently, such accountability systems have been used as leverage by plaintiffs in new finance lawsuits to argue that yet more funding is constitutionally necessary to get students over the bar.[82] As Alfred Lindseth argues, these cases, when successful, can distract political attention from the policies that are most promising in raising student achievement, focusing attention instead on the question of how to pay for large-scale remedies, how to distribute the money, and so on. As Lindseth also points out, if high standards continue to spawn adequacy lawsuits that result in excessively costly judgments, the natural incentive will be to soften the standards.[83]

Finally, Loveless identifies the key issue ahead of us: Unless we can devise and enact policies to bring students up to high standards, the standards movement will collapse under its own weight. As standards are raised from their initial levels toward full proficiency (and as more high-stakes subjects are added to math and English), the issue will become more urgent.

Is it a matter of not knowing what to do, or do we lack the will or the flexibility to do it? Loveless, leaning toward the former, places his hopes in further research to figure out what to do. In reaching this conclusion, he rightly points out that merely trying harder (as standards elicit greater effort) may not be enough. But that in itself does not tell us whether the obstacle is knowing what to do or having the flexibility to do it.

Examples of excellent schools with disadvantaged children are not hard to find. Certainly, the example of the KIPP academies points to fairly straightforward solutions: long days and weekends, parental commitment, high academic and behavioral standards—in short, tough love. These methods are hard to introduce under the rigid work rules in our traditional schools. On the other hand, solutions that may work for students at KIPP might not work for other students, even if the flexibility were there.

Herein lies the tension between flexibility and intervention: should we provide underperforming districts with greater flexibility (by, for example, suspending collective bargaining restrictions on personnel deployment) in the hope that such flexibility will be well used, or should we actively intervene, taking over the school and telling it what to do? In some respects, intervention does not have a good track record, but in areas in which we have a good idea

of what to do (such as making use of test data to inform instruction—one of the concrete results of test-based accountability), intervention might be fruitful. In general, we may need both, enhanced flexibility in underperforming schools along with graduated interventions should research progress far enough to determine effective solutions. There can be little hope, however, that any of these puzzles will be solved any time soon if the stimulus of standards-based accountability is abandoned.

Comment by Larry Cuban

I would summarize Tom Loveless's paper with two one-liners: *Stay the course,* and *More research will dissolve opposition to test-based accountability.* I would offer one statement to summarize my response to Loveless's paper: *Curb your enthusiasm.*

Stay the Course

Loveless claims that the policy of test-based accountability works. Rick Hess's phrase "coercive accountability" better captures both the theory behind the policy and varied state practices of test-based accountability.[84] To support his claim, Loveless cites studies that show states with coercive accountability systems having raised student achievement since the mid-1990s. In short, he argues that those states that have high-stakes tests in place will have a positive effect on students' academic achievement.

Loveless then turns to the matter of harmful, even perverse, effects of the policy of coercive accountability. He examines dropout rates, student retention in grade, misidentification of schools as failures, and the narrowing of the curriculum. He concludes that the evidence on these harmful effects is mixed. Moreover, he claims, astute policymakers can fix these potentially negative effects.

What concerns Loveless the most is that just at the moment when evidence of coercive accountability's positive effects is emerging, political opposition from parents, students, and teachers is also mounting. Moreover, he worries that states have been "backsliding" (his word)—softening some of the coercive components by deferring exit exams to a later date, providing alternative ways for students to pass a course or graduate, and making allowances for students with disabilities. States have responded in this way, according to Loveless,

because the fragmented federal-state-local system of school governance, powerful teacher opposition, and conflicting ideologies of education have created automatic dissenters to test-based accountability.

The Promise of Research

Yet Loveless is hardly discouraged by states' backsliding, the porous nature of school governance, and progressive ideologies that mobilize parents, teachers, and students to oppose test-based accountability. He is, in fact, heartened by ideological attacks upon accountability based on lack of knowledge surrounding what constitutes a good teacher or good teaching. It is precisely the coercive accountability–induced pressures on teachers, students, and parents that enhance the value of educational research. Indeed, Loveless continues, "With teachers' jobs and students' diplomas on the line, we need to know much more than is currently known about how good teaching works and how learning takes place. . . . If research give[s] clear guidance on how students and schools can become better . . . the political obstacles . . . are bound to recede." Thus my second one-liner: more research will dissolve opposition to test-based accountability.

Curb Your Enthusiasm

Why am I less enthused than Loveless about the positive effects of coercive accountability and research? First, the positive results in academic achievement that Loveless cites are, at best, mixed and offer little guidance to policymakers as to what to do. The correlational studies he favorably reports on are recent entries in the history of dueling studies about the worth of high-stakes tests. Although they point in a direction that Loveless favors, they offer no safe bets or assurance that policymakers should stay the course in test-based accountability. Moreover, these studies do not offer evidence for the reduction in achievement gaps among blacks, Latinos, and whites.

Second, none of the studies that Loveless either critiques or admires says why some states do better than others. None of them establishes the causal mechanisms behind the gains in achievement for some groups and not for others. Loveless knows this. He argues passionately that determining why a policy works is crucial and calls for more research to fill this enormous hole. He also knows that determining whether a policy works demands more than a few correlational studies offering positive results. Determining whether a policy works depends upon specifying exactly what was put into practice when, by whom,

where, and under what conditions. Such research draws from both qualitative and quantitative studies at the district, school, and classroom levels. Although Loveless's call for more research uncovers serious weaknesses in the theory and assumptions driving coercive accountability, just what is required to determine clearly whether a policy works in schools and classrooms is missing from Loveless's paper.

Third, in the final section of his paper Loveless lays out carefully and thoughtfully what we do not know about the effects of coercive accountability. In listing the areas that need to be researched, he raises fundamental questions about why we even have federal and state policies that exact strong penalties for students, teachers, and schools when such policies are anchored in so little knowledge of what works in improving teaching and learning.

Fourth, Loveless explains the growing opposition to school and student accountability from teachers, students, and parents by baldly asserting that no one likes to be held responsible for his or her actions. He does not grant opponents of coercive accountability even the possibility that their opposition may be principled, in that there is little evidence that existing policy incentives and penalties drive classroom teaching and learning, as policymakers believe they would (and should). He is stingy in conceding that critics may share the insight that he so nicely sums up in another context: "It is difficult to sanction someone for an unacceptable outcome—and, in democratically governed institutions, to justify the sanctioning as fair—when no one can describe, with reliability and precision, how to produce an acceptable outcome." More specifically, he adds: "We just cannot describe in great detail what makes good teachers good, nor a series of specific steps a bad teacher can take to become good."

This critical lack of knowledge, much less evidence, and the weaknesses in the dominant theory of action driving state and federal policymakers who have adopted coercive accountability lead me to ask why Loveless is reluctant to examine the theory's key assumptions that have driven the standards and accountability movement for the past two decades:

—Strong economic growth, high productivity, long-term prosperity, including a higher standard of living, and increased global competitiveness depend upon a highly skilled workforce.

—Public schools are responsible for equipping students with the necessary knowledge and skills to compete in an information-based workplace.

—All public schools are doing a poor job of preparing high school graduates for college and the workplace, with urban schools doing the worst job of all.

—Schools are just like businesses. The principles that have made businesses successful can be applied to schools to produce structural changes that

will improve academic achievement as measured by standardized tests, end the skills mismatch, and increase public confidence in schools.

—Higher test scores in school mean future employees will perform better in college and in the workplace.

These taken-for-granted assumptions—occasionally contested in academic journals and by lone voices in op-ed columns—have become the basis for the standards and accountability movement over the past twenty years.

Finally, if all these reasons fail to convince Loveless that coercive accountability policies are seriously flawed in their theory, assumptions, and implementation in schools and that all the research he proposes will not correct those defects, I ask one final question: What is the point of test-based accountability? It is clearly an instrumental strategy to achieve larger ends. Exactly what are those ends? Few would quibble with me if I say the end toward which coercive accountability leads is to prepare kindergartners through middle school and high school students for college—getting five-year-olds to dream of Harvard, Wisconsin, and Stanford. And why college? Because an information-based economy demands higher levels of knowledge and skills from workers than does an industrial-based economy. High school is just not enough anymore.

Yet most informed people know that between 20 and 30 percent of all jobs, according to estimates of the Bureau of Labor Statistics, require a bachelor's degree. The rest of the workforce will need short-term training, not a college education. Will we be a nation of lawyers, engineers, managers, and teachers? Who will wait tables, fix broken pipes, assemble computers, sell products, and empty bedpans? Moreover, with the press for everyone to go to college, the dropout statistics are mind bending: fewer than 60 percent of those who enter college finish a four-year degree program, and the rate of failure is even higher among minorities.

I raise these points not because I am opposed to programs that give all students a shot at going to college. My entire career in urban schools has been aimed at precisely that goal. Nor am I opposed to tests. When I was superintendent of schools many years ago, I endorsed and used standardized achievement tests to signal trouble spots and measure progress or regress. I raise these points to get at the basic theory and unquestioned assumptions inspiring test-based accountability. What is the end toward which test-based accountability takes us? If strategies like coercive accountability figure so large in the policy world, then certainly the goals toward which these strategies are bent deserve open and serious consideration.

I had hoped that a paper entitled "Test-Based Accountability: The Promise and Perils" would have at the least considered that question. My comments on Loveless's paper have an edge to them because I am still a believer in rational analysis of problems, examining the logic of a policy, its assumptions, and the available evidence when adopting and implementing policies that have major consequences on others—particularly when children and teachers, not policymakers, are forced to foot the bill for poorly thought out and implemented policies.

Notes

1. For histories of school reform in the twentieth century, see Diane Ravitch, *The Troubled Crusade: American Education, 1945–1980* (New York: Basic Books, 1983); David Tyack and Larry Cuban, *Tinkering toward Utopia: A Century of Public School Reform* (Harvard University Press, 1995).

2. Sam Dillon, "President's Initiative to Shake Up Education Is Facing Protests," *New York Times*, March 8, 2004, p. 12; Amanda Paulson, "An Education Rebellion Stirring," *Christian Science Monitor*, February 11, 2004, pp. 1, 4.

3. Audrey L. Amrein and David C. Berliner, "*An Analysis of Some Unintended and Negative Consequences of High-Stakes Testing,*" December 2002 (www.asu.edu/educ/epsl/EPRU/documents/EPSL-0211-125-EPRU.pdf); Audrey L. Amrein and David C. Berliner, *The Impact of High-Stakes Tests on Student Academic Performance: An Analysis of NAEP Results in States with High-Stakes Tests and ACT, SAT, and AP Test Results in States with High School Graduation Exams,* December 2002 (www.asu.edu/educ/epsl/EPRU/documents/EPSL-0211-126-EPRU.pdf).

4. Debra Viadero, "Researchers Debate Impact of Tests," *Education Week,* February 5, 2003, pp. 1, 12.

5. Martin Carnoy and Susanna Loeb, "Does External Accountability Affect Student Outcomes? A Cross-State Analysis," *Educational Evaluation and Policy Analysis* 24, no. 4 (2003): 305–11.

6. Eric Hanushek and Margaret Raymond, "High Stakes Research," *Education Next* 3 (Summer 2003): 48–55. Also see Caroline Hoxby, "Testing Is about Openness, and Openness Works," Hoover Institution, July 20, 2001 (wwwhoover.stanford.edu/pubaffairs/we/current/hoxby_0701.html).

7. John H. Bishop, Ferran Mane, Michael Bishop, and Joan Moriarty, "The Role of End-of-Course Exams and Minimum Competency Exams in Standards-Based Reforms," in *Brookings Papers on Education Policy: 2001*, edited by Diane Ravitch (Brookings, 2001), pp. 267–345. Also see John Bishop, Ferran Mane, and Michael Bishop, "Is Standards-Based Reform Working? . . . and for Whom?" Working Paper 01-11 (Ithaca, N.Y.: Cornell University, Center for Advanced Human Resource Studies, 2001).

8. John H. Bishop, "The Effect of Curriculum-Based External Exit Exams on Student Achievement," *Journal of Economic Education* 29 (Spring 1998): 171–82. Bishop defines states with curriculum-based external exit exams as those employing both minimum competency and end-of-course assessments.

9. John H. Bishop, "Nerd Harassment, Incentives, School Priorities, and Learning," in *Earning and Learning: How Schools Matter*, edited by Susan E. Mayer and Paul E. Peterson (Brookings, 1999), pp. 231–79.

10. John H. Bishop, "Strengthening Incentives for Student Effort and Learning: Michigan's Merit Award Program," Working Paper 01-10 (Cornell University, Center for Advanced Human Resource Studies, 2001).

11. Brian Jacob, "Getting Tough? The Impact of High School Graduation Exams," *Educational Evaluation and Policy Analysis* 23, no. 2 (2001): 99–122.

12. Bishop and others, "The Role of End-of-Course Exams and Minimum Competency Exams in Standards-Based Reforms," p. 13.

13. Carnoy and Loeb, "Does External Accountability Affect Student Outcomes?" find a statistically insignificant effect on the Hispanic retention rate and conclude that a negative effect of accountability on Hispanics' persistence cannot be ruled out.

14. C. Thomas Holmes, "Grade Level Retention Effects: A Meta-Analysis of Research Studies," in *Flunking Grades: Research and Policies on Retention,* edited by Lorrie A. Shepard and Mary L. Smith (London: Falmer, 1989), pp. 16–33.

15. Jon Lorence, A. Gary Dworkin, Laurence A. Toenjes, and Antwanette Hill, "Grade Retention and Social Promotion in Texas, 1994–1999: Academic Achievement among Elementary School Students," in *Brookings Papers on Education Policy: 2002,* edited by Diane Ravitch (Brookings, 2002), pp. 13–68. Lorrie Shepard's critique is offered on pp. 56–63.

16. Brian A. Jacob and Lars Lefgren, "Remedial Education and Student Achievement: A Regression-Discontinuity Analysis," *Review of Economics and Statistics* 86 (February 2004): 226–44; Jenny Nagaoka and Melissa Roderick, *Ending Social Promotion: The Effects of Retention* (Chicago: Consortium on Chicago School Research, March 2004).

17. Walter Haney, "The Myth of the Texas Miracle in Education," *Education Policy Analysis Archives* 8, no. 41 (August 19, 2000) (epaa.asu.edu/epaa/v8n41/). The analysis is based on ninth-grade enrollments and the declining ratio of the number of twelfth graders to ninth graders three years earlier, a falloff that began in the 1980s and extended into the 1990s. Carnoy and Loeb, "Does External Accountability Affect Student Outcomes?" analyzes Texas enrollment data and finds that the ratio indeed declined until the mid-1990s but then began rising at the end of the decade. Laurence Toenjes, A. Gary Dworkin, Jon Lorence, and Antwanette Hill, "High-Stakes Testing, Accountability, and Student Achievement in Texas," in *Bridging the Achievement Gap,* edited by John Chubb and Tom Loveless (Brookings, 2002), pp. 109–30, compute a ratio of eleventh graders to sixth graders and report flat statistics for 1989–91 (0.82) and then increases in 1992 (0.83) and 1993 (0.84).

18. Linda M. McNeil, *The Contradictions of School Reform: Educational Costs of Standardized Testing* (New York: Routledge, 2000).

19. Brian M. Stecher, "Consequences of Large-Scale High-Stakes Testing on School and Classroom Practice," in *Making Sense of Test-Based Accountability in Education,* edited by Laura S. Hamilton, Brian M. Stecher, and Stephen P. Klein (Santa Monica, Calif.: Rand, 2002), pp. 79–100, 98.

20. Diana Henriques and Jacques Steinberg, "Right Answer, Wrong Score: Test Flaws Take Toll," *New York Times,* May 20, 2001, pp A1, 34–35.

21. D. H. Bowman, "Minn. Extends Testing Contracts despite Scoring Mistakes," *Education Week,* September 6, 2000 (www.edweek.org/ew/ewstory.cfm?slug=01minn.h20).

22. Amy Argetsinger, "State Finds Problems in MSPAP Scoring; Broad Discrepancies Force Outside Review," *Washington Post,* November 7, 2001, p. B1.

23. Thomas J. Kane and Douglas O. Staiger, "The Promise and Pitfalls of Using Imprecise School Accountability Measures," *Journal of Economic Perspectives* 16, no. 4 (2002): 91–114.

24. For a critique of Kane and Staiger, see D. R. Rogosa, "Irrelevance of Reliability Coefficients to Accountability Systems: Statistical Disconnect in Kane-Staiger 'Volatility in School Test Scores,'" October 2002 (www.stat.stanford.edu/%7Erag/api/kscresst.pdf).

25. Brian Stecher and Sheila Barron, *Quadrennial Milepost Accountability Testing in Kentucky,* CSE Technical Report 505 (Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing, 1999).

26. Dan Goldhaber and Jane Hannaway, "Accountability with a Kicker: Observations on the Florida A+ Accountability Plan," *Phi Delta Kappan* 85 (April 2004): 598–605.

27. Bishop and others, "The Role of End-of-Course Exams and Minimum Competency Exams in Standards-Based Reforms."

28. Marc Fisher, "Taking a Stand on Testing," *Washington Post,* March 23, 2003, p. B1.

29. Alan Greenblatt, "One Correct Answer," *Governing Magazine,* May 2002, p. 40.

30. Kathleen Kennedy Manzo, "Protests over State Testing Widespread," *Education Week*, May 16, 2001, pp. 1, 26.

31. David J. Hoff, "Teacher Probed for Role in Anti-Testing Activity," *Education Week*, May 15, 2002 (www.edweek.org/ew/newstory.cfm?slug=36test.h21).

32. Jerry Jimenez, "Teacher Leads Another Protest over Testing," *Californian,* November 22, 2002, p. 1C.

33. "Judge Rules Teacher Who Posted Test Answers Can Retain License," *Associated Press State and Local Wire,* December 18, 2002, p. 8.

34. Gerald W. Bracey, "The 13th Bracey Report on the Condition of American Education," *Phi Delta Kappan,* October 1, 2003, p. 148.

35. John Gehring, "Students Boycott Tests in Mass. to Protest Emphasis on Exams," *Education Week*, April 19, 2000, pp. 1, 37.

36. Manzo, "Protests over State Testing Widespread."

37. John Gehring, "Mass. Teachers Blast State Tests in New TV Ads," *Education Week*, November 22, 2000, pp. 1, 37.

38. John Welsh, "Protesting the Tests, or Just Spring Fever?" *Saint Paul Pioneer Press,* April 18, 2002, p. 1A.

39. Alison Bert, "Parents Drop Testing Boycott," (Westchester County) *Journal News,* March 1, 2002, p. 1B; Alison Bert, "Scarsdale Students Boycott State Test," (Westchester County) *Journal News,* May 9, 2001, p. 1A; Kate Zernike, "Scarsdale Mothers Succeed in First Boycott of 8th Grade Test," *New York Times*, May 4, 2001, p. B5.

40. Nick Chiles, "120 8th-Graders Boycott State Test," *Newsday,* March 6, 2002, p. A14.

41. Nick Chiles, "Critics Blast Use of Test Prep Company," *Newsday,* May 10, 2002, p. A8.

42. Linda Jacobsen, "Ohio Halts K–3 Assessment after Cleveland Teachers Wage Protest," *Education Week*, December 3, 2003, p. 10.

43. Paul Queary, "Parents, Students Balk at Standardized Tests," *Associated Press State and Local Wire,* April 29, 2002.

44. Charles J. Dean, "Diploma Allowed Despite Disability; Disabled Students Can Graduate after Failing Exit Exam," *Birmingham News,* March 12, 2004, pp. 1C, 4C.

45. Katie Pesznecker, "Waiver Rules for Exit Exams Adopted; Others Denied High School Diploma Can Appeal to State Board," *Anchorage Daily News,* December 3, 2003, p. B1.

46. Mel Melendez, "Arizona Joins National Fray on Graduate Testing," *Arizona Republic,* November 30, 2003, p. 1A.

47. Jennifer Coleman, "California Sophomores Tackle High School Exit Exam," *Associated Press State and Local Wire,* March 15, 2004.

48. Randall Chase, "House Approves Delay of Tiered High-School Diplomas," *Associated Press State and Local Wire,* April 7, 2004.

49. Cynthia Kopkowski, "FCAT Threatens Graduation for 1,000," *Palm Beach Post,* April 20, 2004, p. 10A. The thousand students of the article's title refers to those in Palm Beach County only.

50. Brent Kallestad, "FCAT Failure May No Longer Be Trouble," *Miami Herald*, March 23, 2004, p. B6.

51. "More Students Winning MCAS Appeals," *Associated Press State and Local Wire,* March 29, 2004.

52. Maia Davis, "New Jersey High School Students Fail Exam, Graduate Anyway," *Record-Hackensack,* March 4, 2004.

53. Joel Stashenko, "Education Leaders Urge Legislators to Leave Standards Alone," *Associated Press State and Local Wire,* October 22, 2003.

54. Todd Silberman, "N.C. May Revisit Graduation Test; 'Is It an Exam, Is It a Senior Project?'" *Raleigh News and Observer,* March 24, 2004, p. B1.

55. Amber McDowell, "Committee Recommends Revamp of Graduation Testing Requirement," *Associated Press State and Local Wire,* January 29, 2004.

56. Gregg Sherrard Blesch, "Lawmakers Seek Answers for Those Who Fail WASL," *Vancouver Columbian,* February 26, 2004, p. C1.

57. For an account of the progressive and traditionalist arguments in math reform, see *The Great Curriculum Debate*, edited by Tom Loveless (Brookings, 2001). In the same volume, a chapter by E. D. Hirsch, "The Roots of the Education Wars," pp. 13–24, traces the intellectual history of progressive education.

58. These organizations include the California Coalition for Authentic Reform in Education (www.calcare.org [April 2004]), the Massachusetts Coalition for Authentic Reform in Education (www.parentscare.org [April 2004]), and Parents across Virginia United to Reform SOLs (www.solreform.com [April 2004]).

59. Alfie Kohn, "Test Today, Privatize Tomorrow: Using Accountability to 'Reform' Public Schools to Death," *Phi Delta Kappan* 85 (April 2004): 568–77.

60. Frederick M. Hess, "Refining or Retreating? High Stakes Accountability in the States," in *No Child Left Behind? The Politics and Practice of School Accountability,* edited by Paul Peterson and Martin West (Brookings, 2003), pp. 55–79.

61. Students Against Testing, Mission Statement (www.nomoretests.com [April 28, 2004]).

62. Please note that in the following discussion, "principal" does not refer to a school principal.

63. Terry M. Moe, "Politics, Control, and the Future of Accountability," in *No Child Left Behind? The Politics and Practice of School Accountability,* edited by Peterson and West, pp. 80–106, 82.

64. Ibid., p. 92.

65. National Education Association, "President's Viewpoint: ESEA: The Stakes Just Got Higher," March 2003 (www.nea.org/neatoday/0303/presview.html).

66. American Federation of Teachers, "Resolution, 2002: Standards-Based Assessment and Accountability" (www.aft.org/about/resolutions/2002/assess_account.htm [May 10, 2004]).

67. Richard W. Cross, Theodor Rebarber, Justin Torres, and Chester E. Finn Jr., *Grading the Systems: The Guide to State Standards, Tests, and Accountability Policies* (Washington: Thomas B. Fordham Foundation, 2004).

68. James Q. Wilson, *Bureaucracy* (New York: Basic Books, 1989). See pages 158–71 for a discussion of different types of agencies.

69. See S. Paul Reville, "High Standards + High Stakes = High Achievement in Massachusetts," *Phi Delta Kappan* 85 (April 2004): 591–97.

70. For the impact on practices in urban districts, see, for example, Mass Insight Education, "Beyond Tests and Good Intentions: What the Academic 'ER' Looks Like in Boston, Springfield, and Worcester," October 2003 (www.massinsight.org/docs/BeyondTestsandGoodIntentions.pdf [October 2, 2004]). In addition, it is interesting to read the generally positive survey responses of teachers in suburban Brookline, one of the districts most vociferously opposed to the MCAS graduation requirement. See Kevin Lang, "Analysis of Survey of Brookline Educators Regarding the Effect of MCAS on Teaching and Student Learning," October 10, 2002 (econ.bu.edu/lang/mcas-final-report.pdf [October 2, 2004]).

71. Eric A. Hanushek and Margaret E. Raymond, "Improving Educational Quality: How Best to Evaluate Our Schools?" in *Education in the Twenty-First Century: Meeting the Challenges of a Changing World,* edited by Yolanda Kodrzycki (Federal Reserve Bank of Boston, 2002).

72. Jacob, "Getting Tough?"; John H. Bishop, Ferran Mane, Michael Bishop, and Joan Moriarty, "The Role of End-of-Course Exams and Minimum Competency Exams in Standards-Based

Reforms," in *Brookings Papers on Education Policy: 2001*, edited by Diane Ravitch (Brookings, 2001), pp. 267–330, 339–45.

73. Julian R. Betts and Robert M. Costrell, "Incentives and Equity under Standards-Based Reform," in *Brookings Papers on Education Policy: 2001,* edited by Ravitch, pp. 9–55, 66–73.

74. See the series of research reports from Mass Insight Education's "Keep the Promise" project on MCAS remediation (www.massinsight.org/ktp.htm [October 2, 2004]).

75. These ads were belatedly countered by the Department of Education.

76. For details, see the Mass Insight research reports.

77. To be sure, the denominators shrink over time, as fewer students remain in eleventh and twelfth grades, but the rate of attrition has not increased since the advent of high-stakes testing.

78. James Heckman, "Doing It Right: Job Training and Education," *Public Interest* 135 (Spring 1999): 86–107.

79. For example, the appeals process normally requires students to have taken the test at least three times, to have received a minimum score that is very close to passing (except for students with special educational needs), to have maintained a 95 percent attendance rate, and to have availed themselves of tutoring opportunities. The student's grades also have to be comparable with those of other students in his or her school who have passed the MCAS test (so that a school that generally inflates grades will not generate successful appeals).

80. See Miriam Kurtzig Freedman, "Balancing the Needs of Special Education Students," Pioneer Institute 2004 Better Government Competition (www.pioneerinstitute.org/crg/competition/sp_ed04.pdf [October 2, 2004]), pp. 7–8.

81. For an overview of test accountability cases, see Jane Orris Wilensky, Katherine C. Galvin, and D. J. Pascoe, "Educational Accountability Systems: Motivation or Discrimination? A Survey of the Legal Theories Used to Challenge and Defend Educational Accountability Systems," PEPG 04-09, paper presented at conference, Fifty Years after *Brown:* What Has Been Accomplished and What Remains to Be Done? Harvard Program on Educational Policy and Governance, April 23–24, 2004 (www.ksg.harvard.edu/pepg/pdf/BrownConf/PEPG_04-09Wilensky.pdf [October 2, 2004]).

82. In Massachusetts, the same union that ran ads against the MCAS also funded virtually the entire school finance lawsuit, in which the plaintiffs touted their support for the MCAS.

83. Alfred A. Lindseth, "Educational Adequacy Lawsuits: The Rest of the Story," PEPG 04-07, paper presented at conference, Fifty Years after *Brown:* What Has Been Accomplished and What Remains to Be Done? Harvard Program on Educational Policy and Governance, April 23–24, 2004 (www.ksg.harvard.edu/pepg/pdf/BrownConf/PEPG_04-07Lindseth.pdf [October 2, 2004]).

84. Frederick Hess, *Revolution at the Margins: The Impact of Competition on Urban School Systems* (Brookings, 2002), p. 234.