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 on "Test-Based Accountability: The Promise and the Perils"

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.\textsuperscript{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.\textsuperscript{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.\textsuperscript{71}

\textit{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.\textsuperscript{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.\textsuperscript{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.  

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. 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. In addition, as figure 3
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
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. However, questions have been raised regarding some of the accommodations for students in special education on the regular MCAS exam.

**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
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,


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.parentscares.org [April 2004]), and Parents across Virginia United to Reform SOLs (www.solreform.com [April 2004]).


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


64. Ibid., p. 92.


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]).


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


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.


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).


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.
