Incentives and Equity
under Standards-Based Reform

JULIAN R. BETTS and
ROBERT M. COSTRELL

Standards-based reform is a strategy that includes specifying what is to be learned, devising tests to measure learning, and establishing consequences of performance for students and schools (for example, setting cut scores for grade promotion and high school graduation). The goal of this strategy is to raise student performance across the spectrum, especially for students from those schools, often heavily minority, where expectations are chronically low. The point is to alter incentives and change the behavior of students, teachers, administrators, and parents in a way that improves learning.

Popular support remains strong for standards-based reform, according to national polling data as well as local data in the states implementing this strategy. For example, a recent poll in Massachusetts, which is implementing one of the more rigorous sets of exams (effective for the class of 2003), indicates that 70 percent of the general population favors graduation exams. Support is slightly more emphatic from urban than suburban respondents, and somewhat broader (75 percent) from those with income under $25,000. When respondents are asked if they would still support the exams should 25 percent of students in their communities fail on the first try, support remains unchanged overall at 70 percent and rises to 81 percent among those with income under $25,000.

Nonetheless, vocal, if not yet necessarily wide, opposition has emerged in several states, in the run-up to full implementation of standards-based reforms. Objections fall into different categories. One source of discord concerns the content of what should be learned. The battles of the mid-1990s over national
content standards in history and English, and more recently in science and mathematics, have had their counterparts in the states.\textsuperscript{3} Despite continuing conflicts, certain broad (if not universal) agreement can be obtained in basic content areas (at least mathematics and English). The focus here will not be on content disagreements, but on disputes over testing and cutoff scores. However, at least some of the more vocal opposition to testing is based (if not always explicitly so) on unresolved disagreements over content standards, because tests give force to the content standards.

Opposition to testing-with-consequences is based on a simple, fundamental fact of life: Almost any change creates winners and losers. For example, technological progress has always had its losers, from the hand-loom weavers to the buggy-makers to current-day bricks-and-mortar retailers, computer illiterates, and those of low cognitive skills. The technologically caused losses of those with low cognitive skills over the last two decades have driven much of the standards-based reform movement. So, too, may standards-based reform create losers (at least in the short run) in the attempt to create more winners from technological progress. The fact that there are losers, along with winners, is not, in itself, a compelling reason to roll back the standards any more than it would be a reason to try to halt technical progress (by, say, shutting down the U.S. Patent Office). Instead, the nature of the losses must be examined and an appropriate set of policies must be crafted to minimize them.

The most obvious potential losers are those who may not meet the standard, and who may not earn a high school diploma as a result. But this is only the beginning of the analysis. For example, whether the failure rate rises as a result of sorting or whether it also reflects adverse incentive effects makes a great deal of difference. The distinction is important both for evaluating the costs of increased standards and for focusing policies to mitigate costs. Similarly, distinguishing sorting and incentive effects among the winners from various points on the educational spectrum is important.

Standards generate a mix of sorting and incentive effects. How are incentives altered by standards-based reform, for better or for worse, to encourage or discourage achievement? What are the trade-offs between some students’ losses and others’ gains, in learning or income? Do these trade-offs adversely affect equity, as opponents to standards-based reform often claim? Or is equity enhanced by raising standards in schools attended by disadvantaged students? Why does some of the most vocal opposition often come from the most advantaged districts? Finally, and most important, what steps can and should be taken to minimize the losses and spread the gains most broadly from standards-based reform?
The Economic Theory of Educational Standards

The economic theory of educational standards attempts to elucidate the likely effects on learning incentives and economic outcomes by means of a simplified model. Economic theory is applied to the subject of standards because economics offers a well-developed framework for the study of incentives, which lie at the heart of standards-based reform. It also offers a systematic method for identifying likely winners and losers, and, more important, the reasons behind and nature of the gains and losses. Finally, economic theory helps point to policy measures that might ameliorate trade-offs (a familiar phenomenon in economics). However, the economic analysis of standards also has limitations.

The analysis largely focuses on the passing score required for an educational credential, for a given test, covering a given set of content standards. Consider the effect of a rise in the cutoff in a simple pass-fail situation, with a single undifferentiated diploma. All the theoretical models that we are familiar with predict a rise in the failure rate, along with other, more salutary, effects. This literature is silent on the magnitude of the rise in the failure rate (which is critical in comparison with the beneficial effects), but it does help distinguish between more and less compelling reasons for concern.4

Sorting Effects of Graduation Cutoffs

Consider first a simple sorting model, in which behavior and thus learning are held constant, independent of the standard. A rise in the cutoff merely relabels some students as failers who would otherwise be considered passers. There is, by assumption, no effect on learning or productivity. The aggregate income generated by the students is unchanged, but the distribution of it does change. The students who pass are now a more elite group, so their average productivity is higher. To the extent that graduates are pooled together in the eyes of employers (who may make only limited use of individual information, as economist John Bishop has long argued), their wages tend to rise. This point is well understood: Higher standards raise the value of a high school diploma.

Less widely understood, however, is that higher standards also raise the average quality of the pool of nongraduates, insofar as some students who would previously have passed now fail. Because nongraduates (like graduates) are evaluated by employers in part on the average quality of their pool, their wages also tend to rise. This is not a minor point. The reason nongradu-
ates typically fare so poorly under the existing system is that the ease of social promotion exacerbates the stigma attached to nongraduation. Thus, it is a logical fallacy to argue, as many do, that higher standards will reduce more students to the current economic level of nongraduates; the stigma on nongraduates depends on their average quality, and that depends critically on the standard itself.

A rise in standards thus leads to gains for two of the three groups—those at the top, who graduate, and those at the bottom, who would not have graduated anyway. The losers are those in the middle, who would have graduated under a less stringent standard, but who now fail. Those individuals suffer from being pooled with a group that includes those less skilled than themselves (those without the diploma) instead of with those more skilled than themselves. No efficiency loss has occurred in this pure sorting model, only a distributional effect stemming from the individuals’ relabeling. Do these losses constitute a compelling case against higher standards? The answer is no, for two reasons.

First, in terms of the narrow choice between high and low cutoffs, a high cutoff does not necessarily lead to less egalitarian outcomes. The redistribution is from the losers in the middle to the winners at both the top and the bottom. Those with the most egalitarian preferences (so-called Rawlsians, after the philosopher John Rawls) place the highest priority on raising incomes at the bottom, so they should favor a rise in standards. The equity implications of higher standards are not limited to those who are at increased risk of failing but include those who would fail in any case, and whose stigma stands to be reduced.

Second, the standards themselves are not at the heart of the losses from adverse pooling. The crux of the matter is the imperfect information that underlies such pooling. How disturbing is it if someone loses from no longer being confused with those of greater skill? While those able students who are now pooled with those of lesser talent should be of concern, the answer is not necessarily to reverse the rise in standards and reclassify them with those of greater talent. Perhaps, instead, information flows should be improved, if possible, such that individual talents are more accurately conveyed than with a simple binary pass-fail credential, as Bishop argued.

**Incentive Effects of Graduation Cutoffs**

The losses incurred from sorting may not be of first-order policy importance, but neither are the gains from sorting the reason for implementing
standards. The rationale for standards is to alter incentives of students, parents, teachers, and administrators to change behavior in a way that advances learning. Microeconomic analysis, the study of how rational actors respond to incentives, may offer some insights.

Economic theory predicts that the effect of raising the graduation cutoff depends on where students lie in the distribution of ability and attitudes toward study. Suppose the cutoff is raised from a level at which 10 percent fail to one at which 20 percent would fail under existing behavior. Under a pure sorting model, in which behavior is held constant, this rise in standards would lead to a doubling of the failure rate. Under a more realistic model, students (and their parents) respond to higher standards by reevaluating the costs and benefits of student effort.

How are the incentives for student effort affected at different parts of the distribution? Consider first those students at or near the twentieth percentile under the original distribution of achievement. In this example, these are students who passed under the old standard by a margin of ten percentiles, but who are now just on the margin of passing under the new standard. It would take only a small increase in their effort for a number of them to pass. The cost in doing so would be less than the substantial benefit of passing instead of failing, and so the higher standard will have a positive incentive effect on utility-maximizing individuals in this part of the distribution. As a result, one can predict with some confidence that the failure rate will not rise as much as would be naively predicted under the preexisting distribution of student achievement, because students in this part of the distribution will rise to the challenge.

These students, the ones for whom the most positive response is predicted and who have the most to gain from higher standards, are not the elite (they are near the twentieth percentile in this example). Unlike the elite, who will easily pass the higher standard with unchanged effort, these students are stimulated to higher effort because otherwise they will fail. These students are typically non-college-bound or marginally college-bound. For those non-college-bound students who rise to the challenge, the benefit is a high school diploma of enhanced value—a matter of great importance for those who will not have a college degree with which to distinguish themselves. For the marginally college-bound, the benefit of being prodded to meet a higher standard is better preparation for college, which, in turn, raises the probability of successful college completion.

However, the incentives are different farther down the distribution. Specifically, consider those students who are on the margin of failing under the old
standard (students at or slightly above the tenth percentile in this example). The effort they are exerting yields expected benefits that barely exceed the costs of the effort. A rise in the standard reduces the probability of passing with that level of effort and thereby reduces the expected benefit below the cost. For these students, the rise in standards has a negative incentive effect, leading them to reduce their effort, discouraged by the low prospects of success. They may drop out of school, as critics of standards-based reform warn. This effect is more troubling than the sorting effect, because it reduces the amount of learning in this portion of the distribution.¹¹

Thus, standards have different effects on students in different parts of the distribution, even among those of lesser achievement.¹² Four groups of students are at risk of failing under the higher standard (see figure 1):¹³

1. Some students who met previously low expectations will be stimulated to greater effort by a rise in standards, with the help of teachers and parents. (In figure 1, the dashed distribution of productivity depicts a rightward shift from just left of the new standard.) These are the most important gains from high standards.

2. Other students who would have passed under low standards will not change their behavior and will now fail. (In figure 1, these students remain
between the old and new standards, on the dashed distribution.) These students lose, but only by virtue of being relabeled.

3. Other students, farther down the distribution, will be discouraged and reduce effort or drop out. (In figure 1, the dashed distribution of productivity depicts a leftward shift from just right of the old standard.) These are the most important potential losses from high standards, toward which mitigating policies should be aimed.

4. For those students at the very bottom (the left-most portion of figure 1), who would not pass anyway, behavior is unaffected, but they may passively gain from the sorting effect.

Policymakers and others may differ on how to weigh the fortunes of these groups in arriving at the optimal set of standards. The way out of this dilemma is not necessarily to forgo the benefits of higher standards, but, if at all possible, to craft accompanying policies for those students whose efforts may flag, especially those who might drop out.

Much of the most vocal opposition to standards-based reform, however, comes from a completely different segment of the population—that of generally high achievers. For example, according to recent reports, “Wisconsin scuttled plans for a high school exit exam after a protest lodged mainly by more-affluent parents.” Similarly, efforts in Massachusetts to boycott the statewide exams have been concentrated in affluent and high-achieving suburbs, as well as high-spending communities such as Cambridge, rather than such urban areas as Boston. State representative Ruth Balser told a group of Brookline test critics that most of her legislative colleagues support the exams. “It’s just those of us from districts that were already doing really well, like Lincoln-Sudbury, Brookline, and Newton, who feel that our systems are at risk of being dragged down by ed reform,” she said.

Perhaps the most plausible claim that suburban critics have to offer is that higher-order skills may be deemphasized by teachers of high-achieving students, students who are at relatively low risk of failing. It is not entirely clear why this would be so at the high school level, if students are sorted among basic and honors classes. The more elite students, aiming for selective college admissions, are more likely focused on the Scholastic Assessment Test (SAT), Advanced Placement (AP) exams, and a high school transcript enhanced with high grades in honors courses than on high school exit exams. However, if the school reallocates resources, or changes its teaching methods to bring up those at risk of failing, these equity-enhancing efforts could adversely
affect those of high achievement. If so, these objections to standards-based reform are not based on equity concerns, but the opposite.

The policy implication is not necessarily to forgo the benefits of higher standards, just because they may be concentrated among those for whom expectations are low, relative to the high-achieving critics. The challenge is to meet these objections by accompanying the standards with policies addressed toward the high achievers as well. This is an easier and less pressing challenge than the one concerning lower achievers, who might be discouraged from continuing academic effort.

Centralized versus Decentralized Standards

What is the proper locus of standard-setting—federal, state, or local? The movement toward standard-setting began with the states in the late 1970s (minimum competency testing), shifted toward the federal level from the late 1980s to the early 1990s, and has shifted back to the states since the mid-1990s, where it has made its greatest strides. Leaving aside the question of where content standards should be set, economic theory does have something to say about whether graduation cutoffs should be set locally or centrally.

In the simplest case, where all districts are alike, decentralization probably would lead to inefficiently low standards. Suppose each district’s non-college-bound graduates are pooled to some extent with graduates of other districts in the labor market. That is, employers do not fully distinguish graduates of any district that chooses a different standard. The reward to raising standards in any given district is thus attenuated. The district’s graduates would be of higher quality, but would not be fully identified as such, and so would only reap some of the benefits; the rest of the gains would spill over to graduates of other districts, with whom they are pooled in the labor market. As a result of this externality, local standard-setters have an incentive to free ride on the standards of other districts, establishing cutoffs that are too low to maximize their collective welfare. A centralized standard-setter would avoid this problem.

Even in this simple case, with identical districts, there are winners and losers in the choice between decentralized and centralized standards. Given that centralization raises standards, the winners are those who rise to the challenge, and the losers are those who become discouraged from exerting effort. But each district would, on the whole, be better off with a centralized standard-setter choosing the same cutoff for all districts. This logic is independent of the weights attached to winners and losers. Even the most egalitarian collec-
tion of standard-setters would prefer standards set centrally, rather than each of them riding free in a standard-cutting race to the bottom.\textsuperscript{24}

Heterogeneity across districts makes things more complicated but is also an important factor in understanding current controversies.\textsuperscript{25} For example, centralization typically raises standards in low-achieving districts but may lower them in high-achieving ones. To the extent that diplomas reflect some degree of district reputation (that is, pooling is not total), this means low-achieving districts’ graduates benefit from the rise in their standard while those from high-achieving districts lose from the drop in theirs.\textsuperscript{26} Thus, a conflict of interest may arise between those high-striving urban black students whose diploma is enhanced in value and those suburban students whose diploma could be depreciated from that obtained under decentralized standards.

With heterogeneity across districts, centralization need not always outperform decentralization.\textsuperscript{27} However, suppose the centralized standard serves as a minimum requirement for graduation, with the localities retaining the option of setting a higher standard. This arrangement outperforms decentralized standard-setting and is at least as good as central standard-setting without the local option. The result is the best of both worlds, with the centralized minimum standard putting a floor on free-riding by districts, while the high-achieving districts retain the option of exceeding that standard, if enough of the benefits accrue to their own graduates.\textsuperscript{28}

The model considered here helps frame questions that arise from current controversies. For example, in Massachusetts (among other states), the demand for local control of graduation requirements is strongest in the suburbs, while urban superintendents are generally the biggest supporters of rigorous state standards (even though their students are most at risk of failing). The urban districts suffer from a poor reputation but have still found it difficult to unilaterally raise it. One possible explanation that goes beyond the simple model but is consistent with its spirit is that a district’s reputation adjusts only slowly to its own actions. A long period of low standards will result in a low reputation, but a unilateral rise in standards may only raise the reputation over time, increasing dropouts in the short run with no reward. On this view, the imprimatur of state standards promises to be a more informationally powerful signal, more readily recognized, than the urban districts could establish on their own. Political considerations beyond the model probably are also important. The state mandate provides valuable cover to superintendents who would like to raise standards but who face local political and union obstacles to doing so and to taking steps necessary to meet them.
The model assumes some pooling, or blurring of credentials across districts even in the long run. If there is no such blurring of credentials—if each district’s diploma is fully understood by employers to represent that district’s own graduation cutoff—then the model’s case for decentralized standard-setting is stronger. But even then, high-striving students in low-achieving districts suffer from having their accomplishments depreciated by the low standards that local authorities tend to set in those districts. If policymakers are able to reduce the degree of cross-district pooling to lessen the need for centralization, then why not reduce intradistrict pooling as well, so that high achievers in any district can be evaluated by their individual accomplishments?

**Binary Credentials versus Fuller Information**

John Bishop has long argued that credentials such as a high school diploma, which convey only a binary signal to employers, are far inferior to richer and more finely graded information flows, such as those conveyed in high school transcripts. Economic theory speaks to the incentive and equity implications of improved information flows and largely bears out Bishop’s argument. A difficult question, however, is why employers often choose not to use the fuller information flows that are available. This question, to which no satisfactory answer has been found, is important in designing policies to ameliorate the trade-offs carried by a system of binary credentials.

In understanding the effects of improving information flows over that of binary credentials, sorting effects must be distinguished from incentive effects. Consider the simplest case, in which a single measure of productivity (such as a test score) is available, but a credential truncates that measure into a pass-fail signal. In a simple sorting model, in which behavior is assumed constant, the truncation of full information redistributes income by pooling. Among those who fall below the cutoff, the average income is unchanged, but it is redistributed from those just below the cutoff toward those at the very bottom, with whom they are pooled. Similarly, among those above the cutoff, the truncation of full information redistributes from those at the very top downward to those just above the cutoff. Thus, in the simplest sorting model, binary credentials generate outcomes that are more egalitarian than full information. However, even within the confines of these assumptions, the case for redistribution by blurring of differences is not compelling, unlike a case based on improved incentives.

Even before considering incentive effects, however, another aspect of sorting bears examination—the issue of job matching. Better sorting improves the
match between workers and jobs. Truncating information with a binary credential reduces the efficiency of the match and reduces output. Who bears the brunt of the lost efficiency: those at the top or those at the bottom? In one recent model the answer depends on where in the job ladder accurate sorting is most important.29 Suppose it is most important at the top; that is, getting the very best people into the very top jobs is more important than getting the least productive people into the very bottom jobs. Then the burden of the efficiency loss from truncating information will tend to fall on the least skilled, and this can outweigh any beneficial pooling effect they may enjoy. The wage earned by the least skilled depends on the ability of those higher up the job ladder who can only do those top jobs with the support of those lower down. If those who will fill the top jobs are not as well identified, because of truncated information, then the reward to the least skilled for supporting those in the top jobs will fall. In this case, the use of full information enhances both efficiency and equity.

Now consider the incentive effects of full information.30 If employers have and use individual information, diplomas and standards become irrelevant, because they add nothing to it. Each student chooses his or her own preferred level of achievement and is rewarded accordingly. More realistically, information flows can be improved by generating a discrete number of differentiated credentials. Either way, fuller information affects incentives in different ways across the spectrum of students.

Compared with a coarse pass-fail signal, better information about high achievement is a stimulus to those at the top of the distribution, who would otherwise find no payoff in exceeding the cutoff. This provides much of the answer to the criticism that high-achieving districts are dragged down by standards-based reform. High-achieving students are already motivated to excel by an array of credentials over and above high school graduation exams (for example, SAT and AP exams). If these are insufficient, differentiating diplomas based on the level of performance on the graduation exams, as a number of states do, is a relatively simple matter.

Moreover, differentiated consequences for differentiated credentials seem particularly straightforward to arrange for college-bound students. Admission to public higher education can be made contingent on higher performance levels than are required for graduation; scholarships can be based on higher levels yet. These credentials may be multidimensional, for those who find traditional graduation requirements overly narrow. For example, students can place many credentials based on artistic and musical talent on their college applications. There are literary contests, outlets such as the Concord Review (for historical essays),
and science fairs, to name just a few more credentials that high-achieving students can aim for, with confidence that they will be recognized.

Schools might arguably be under pressure to divert attention from these types of credentials toward the graduation exam, even for those students who are at no risk of failing. There could be some truth to this, insofar as districts reap rewards based on mean exam scores, instead of pass rates only (for example, the real estate market may tend to do this). However, this effect should not be exaggerated, because districts will continue to be attuned to how well their students do in college admissions, which still rests on these other types of credentials. That is why some high-achieving districts choose not to teach to the graduation exams any more than is necessary to achieve passing performance. In short, the introduction of graduation exams only adds information to the existing array of high-end credentials and should not pose any serious incentive problems for high-achieving students.

At the bottom of the distribution, the incentive effect from fuller information should also be positive. Those students who have no other way to convey their skills short of a graduation standard that is beyond their will or ability to meet would certainly gain from finer signals. As John D. Owen points out, fuller information at this end of the distribution advances egalitarian goals by giving students less extreme alternatives to dropping out.31

This rationale is behind the proposal that students who repeatedly fail the state graduation exam might receive instead a local diploma or a local certificate of completion. Such a credential could convey the achievement of noncognitive skills such as persistence, punctuality, and discipline that are also important and rewarded in the labor market.32 The general equivalency diploma (GED) already exists as an alternative credential and should continue to signal a certain level of cognitive skills. But its payoff in the market is considerably less than a high school diploma, probably because it does not convey the same level of noncognitive skills as even a diploma based on seat time alone.33 So room remains for a credential to certify such noncognitive skills (which may be particularly important for some special education children).

The challenge is to make sure that such a noncognitive credential is properly differentiated from a standards-based credential that signifies both cognitive and noncognitive skills, and that it is treated as such by end-users (employers or colleges). This is at the heart of the dispute between those who would grant a local diploma option and those who would allow only a local certificate of completion. For reasons perhaps better understood by psychol-
ogists than economists, such terminological distinctions seem to be empirically important.

The concern is that a local diploma would not be treated with sufficient differentiation from a state diploma and would thereby undermine incentives for those students who would otherwise meet the state standard. (This seems to have been the rationale for New York’s decision to phase out the local diploma option, leaving only the Regents diploma.) A certificate of completion could and perhaps should convey the same information that a high school diploma currently conveys in those states where the requirements are almost entirely local (such as Massachusetts, until the state standards bind in 2003). Once employers recognize that a certificate of completion is equivalent to the old local diploma, there should be no basis for objecting that students are denied a diploma by the higher state standard. Diploma is only a word. If it takes a different word—certificate versus diploma—to differentiate those who have met the old local standards from those who meet the new state standards, then this would provide the finer information flows that are called for. Remaining will be those who object to such differentiation, as to all differentiation, on the grounds (perhaps unstated) that it will deny certificate-holders the benefits of being pooled with those who hold diplomas. But such sorting arguments are not persuasive.

Finally, consider students between those near the top and those near the bottom, those who would meet the state standard, but not by much. These are students for whom the incentive effects of full information are negative. They are students who rise to the challenge of the standard only because the alternatives are so much worse. If information flows are improved, these students would choose to meet a lesser level of achievement that has a lesser payoff, but not as dramatically as dropping out. The problem here is that too many students evaluate the payoffs to higher achievement differently from adults, such as their parents or state standard-setters or from the adults that they will become themselves. The labor market signals to students are somewhat remote, and many students are notoriously present-oriented. Furthermore, schools likely have a greater incentive to bring students up to a given standard when the alternative is dramatically worse than simply meeting a lesser standard. In short, while the coarse instrument of pass-fail blunts incentives for those at the bottom and the top, it does elicit greater effort from those near the passing margin.

A key policy dilemma thus emerges from our theoretical analysis: How much differentiation should exist between the state-certified standards-based
diploma and any lesser credentials? If the differentiation is too large, then students near the bottom will have no incentive to achieve beyond the low level certified by the lesser credentials. If, alternatively, the gap between the lesser credentials and the state diploma is too small (as with continuous measures, such as the test score itself, affixed to the diploma or the transcript), then too many students who might meet the state standard would be willing to settle for less, especially if employers ignore the differentiation.

Our theoretical analysis shows that some problems alleged by critics of standards-based reform are not particularly compelling, notably those based implicitly on the logic of pooling and those concerning incentives for high-achieving students. But it also points to a trade-off between incentives for those lesser achieving students who will be stimulated to meet high standards and those low-achieving students who will be discouraged. The analysis clearly indicates that the key to ameliorating this trade-off is not so much one of setting the standard high or low as it is one of filling in the information spectrum with credentials that allow lesser achieving students to demonstrate their cognitive and noncognitive skills. The optimal degree of differentiation among these credentials can probably be worked out only in practice over time, by trial and error, because it depends much on the way employers will treat different credentials, which is not something that is easily foretold.

A Description of Current State Educational Standards

Effective educational standards require the following three components:

1. Content or curriculum standards that clearly delineate what students should learn in each grade.
2. An assessment system that measures student progress toward mastery of the content standards.
3. An accountability system that stipulates a set of rewards and interventions based on student progress. Such a system should hold accountable not only students but also teachers, principals, and entire school systems for the rate of learning of students.

How close are the states to implementing educational standards that fit these criteria, and how do states vary in that regard? Complicating the analysis is that, even though standards in practice typically resemble the binary pass-fail model, these standards have taken many forms. Some states have implemented high school exit exams. Other states have left the task of assessment to indi-
individual schools but have set minimum sets of courses that students must complete before graduating from high school. Some states also use achievement scores to make decisions about whether to promote students from one grade to another, or to assign students to remedial or other courses.

Throughout the 1990s states’ graduation requirements varied radically. For instance, according to the Department of Education, in 1993 the number of courses states required students to complete before graduating with a standard diploma varied from thirteen in California and Wisconsin to twenty-four in Florida and Utah. By 1996, California still required only thirteen courses to graduate, but Wisconsin had increased its graduation requirements from thirteen to twenty-one and a half. At the top end, three states—Alabama, South Carolina, and Texas—either had joined or were about to join Florida and Utah in requiring twenty-four courses for high school graduation.35

These variations in course requirements become stronger once the specific courses required to graduate are examined across states. For instance, in 1996, over half of the states required that high school students take at least two math courses to graduate. Another fifteen states required three courses, and two states (Alabama and South Carolina) required four. A number of states’ requirements defy a simple categorization. Colorado, Iowa, Massachusetts, Michigan, Minnesota, and Nebraska rely mainly on local boards to set graduation requirements.36 In other states, including perhaps most notably California, districts are free to impose their own additional requirements.

Several states have more than one class of diploma, to recognize advanced achievement. The American Federation of Teachers (AFT) reported in 1999 that twenty states offered advanced diplomas, up from only eight in 1996.37 Perhaps most famously, New York for over a century has offered the Regents examinations and the Regents diploma as an advanced diploma to supplement local diplomas. The creation of multiple credentials can increase the efficiency with which schools transmit information on students’ strengths and weaknesses to the labor market, provided the credentials are sufficiently differentiated from one another.

In the late 1990s, New York decided to begin phasing out local diplomas in favor of requiring all students to acquire a Regents diploma. This transition process had not yet finished by 2000. By moving to eliminate the lower tier of high school diplomas, the state of New York will in a sense be restricting the flow of information between schools and the labor market. Most other states have been moving in the opposite direction, providing additional credentials or recognition to students who surpass the minimum achievement
levels required for graduation. New York deserves to be closely studied over the next few years. The abolition of local diplomas may make it more difficult for employers to evaluate the skills of the middle group of students—high school graduates who currently do not qualify for Regents diplomas. Alternatively (although authorities have given no indication of this), New York may yet decide in the future to award certificates of completion to students who would previously have received a local diploma. If so, they will merely be relabeled. But it will be important to ascertain how employers and institutions of higher education respond to such relabeling, for that will govern the incentives generated for students. Clearly policymakers in New York are working on the assumption that eliminating the local diploma option will generate positive incentive effects for most students to work harder.

Educational standards will in practice include far more than stipulations about the number of courses required. For instance, standards must also include descriptions of the content that schools expect students to master. The AFT has published an annual review of each state’s content standards, assessment, and accountability systems. Recent trends in the number of states “with clear and specific standards,” “with assessments aligned with the standards,” and “with promotion policies based on achievement toward the standards” are represented in table 1. For a state to qualify as having clear and specific standards, AFT researchers had to determine that the state had clearly worded and specific content descriptions in English, math, science, and social studies at the elementary, middle school, and high school levels. The second AFT variable measures the quality of states’ assessment systems, while the third partially describes the states’ student accountability system. (Unfortunately, the AFT report does not include as detailed information on the ways, if any, in which teachers, principals, and district administrators are accountable for the performance of their students.)

The data in table 1 reveal that by all three measures—content standards, assessments, and student accountability—the national trend is toward more stringent requirements. The data also indicate large variation across states in these three components of educational standards and accountability.

Equally important, the AFT study shows a disturbing pattern: All states but Iowa, Montana, and North Dakota have implemented or plan to implement tests or other assessments that are aligned with their standards, yet only twenty-two states have implemented content standards that the AFT deems clear and specific. Lack of clarity in standards will create difficulties for teachers. In
many cases, states have purchased off-the-shelf standardized tests that do not necessarily link well to the content standards.

For example, beginning in spring 1998, California required that all students take the Stanford 9 tests. In the first year, the test items were not altered to reflect the state’s newly developed content standards. In spring 1999, the state added a battery of questions that more closely reflect content standards but is not yet using results from this add-on to the Stanford 9 tests to evaluate schools.

In 1996 almost no states based decisions to promote students to the next grade on standards, but, by 1999, thirteen states had such policies in place. This number underestimates the extent to which schools base promotion decisions on objective assessment measures such as achievement tests. Many school districts have gone beyond existing state promotion policies and implemented their own criteria, and interventions, for student promotion. Particularly well known is the ambitious program implemented by the Chicago public schools in 1996–97. Other districts have followed suit. For instance, the San Diego Unified School District, one of the ten largest in the country, in 2000 implemented its own radical program for assessment, additional spending on students lagging behind in reading, and, if necessary, summer school and grade retention.

Promotion policies represent only one of the many ways in which policymakers can link standards and assessment to overall accountability. Another incentive for students that a large number of states have adopted is high school exit exams. According to the AFT, twenty-eight states currently have or plan to implement graduation exams that are aligned with the state’s curriculum standards.

---

Table 1. Number of States with Various Components of Standards in Place, by Year

<table>
<thead>
<tr>
<th>Year</th>
<th>Clear specific standards</th>
<th>Assessments aligned with standards</th>
<th>Promotion policies based on standards</th>
</tr>
</thead>
<tbody>
<tr>
<td>1995</td>
<td>13</td>
<td>33</td>
<td>n.a.</td>
</tr>
<tr>
<td>1996</td>
<td>15</td>
<td>42</td>
<td>3</td>
</tr>
<tr>
<td>1997</td>
<td>17</td>
<td>46</td>
<td>7</td>
</tr>
<tr>
<td>1998</td>
<td>19</td>
<td>47</td>
<td>7</td>
</tr>
<tr>
<td>1999</td>
<td>22</td>
<td>49</td>
<td>13</td>
</tr>
</tbody>
</table>

Note: The counts include the District of Columbia and Puerto Rico. n. a. = Not available.
The most difficult aspect of implementing a graduation or exit exam apparently is to design the exam so that it links well to curriculum standards. It can be done, but it cannot be done quickly. For instance, California published science and history-social science content standards in 1999, on the heels of adoption of language and math standards the year before. The state plans to require that all students pass a high school exit exam before leaving school, beginning in the year 2003–04. California’s efforts to implement a school-leaving exam that is well articulated with its content standards have led to delays in the program. Not a single commercial test-preparation firm submitted a bid in response to the state’s tender in fall 1999, apparently because of concerns that it was not possible to prepare a specifically tailored test for a trial run in spring 2000.

A third aspect of accountability is whether states complement the stick of grade retention with the carrot of incentives for students to excel. The AFT reports that twenty states offer advanced diplomas to recognize exceptional achievement. Eight states also grant preferential college admissions or college financial aid to top-performing students. Others, such as California, are in the process of implementing such policies. A probable weakness of the carrot-and-stick system of educational incentives for students is that the students who vie for the carrots are a different group from those who face grade retention. By the start of high school, some students are likely to view college attendance as a somewhat dim prospect. What positive incentives can be created for such students remains to be seen, especially given the possibility open to high school students to drop out of school altogether.

A state educational policy that focuses on only one or two of the three pillars of educational standards—content, assessment, and accountability—is likely to achieve little. How many states have passed muster, at least according to the AFT, in all three of these categories? Because student accountability can take many forms, we list a state as having implemented student accountability if it has implemented or has plans to implement either promotion policies based on content standards, high school exit exams, or differentiated graduation diplomas to recognize students achieving beyond the requirements for a basic high school diploma. We categorize a state as having succeeded if the given accountability measure was implemented in either elementary, middle, or high school. (For this reason, the numbers in our state-by-state calculation differ somewhat from the aggregate results reported by the AFT and shown in table 1.) Our calculations of the number of states that fit into each of eight possible categories are presented in table 2. The results are revealing: Only a
handful of states—California, Georgia, North Carolina, South Carolina, and Virginia—have succeeded in all three categories so far. Moreover, seven states had not implemented any of these three types of educational standards to the satisfaction of the AFT researchers. These states were Connecticut, Iowa, Montana, North Dakota, Pennsylvania, Rhode Island, and Wyoming.

**What Explains Variations in State Standards?**

Given the considerable variations in standards across states, knowing what causes these variations becomes important. Proponents of national standards may worry that as states set their own standards, states in which student performance lags the most will have an incentive to do the least to implement educational standards. After all, not many incumbent politicians will want to create an assessment system that might show that most of the state’s children are failing to meet expectations. However, the existing federally mandated National Assessment of Educational Progress (NAEP) data, which beginning in the 1990s released results by state, may have induced legislators in states that fared poorly to implement content standards, state testing, and student accountability.

State population represents a second factor that might influence the extent to which states have implemented standards. Smaller states will have less incentive to set standards high, because of free riding. Larger states are also likely to have progressed further simply because in such states the fixed cost of developing content standards, tests, and accountability mechanisms can be spread over a greater number of taxpayers.

---

### Table 2. Number of States Meeting Three Criteria in at Least One of Elementary, Middle, and High School Grades, 1999

<table>
<thead>
<tr>
<th>Number of states</th>
<th>Clear standards in all core subjects</th>
<th>Assessments aligned with standards in all core subjects</th>
<th>Promotion or exit policies</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>12</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>1</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>9</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>4</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>3</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>11</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>7</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>


Note: The counts include the District of Columbia and Puerto Rico.
The degree of socioeconomic homogeneity, and the overall socioeconomic status (SES) of the state population, may also influence standards. On the one hand, states with fewer disadvantaged families may set higher standards in the belief that most students will be able to fulfill them. On the other hand, those states with greater socioeconomic heterogeneity, and lower socioeconomic status more generally, might do more to implement standards in the conviction that such policies can improve the life outcomes for the most disadvantaged students.

To test these three propositions informally, we first calculated an overall measure of the quality of standards based on the three measures listed in Table 2. Each state (but not the District of Columbia or Puerto Rico) was allocated from 0 to 1 point for each of the three components of standards. For content standards, we calculated the proportion of the four core subject areas that according to the AFT have clear and specific content standards in at least one grade-span. Thus this measure can equal 0, 0.25, 0.5, 0.75, or 1. Second, each state earned either 0 or 1 point depending on the AFT judgment on whether it had implemented student assessment sufficiently linked to the content standards. Third, to capture the extent to which states have established student accountability, each state earned either 0, 0.5, or 1 point based on whether it had implemented promotion criteria based on the standards or exit exams aimed at grade-ten standards or a higher level. These three measures were then added together. A state that had failed by 1999 to satisfy any of the AFT criteria would receive a score of 0; a state that had satisfied all the criteria would receive a perfect 3.

We then calculated the relation between this overall measure of the quality of state standards and measures of student achievement in the mid-1990s when most states were just beginning to implement rigorous standards. We used three different measures: the percentage of public students scoring at the basic or higher levels in the 1994 fourth-grade reading assessment on the National Assessment of Educational Progress, the analogous percentage in the 1996 fourth-grade math assessment, and the average of these two achievement measures. We also calculated the correlation between our overall measure of standards and the natural log of population in the state in July 1995, and three measures of socioeconomic status.39

In Figure 2, the states’ scores on our measure of overall quality of standards are plotted against the average of the percentage of public school students at or above basic levels on the reading and math assessments. A negative relation emerges strongly. States that in the mid-1990s had weaker student
performance tend to have implemented more fully articulated systems of content standards, assessment, and accountability by 1999. Thus the large variations in state standards to some extent reflect greater efforts by states with lagging test scores to use standards to reform the existing educational system. This is likely to engender greater equality in student outcomes across states.

In figure 3, the extent to which each state had implemented standards by 1999 is plotted against the natural log of population in 1995. Here a strong positive relation is apparent. As predicted, larger states have gone further in implementing content standards, assessment, and accountability.

The correlation coefficients for the relationships depicted in figures 2 and 3 and for more disaggregated relationships are presented in table 3. The table gives the correlations between the three components of our overall measure of standards, as well as their composite, versus the individual measures of student achievement in reading and math, the average of these measures of achievement used in figure 2, and the natural log of population. In all cases, the standards measures are related to achievement and population in the same direction as indicated above, although the strength of the relation varies. Initial student achievement and population in the state do not determine all of

---

*Figure 2. Quality of State Standards in 1999 versus Average Percent of Fourth-Grade Students at or above Basic Level, 1996 Math NAEP and 1994 Reading NAEP*

Note: NAEP = National Assessment of Educational Progress.
the variation across states in the standards that they have set, but these variables do seem to matter in important ways.

The data in table 3 also depict the correlation between the individual and overall measure of standards with three measures of socioeconomic status: the percentage of the population that is white (non-Hispanic), the percentage of adults age twenty-five and higher who hold at least a high school diploma, and the percentage of the population living above the poverty line. These measures of socioeconomic status are weakly negatively related to the quality of the states’ educational standards. That is, states with a greater proportion of disadvantaged residents have set slightly higher standards on average. This finding should come as good news. It suggests that decentralized (state-level) standard-setting (versus nationally mandated standards) might over time lower inequality in educational outcomes across the country. The level of standards is more strongly related to initial student achievement than to the three measures of socioeconomic status. Low student achievement rather than socioeconomic disadvantage seems to have been the more important factor driving the move to higher standards.

We have documented a rise in courses required for graduation in many states in the 1990s, a rapid expansion of state content standards, assessments linked to these standards, and student accountability and incentives in the form of exit exams and grade promotion and retention policies. A trend toward tougher
Educational standards and accountability is sweeping the country, even though some states lag behind. States in which student performance on the NAEP was relatively low in the middle of the 1990s tend to have done more to implement content standards, testing, and accountability since that time. Similarly, larger states and states with relatively disadvantaged populations tend to have made more progress.

### The Evidence on Effects of Educational Standards

How will the new educational standards affect student achievement? The literature on what happens to student outcomes under different sets of academic standards is small but growing.

#### Graduation Requirements

Given that all the published theoretical models agree that a rise in educational standards must, other things being equal, cause fewer students to meet the standard, it makes sense to begin by examining how many students lose from higher standards in this way. Dean R. Lillard and Philip P. DeCicca

---
**Table 3. Correlation Coefficients between Measures of Quality of State Standards in 1999, and Measures of Student Achievement in 1994 and 1996 and State Population**

<table>
<thead>
<tr>
<th>Measure</th>
<th>Content standards</th>
<th>Assessments</th>
<th>Accountability</th>
<th>Overall standards</th>
</tr>
</thead>
<tbody>
<tr>
<td>Math percent at basic level, 1996</td>
<td>-0.31</td>
<td>-0.22</td>
<td>-0.49</td>
<td>-0.46</td>
</tr>
<tr>
<td>Reading percent at basic level, 1994</td>
<td>-0.32</td>
<td>-0.19</td>
<td>-0.55</td>
<td>-0.47</td>
</tr>
<tr>
<td>Average percent at basic level</td>
<td>-0.32</td>
<td>-0.19</td>
<td>-0.54</td>
<td>-0.47</td>
</tr>
<tr>
<td>Natural log population</td>
<td>0.28</td>
<td>0.27</td>
<td>0.31</td>
<td>0.42</td>
</tr>
<tr>
<td>Percent population white non-Hispanic, 1997</td>
<td>-0.16</td>
<td>-0.04</td>
<td>-0.49</td>
<td>-0.31</td>
</tr>
<tr>
<td>Percent with high school diploma or higher, age 25 and above</td>
<td>-0.19</td>
<td>-0.32</td>
<td>-0.22</td>
<td>-0.37</td>
</tr>
<tr>
<td>Percent of population above poverty level</td>
<td>-0.20</td>
<td>-0.12</td>
<td>-0.21</td>
<td>-0.25</td>
</tr>
</tbody>
</table>

Source: Authors’ calculations based on American Federation of Teachers data on standards, National Assessment of Educational Progress test scores, and Bureau of the Census demographic estimates.

Note: In a small number of cases, only one test score was available, so the average percent of students at or above basic levels was set using the one available test score.
compare high school dropout rates and attrition rates among states in 1980 and 1990, and individual-level data from about the same times. Overall, the authors conclude, a one standard deviation increase in graduation standards, which corresponds to an additional 2.5 courses, is correlated with a 0.3 to 1.6 percent rise in the share of high school students who drop out. The basic finding that past increases in graduation requirements have led graduation rates to be lower than they otherwise would be meshes with theoretical predictions and needs to be taken seriously. Policymakers will require much more detailed information on what measures, if any, were targeted toward students who were at risk of dropping out as a result of the move to more rigorous standards. Policymakers will also want to know why some students appear to have been induced to drop out, as well as what alternative credentials and career paths might reasonably be made available to those students who will drop out in any event. (The companion paper in this volume by John H. Bishop, Ferran Mane, Michael Bishop, and Joan Moriarty provides a more detailed summary of existing work as well as extensive new findings on this important issue.)

Homework and Grading Standards

A number of papers that do not explicitly address the impact of changing standards over time nonetheless provide relevant insights. These papers consider the impact of variations in homework and grading standards.

Some papers examine the correlation between homework and test scores. Harris Cooper provides a detailed review of earlier research on the link between homework and student achievement. He cites numerous experiments, some but not all of which suggest a positive link. However, the sample sizes in these studies are very small (thirty-nine to four hundred students) and the studies examined only one to eight schools each. A larger literature examines the correlation between achievement and time spent on homework in a nonexperimental, cross-sectional framework. For instance, Cooper reports the results of eleven studies that model student achievement as a function of homework while controlling for background variables. Most of the eleven studies indicate a positive link between homework and achievement. But in some cases the research used small samples that are not nationally representative. In other cases, researchers used national samples but did not control well for prior achievement, thus increasing the risk of omitted variable bias. However, two notable exceptions are research done by Timothy Z. Keith and others and by
Herbert J. Walberg, Barry J. Fraser, and Wayne W. Welch, who use High School and Beyond and the National Assessment in Science, respectively, to establish a correlation between student test scores and the amount of homework that the student reported doing per time period, while controlling well for prior achievement and characteristics of the school environment.43

Unfortunately, these studies, like the vast majority of the literature, use a student report on hours of homework done per week. This is not a policy variable that a school administrator or teacher can directly control. In particular, much of the variation in homework performed by students might reflect unmeasured differences in student ability or attitudes. Another typical problem in the literature is that achievement in a given subject is regressed on homework performed in all subjects. The ideal measure of homework would be the amount of homework assigned by the student’s teacher in the given subject.

Julian R. Betts attempts to get around these problems by analyzing a nationally representative sample of students attending grades seven through twelve.44 Because teachers indicate the amount of homework they assign per week, chances are reduced that the analysis merely picks up more highly achieving and more highly motivated students choosing to do more homework. The results, for models of math test scores, are very strong, indicating that math homework is a more important determinant of gains in achievement than any of the standard measures of school quality, such as teacher education and experience or class size. The results are robust to the addition of a dummy variable for each student to control for omitted ability or motivation among students.

Betts also addresses the questions of how much homework is too much and whether only the best students respond to additional homework. Homework assignments ranged from zero to roughly eight hours per week. Within this range, no trailing off of the effectiveness of math homework emerged. (This study, focused on math homework, cannot indicate the optimal amount of homework that schools should assign across all subjects.) Betts also concludes that additional math homework appears to be equally effective in increasing the rate of learning across all students, regardless of their initial level of achievement. This is an important finding, given that one of the chief criticisms of higher standards and higher expectations has been that some students will respond by simply giving up.

Another Betts study examines variations across schools in math and science grading standards.45 It estimates the stringency of grading standards in each school by comparing test scores in these two subjects with grades in math and science courses, while controlling for the type of course taken, student
demographics, and school resources such as class size and teacher preparation. In the second stage, the analysis tests whether students learn more quickly if they attend schools with more stringent grading standards. The answer appears to be yes. However, in this case, unlike the case of homework, a policy of higher grading standards might help all students, but it seems to help most those near the top, increasing inequality in the distribution of student achievement.

**Grade Retention and Summer School**

Our theoretical analysis focused on a pass-fail standard in which there are repercussions for students who do not fulfill the academic requirements established by the educational standards. An increasingly common implementation of this idea calls for students to repeat a grade if they lag too far behind established standards for the students’ grade level. Grade retention differs from the theoretical analysis in that students receive a second chance to meet the standard. Another variant requires students who do poorly on achievement tests to attend classes after school, on weekends, or in summer school. These approaches provide additional resources to the students most in need.

The impact of grade retention has received considerable attention. In a review of the literature, Thomas Holmes reports that grade retention is typically associated with poorer student performance after the student is held back a year. Only nine of sixty-three studies found that retention improved the students’ performance. Holmes indicates that in most of these nine studies, the treatment of students was not simply retention but retention accompanied by intensive remediation. Additional attention to the students who lag furthest behind is likely to be necessary in a system that sets strict content standards.

Summer school for students who have fallen well behind grade level seems to offer an alternative, and perhaps less stigmatizing, option. The Chicago public school system has received national attention for a bold program called Summer Bridge. As reported by Betts, beginning in the 1996–97 school year, students in grades three, six, eight, and nine whose performance lagged behind national norms on either the reading or mathematics portion of the tests were required to attend summer school. The cutoff points below which students were required to attend summer school were 2.8 for grade three, 5.2 for grade six, 6.8 for grade eight, and 7.9 for grade nine. (The tests were given in spring, so that a student progressing at the normal rate should have attained a grade equivalent of about 3.8 by May of the third-grade school year.)
summer school, students were tested again and were promoted to the next grade if they then met the standard. Betts calculates that, in the initial testing, 27.1–62.2 percent of students failed at least one of the two tests, depending on the grade level. Unfortunately, not all students who should have attended summer school did so. But when calculated as a percentage of those who took the summer tests, the success rate at the end of summer ranged from 38.4 percent to 49.6 percent, with the highest success rate among eighth-grade students.

The first-year evidence suggests that the summer school program provided an extremely cost-effective way of improving student performance. The mean increase in students’ grade equivalent during summer school varied by grade from about one half to a full year. These increases hint at large incentive effects on the students and their teachers. But important questions remain. If the Summer Bridge program merely drilled students on testing techniques, then much of the gains over the summer should disappear during the following school year. Further, improvement over the summer might in part represent regression to the mean after some students on the spring test had an off day. A longitudinal analysis should be able to provide direct information on some of these issues, including whether the creation of high-stakes tests increased student effort.

Melissa Roderick and others present the results of a two-year study of Chicago students. Among the important findings:

— Students who attended Summer Bridge in the summer of 1997 retained most of their large achievement gains. However, their rate of improvement during the 1997–98 school year was much smaller than for other students, so that part of the achievement gap reemerged during the 1997–98 school year.

— To test for the incentive effects, the authors compared scores for students in spring and summer 1997, during the first year of the program, with scores of students in spring 1995, before the new summer school and grade retention policy was in place. Gains in grade three were fairly muted. However, the percentage of students making the grade cutoffs during spring testing increased considerably between 1995 and 1997 in grades six and eight. The largest gains accrued to students who were particularly far behind at the start of the school year.

This latter finding suggests that the imposition of new standards and accountability led to significant increases in student effort, teacher effort, or both, at least in grades six and eight. Results for the reading test in grade six—the percentage of students in various categories who met the reading cutoff at stated times—are listed in table 4. Students were divided into groups based on how
many grade equivalents they would need to gain during grade six to reach the stipulated cutoff. The table has the results for students who needed to gain at least some positive fraction of a grade equivalent by May of their year in grade six to be promoted to grade seven. The first column of numbers shows the percentage of students making the cutoff in spring 1995. These students provide a benchmark case because the Summer Bridge and promotion policy were not yet in place. The second column shows the percentage of students making the cutoff in May 1997, the first year of the new policy. The third column combines this percentage of students who met the cutoff in May 1997 with those who failed in May but met the cutoff during a second test after participating in Summer Bridge.

The data in the table show a marked increase in the percentage of students making the cutoff in May 1997 relative to May 1995, with the largest gains among the students who were initially furthest behind. For example, among students who needed to improve their test scores by more than 1.5 grade equivalents, only 20 percent met the cutoff by May of the following year in 1995, compared with 31 percent in 1997. Because these two groups of students had similar initial achievement, the 11 percent gain suggests that the replacement of social promotion with strict grade promotion policy in the 1996–97 school year induced strong incentive effects. Weaker incentive effects are apparent among students whose initial grade equivalents were higher, as shown in the table.

The data in Table 4 also make clear that summer school for at-risk students led to major gains in achievement. Roderick and others report that these impressive gains persisted in the second year, but Summer Bridge did not lead to greater rates of learning for these children during the subsequent school year, so that part of the achievement gap reemerged over time.

Table 4. Percentage of Sixth-Grade Students Meeting Reading Test Score Cutoff in 1995 and 1997 in Chicago Public Schools by Number of Grade Equivalents Behind in Previous Year

<table>
<thead>
<tr>
<th>Initial number of grade equivalents behind</th>
<th>May 1995</th>
<th>May 1997</th>
<th>After summer bridge, August 1997</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt; 1.5</td>
<td>20</td>
<td>31</td>
<td>52</td>
</tr>
<tr>
<td>1.5 to 1</td>
<td>36</td>
<td>43</td>
<td>65</td>
</tr>
<tr>
<td>1 to 0.5</td>
<td>50</td>
<td>57</td>
<td>79</td>
</tr>
<tr>
<td>0 to 0.5</td>
<td>65</td>
<td>71</td>
<td>88</td>
</tr>
</tbody>
</table>

Source: Melissa Roderick and others, Ending Social Promotion: Results from the First Two Years (Chicago, Ill.: Consortium on Chicago School Research, 1999).
Whether the apparent incentive effects derive from greater effort among students, teachers, or parents of at-risk children, or all three, is not certain. In addition, as Roderick and others note, the simple comparison they make across two cohorts cannot establish whether the new grade promotion policy or some other unobserved change in the Chicago schools was the main cause.

Still, the results provide indirect evidence in favor of strong incentive effects related to the raising of standards, as posited in our theoretical analysis. Four groups of students who are at risk of failing should be considered. In order of increasing achievement, these groups are: those at the very bottom who exerted no effort with or without the new standard, slightly more highly achieving students who reduce their effort after the standard is raised because they believe that they cannot meet the new cutoff, students who do not change their effort and fail under the new system, and students who work harder after the standard is raised. (At the very top are top-achieving students who can easily meet the new cutoff without increasing effort.) Our main concern is the size of the bottom three groups compared with the fourth group, which increases its achievement. The Chicago results summarized by Roderick and others yield no trace of the bottom three groups of students who either do not change their effort or reduce it. Students who had to improve by more than 1.5 grade equivalents showed the strongest improvement relative to similarly weak achievers who entered grade six before the standard was raised.

Caution must be exercised in inferring the cause of the large achievement gains observed in Chicago. But the finding that higher standards help the lowest achieving students the most is potentially of great importance. It also squares well with the finding by Betts that additional math homework has strong positive effects on the achievement of all students, regardless of their initial level of achievement.

The Case of Massachusetts

The 1993 Massachusetts Education Reform Act (MERA) established two prongs in a seven- to ten-year plan. The first prong, in response to a state court ruling in a district finance adequacy case, established a seven-year schedule for a massive rise in state aid to bring all localities up to a newly formulated foundation budget by 2000. Real state aid more than doubled over this period. The annual growth rate of state aid in current dollars averaged 12.4 percent, exceeding inflation plus enrollment growth by 7.7 percent.
As a result, all districts were successfully brought up to foundation budget, and the gaps in spending were markedly narrowed. Low-income districts (bottom quartile) now spend more per pupil than middle-income districts (middle two quartiles). At the same time, even the higher spending communities received some increase in state aid, over and above inflation. Per pupil spending in districts at the tenth percentile (that is, low spending districts) rose $862 (in 1999 dollars) from 1993 to 1998, and by $449 at the ninetieth percentile, because of a combination of local and state funding. This achievement of raising all districts to foundation budget is widely viewed as remarkable, thanks to the surprisingly robust growth of the economy and the bipartisan commitment to education reform.

The other prong of MERA was standards-based reform. The law stipulated the development of state curriculum frameworks, to be followed by aligned assessments, which would be administered for a few years before triggering consequences. Accountability would first apply to school officials, through a school accountability program, and finally to students. MERA stipulated that a Massachusetts diploma would become contingent on demonstrating tenth-grade proficiency in the core subjects.

Both prongs of MERA were essential to the broad, bipartisan consensus among the Democratic legislature, Republican governor, the press, and the public, in an otherwise politicized state. Of note is that the money came first, while the accountability measures were being developed, and the consequences of the standards were scheduled to be the last step. The wisdom of this approach (facilitated by good economic times) is that it not only provided the wither- withal to localities, but also strengthened the backbone of public officials for phase two: They are now committed to follow through on accountability measures to justify the massive increase in funding that has taken place over the previous seven years.

The MCAS Exams

The curriculum frameworks took longer to develop than originally scheduled, in part because of changes in leadership of the Massachusetts Board of Education. Some of the more contentious frameworks, notably history and social science, went through many twists and turns before being adopted. This delayed the development of some of the exams in the Massachusetts Comprehensive Assessment System (MCAS), given that they are specifically aligned with the state frameworks. Unlike some states, which have purchased
off-the-shelf tests, Massachusetts spent the time and money to develop its own exams.

The first exams were administered in the spring of 1998 to students in grades four, eight, and ten, without high stakes attached to them. In the fall of 1999, the Massachusetts Board of Education voted to go ahead with the scheduled graduation requirement for the class of 2003, ten years after the enactment of MERA, but on a temporarily more limited basis than was originally envisioned. Instead of requiring students to pass exams in all the core subjects, only math and English language arts (ELA) would initially be required. The board also voted to set the initial cutoff for graduation on these exams at the bottom of the Needs Improvement category, instead of the originally intended cutoff at Proficient, because the initial tenth-grade failure rates exceeded 50 percent.55 Students will have at least four opportunities to retake the tests before the end of twelfth grade.

Both math and ELA exams include sizable open-response and essay sections, in addition to multiple-choice questions. Specifically, the ELA exams for each of the three grades include two sessions for a long composition (one for drafting and one for revising, as well as extra time granted upon request), four open-response questions, and thirty-two multiple choice. The spring 1999 fourth- and tenth-grade compositions were as follows:

“Some days are more fun than others. Describe a day that was great for you and tell WHY it was great. Include details so the reader can enjoy the day as much as you did.”

“In literature, as in life, things are not always as they appear to be. Identify a work of literature that you have read in or out of class in which this is true. Select one event, scene, or episode from this work of literature and explain in an essay what the situation appears to be and what the situation really is.”

The grading standards for passing performance on such essay questions are not overly demanding, to judge by the examples of student essays released by the Department of Education (DOE).56 Essay exams are graded by teachers in a summer program that converts many initial skeptics into true believers, according to the DOE.

Each year all of the questions that student scores are based on are publicly released, and they are not used again. This greatly reduces the problem of artificial test-inflation compared with exams in which the questions on existing forms become more widely known over time.57 This raises the cost of testing, but at about $15 a head, it is still cheaper than AP and SAT exams.
Early Test Results

The 1998 and 1999 failure rates were high on math in grade eight (over 40 percent) and grade ten (over 50 percent), as well as on ELA in grade ten (about 30 percent). The failure rates are much higher in most of the urban districts (over 75 percent in Boston and over 80 percent in Springfield). Moreover, the tenth-grade scores did not improve in the second year of the test. Two math examples illustrate some of the range in level of performance.

From the 1998 eighth-grade test:

According to the 1990 census, the population of Massachusetts was 6,016,425. Approximately what percent of those people lived in Boston?

<table>
<thead>
<tr>
<th>City</th>
<th>Population</th>
</tr>
</thead>
<tbody>
<tr>
<td>Boston</td>
<td>574,283</td>
</tr>
<tr>
<td>Cambridge</td>
<td>95,802</td>
</tr>
<tr>
<td>Fall River</td>
<td>92,703</td>
</tr>
</tbody>
</table>

A. 10%
B. 20%
C. 30%
D. 40%

Only 28 percent of Massachusetts’s eighth-graders answered correctly, barely more than the 24 percent that would obtain if those who answered the question guessed randomly. This was a particularly low scoring question, but performance on the following question was slightly better than most.

From the 1999 tenth-grade test:

Which of the following functions will yield the largest value for \( x = 50 \)?

A. \( f(x) = 5 + x \)
B. \( f(x) = 5x \)
C. \( f(x) = x^2 \)
D. \( f(x) = 5^x \)

Students were allowed to use calculators during this part of the exam, but still only 52 percent got it right. Other questions were harder, primarily because they demand that students know how to apply mathematical concepts, including multistep problems.

Some factors contributing to the high failure rates have been identified in a study for Mass Insight Education, which examined records of a sample of urban and nonurban students who failed one or both tenth-grade exams.
Approximately one-fourth of these students were absent more than five weeks of the school year. Many of these students, clearly disengaged, are likely to become dropouts independent of the MCAS. It seems unlikely that MCAS would have negative incentive effects on such students once it starts to count and may have positive incentive effects for some, once students realize they will have to attend school to pass.

A number of students left entire sections of the exam blank, including 13–19 percent of the failing urban students in this sample who answered no multiple-choice questions at all, and 20–23 percent who left all the open-response questions blank. It seems reasonable to predict that a significant number of these students, and others as well, would behave differently once the test starts to count for graduation.60

Other factors that give some reason to believe the failure rates will drop once the exam starts to count include the fact that about 10 percent of the failing students in math came close to passing on the first try and will likely do so with multiple retake opportunities in grades eleven and twelve.61 Also, about 20 percent of the students who failed the math exam are special education students, some of whom will be eligible for test-taking accommodations or alternative examinations starting in 2001.

A quarter or more of these failing students were also failing the math or English course they were taking at the time. For the majority who were passing these courses, a big part of the problem is the level of the math course. Well over half of students failing the tenth-grade math exam were enrolled in remedial or basic math or algebra 1, so they have not been taught much of the tenth-grade material expected from them on this exam. The math exam is a much greater hurdle than the ELA exam, and a huge part of the challenge will be to get students completing algebra 1 by ninth grade at the latest.

In short, good reason exists to believe that the failure rates will be substantially lower once the exam starts to count, but they still threaten to be high on the math portion. Consequently, a full array of remedial measures is currently being implemented in a number of districts. As in other states, these include after-school, summer school, and in-school programs, to provide short-term help for students who have fallen behind.62

But deeper changes are also called for, reaching farther back in the curriculum, so that students will be ready in the normal course of study for the exams they will face. This is definitely happening, at an accelerated pace because of MCAS, according to many superintendents across the state. Widely
noted changes include greater emphasis on writing and on open-ended math problems. Scores on the fourth-grade MCAS exams have already shown improvement in the second year of testing.

*An Econometric Analysis of ELA-4 and ITBS-3 Scores*

In 1999, the second year of the MCAS, the mean score on ELA-4 rose approximately 4.0 percentiles, and the median score rose 4.5 percentiles over the scores of the previous cohort. The question arises as to how much of this improvement stemmed from a change in the quality of the cohort (a better group of students), as opposed to more fundamental change, in the amount of learning in grade four. Fortunately, the Massachusetts DOE has assembled a useful micro data set that allows one to answer this question for the ELA-4. The state required all school districts to administer the third-grade Iowa Test of Basic Skills (ITBS) reading test for the years 1997–99. The ITBS scores are far and away the best predictor of the following year’s MCAS scores. But the third graders in 1998 scored worse on the ITBS than their predecessors in 1997 and then, the next year, scored better than their predecessors on the MCAS. This suggests that the MCAS improvement was not the by-product of a higher quality cohort. The cohort effect worked in the opposite direction, masking an even larger MCAS improvement, apparently reflecting more fundamental change in fourth-grade learning.

More rigorous statistical analysis bears this out. The DOE has linked the third-grade reading scores with the fourth-grade MCAS ELA scores for over two-thirds of the state’s seventy-five thousand fourth graders, to validate the MCAS exam. The ITBS score accounts for 56 percent of the variance in individual MCAS scores a year later. We ran regressions with additional controls for race and gender, plus indicators for the nearly one thousand schools in the sample, for MCAS scores from 1998 and also from 1999. This allows us to decompose the mean gain in MCAS scores into that part which comes from changes in the explanatory variables (especially ITBS scores) and that part which reflects changes in the effects of those variables, the regression coefficients (especially the school effects). This decomposition (known as an Oaxaca decomposition) suggests that the adverse cohort effect (from lower ITBS scores) masked an underlying improvement in mean MCAS scores of over 5 percentiles (versus 4.0 in the raw data).

We take the analysis a few steps further, to shed some light on whether the improvement in MCAS scores represented a superficial test-specific improve-
ment or whether broader skill improvements were set in motion. We begin with a decomposition of changes in the ITBS scores, analogous to that of the MCAS. Controlling for race, gender, special education, limited-English proficient (LEP), and free-lunch status (but without a prior test score to control), we find that ITBS scores improved dramatically from 1998 to 1999, despite an adverse cohort effect. The underlying improvement in mean ITBS scores was more than eight percentiles, after correcting for the cohort effect.64

Was it a coincidence that third-grade ITBS scores rose dramatically the same year that fourth-grade MCAS scores rose by 5 percentiles? If both events reflect improved practices or curriculum, stimulated by the introduction of MCAS the year before, this would be a finding of great interest. It is impossible to test this hypothesis directly, but some suggestive circumstantial econometric evidence is available. Roughly speaking, schools that added more to 1999 student performance on their third-grade ITBS scores than would have been predicted based on how much the school added in previous years also tended to add more to their 1999 fourth-grade students’ MCAS scores than would have been predicted.65 This is consistent with, though it does not prove, the hypothesis that those schools that were stimulated most to action by the introduction of MCAS were likely to have made improvements in third-grade reading instruction as well as fourth-grade reading and writing. If so, this would indicate the positive effects of MCAS go beyond superficial test coaching to more pervasive improvements. These improvements seem to go back to earlier grades, providing the foundation on which to build.

**Controversy over MCAS**

In the third year of MCAS, controversy has escalated. Media attention has focused on student and teacher boycotts, even though the number of boycotters is small (about two hundred to three hundred students). Students are by tradition averse to exams, so the more important question is why some adults are encouraging them.66

Objections fall into several categories. The protestors (and groups such as FairTest and the American Civil Liberties Union) claim the test is unfair to disadvantaged students in low-income, poorly funded districts. But funding gaps have narrowed markedly, and the largest urban districts spend above the state average per pupil. The opposition is “mostly in the affluent suburbs west of Boston and in pockets of progressivism like Cambridge.”67 With a few exceptions (such as the local chapter of the National Association for the Advancement
of Colored People), representatives of the minority communities have largely targeted their anger at the failure of the school system to bring up the skills of their children, rather than at the MCAS. They already knew the general message MCAS was bearing.

A disproportionate number of the teacher opponents to MCAS comes from the history and social studies departments. They object to the MCAS history exam. It will not yet be required for graduation for 2003 but is being administered because MERA includes history in the core competencies. These teachers believe it narrows the scope of what they teach. One prominent and vocal group of opponents is employed by Facing History and Ourselves, a company that sells history curriculum to the schools (built around the Holocaust) and argues that its curriculum will be squeezed out by MCAS.

Some of the opposition in the higher achieving localities is based on the concern that the exam is too long and takes too much time from other activities. The state is responding to this concern by spreading out the testing over more grades, such that no student in grades one through seven will spend more than seven hours a year in MCAS testing, from 2001 on.

Another objection, common elsewhere as well, is to the idea that a student may be denied a diploma on the basis of a single test. However, MCAS is an extensive set of examinations, so that students who write strong essays or excel in open-response questions can offset poor performance on multiple-choice sections (or vice versa). The objection seems not so much directed to a single test, but to a set of external common assessments versus a set of local and possibly idiosyncratic criteria.

The Massachusetts Teachers Association (MTA), the state affiliate of the National Education Association (NEA), has also joined in opposition to the MCAS. The MTA recently announced its intention to file legislation to eliminate the MCAS graduation requirement. The MTA also began a $700,000 TV ad campaign explicitly designed to counter the perceived attack on public education by those who point to low MCAS scores.

What seems to be at issue here is that the MCAS is the key component in the accountability phase of Massachusetts’s education reform. The MTA is understandably threatened. Thus far, however, with few exceptions, the legislature and administration stand firm behind MCAS. Too much money has been spent over the last seven years to lightly abandon the insistence on results.

Meanwhile, in the school districts that face the highest failure rates, the most important story is unfolding:
Little of this [anti-MCAS] grumbling . . . is coming from the urban districts and poor communities that are the true targets—and primary beneficiaries—of education reform. In places filled with the neediest, low-income, immigrant and transient student populations, school leaders have, by and large, embraced the state’s regimen of standards and accountability. For districts that, prior to 1993, hadn’t been pushed to serve all students well or didn’t have the resources to do so, the $5.6 billion spent statewide has been a godsend. From Boston to Springfield, city school chiefs have latched on to standards-based reform not only as a quid-pro-quo for the new dough, but as their preferred vehicle for improving instruction.70

The ways in which school chiefs are using MCAS to improve instruction go beyond changes in curriculum and remedial programs to more general leverage (the term commonly used by superintendents) over those teachers and administrators who resist changes such as the reorganization of the school day, revamped professional development, and so on.71

One of the most striking instances of this leverage arises in the hard bargaining stance taken in the spring and summer of 2000 by the Boston School Department over the issue of seniority. As is commonly the case, the union contract (of the AFT affiliate) grants senior teachers first refusal of new jobs and the right to apply for jobs held by new teachers. In an unusual development, a broad coalition of about thirty parent and community groups, such as the Urban League and the Black Ministerial Alliance, joined together to side with school officials in limiting seniority rules. As the Boston Globe reports, “Parents say the drumbeat of reform—from stiffer curriculum standards to a standardized test as a graduation requirement—underscores the importance of this year’s negotiation.”72 One cannot help but note the contrast between the Boston parent groups whose response to standards-based reforms is to challenge problematic union rules, while efforts to derail the standards are largely confined to the more affluent and progressive districts, along with the state NEA affiliate.

**Obstacles to Strengthening Educational Standards**

Based on our knowledge of reform efforts in California, Massachusetts, and other states, and the theoretical and empirical research on standards, we identify four key obstacles that can stand in the way of higher educational standards: opposition arising from concerns about the distribution of student achievement, problems in defining standards and assessing students’ progress...
toward those standards, the need to align the incentives of all participants in public education, and equity concerns created by the large gap in school resources that currently exists among students from various socioeconomic groups in some states.

Opposition to Standards Based on Distribution of Student Achievement

Opposition to higher educational standards can arise for many reasons, but in our judgment the source of opposition that resonates most strongly (if not always most convincingly) derives from concerns about equity. According to our theoretical analysis, any change in standards typically leaves some students worse off. This makes the politics of higher standards inherently divisive. Legislators in most states have determined that a movement toward higher educational standards is worth the effort. However, as parents become more fully aware of the gap between published standards and the performance of their children, opposition could swell.

Many parents and legislators might be surprised to learn just how much variation there is in student performance at present. The twenty-fifth through seventy-fifth percentiles and the minimum and maximum in student performance on a standardized math test by grade level, in the Longitudinal Study of American Youth (LSAY), are depicted in figure 4. The LSAY sampled a representative population of American school students between 1987 and 1992. Particularly striking is how large the variation in achievement is within grades, compared with the average rate of improvement between grades. Betts uses these data to calculate the percentage of students who would be held back a year if the school’s policy were to retain students whose test scores were below the national average for students one or two grade levels below the student’s current grade.73 In other words, what percentage of ninth-grade students would be held back if their math scores were below the national average for students in grade eight or even grade seven? The predicted percentage of students who would be held back if their achievement lagged by a year ranged from 37 to 46 percent, depending on grade, in grades eight to twelve. If, instead, students were retained only if their scores lagged national norms by two years, then 26–40 percent of students would have been retained. These are very large shares of the student population.

The estimates are an upper bound in the sense that if strict grade promotion policies based on test scores were implemented, it would provide an incentive for students to study harder and for schools to reform curriculum
Figure 4. The Distribution of Test Scores by Grade, 1987–92

Math scores

Grade

and teaching practices. Evidence from the Chicago public schools suggests that the development of standards, testing, and accountability can spur much greater effort among students at risk of failing. Nonetheless, early experiments with grade promotion linked to test scores suggest that these discouraging numbers are not outlandish.74

Given the large variations in student achievement at present, what policies might reduce the chance that political opposition will overturn recent moves to institute standards? One solution might be to devote additional attention to marginal students including those who are most likely to give up after standards are raised, in a bid to ensure that no student’s achievement falls after standards are raised. The Summer Bridge program in the Chicago public schools represents one example of an effort to supplement higher standards with programs aimed specifically at the students most in need.

However, opposition to standards appears to come not typically from families whose students are most likely to fail when the standards are raised, but from families in areas served by good schools. (In Massachusetts and Wisconsin, at least, the most vocal opposition to tighter standards has come from affluent communities.) Parents in more successful schools may fear that districts will shift resources from their schools to underperforming schools in the district. Parents’ fear that administrators will reduce funding at top schools is a legitimate one, especially in systems with large heterogeneous districts. The only evident solution is to expand total funding so that no school suffers a reduction in programs, while at the same time the schools most in need receive additional resources. Thus, implementing higher standards makes sense at a time when state budgets make higher funding a real possibility. Massachusetts appears to have followed this policy prescription closely.

Some affluent parents might worry that higher standards will make standing out from the crowd more difficult when their children apply to university. Such concerns become potentially relevant when a state imposes a single standard, but the existence of other high-end credentials (AP exams, SATs, and so on) renders this concern less compelling. Further, if the existing array of credentials is insufficient to differentiate high-end performance, the state can create a range of high-end standards, which in turn create incentives for a wider range of students to excel. If a multitiered set of standards induces almost all students to work at least as hard as they had without the standards, and if the minimum standard is set to ensure that even the weakest students leave school with a good set of basic skills, a multitiered set of standards makes good sense. It provides incentives for a wider range of students than the group of students
near the margin under a simple pass-fail standard, while providing top students with a means to signal their high effort levels to universities and employers. Many states have taken this lesson to heart, creating differentiated advanced diplomas for students who meet strict standards.

Problems in Defining Content Standards and Assessing Student Achievement

Implementation of content standards and assessment of student progress have often proven difficult. The design of content standards has been contentious in many states. Perhaps this is best seen in the history of the movement for national content standards in public schools. In brief, the National Council of Teachers of Mathematics (NCTM) developed national math standards during the 1990s. These standards have provided an influential framework for individual states as they have striven to develop their own standards in math. However, certain elements of these standards have elicited objections from parents and many prominent mathematicians. Similarly, when California first attempted to develop science standards, two rival groups, one led by Nobel Prize–winning scientist Glenn Seaborg, and a second led by educators from state schools of education, clashed. In the end, the state urged the two sides to come together, with some success.

The care and attention to detail that are required to develop a set of content standards suggests that for reasons of cost, it probably makes no sense for individual schools or smaller districts to write their own set of standards. But given the limited success of the movement to create nationally adopted standards, the states will continue to play a paramount role in standard-setting.

Similarly, several problems arise in the creation of tests. First, most commercially available tests may be related only weakly to the given state’s curriculum standards. It will take time for all states to develop more suitable test instruments. For example, California adopted the Stanford 9 test for use in spring 1998 and is now moving this off-the-shelf test toward the new state content standards by adding several components.

Second, preparing tests that provide both in-depth and sufficiently wide coverage of a subject creates challenges. The solution would appear to be to lengthen existing test instruments so that they provide an in-depth coverage of a wide area within a subject. Essay and open-response questions, of the sort used in the MCAS test in Massachusetts, represent a step in the right direction in that they gauge students’ level of mastery of written expression and
problem solving that no pure multiple-choice exam could approach. However, broadening the test then evokes the complaint that it is too long, diverting student time from other learning activities. Often the same critics who object to a single test being used for high stakes also object to the length of a multifaceted set of exams, indicating that the objections are not being accurately framed. The external nature of the assessments is really at issue.

A third problem can arise from the natural tendency of teachers to teach to the test. This is compounded by the fact that, in many cases, the same form of the test instrument is given several years in a row, so that teachers, and perhaps students, become familiar with the specific questions over time. This can lead to inflation of test scores without accompanying gains in true student achievement. Daniel Koretz summarizes earlier work he conducted with coauthors in which a school district had introduced a new test form in 1987, only to find a significant drop in the average grade equivalent of students on the test. Over the next three years, however, successive cohorts of students improved in this test, to the point where students were performing at about the same level as students had the year before the switch to the current form. Two questions arise: First, did the large drop in test scores in 1987, the year that the new form was introduced, represent a true drop in achievement? Second, did the steady improvement over the next three years that the same form was used represent true gains in performance of students, or merely teaching to the test as teachers became better acquainted with the new questions? To test the latter hypothesis, Koretz and coauthors arranged to test students in the district during 1990 using the same test form that had last been used four years earlier, in 1986. Their findings suggest that the large drop in achievement in 1987 and the subsequent gains reflect the switch to a new test form and then teaching to the test on the new form. Little change in true achievement occurred.

There seem to be two solutions to this problem. First, annual changes in the test form should reduce gains in test scores that result from teaching to the test. This may raise the cost of testing but seems worth the price if policymakers and parents want a reliable indicator of trends in student achievement. Second, it seems inevitable that teachers will teach to the test, especially if schools and teachers are held accountable for student performance. This tendency can be transformed from a vice into a virtue as good tests that accurately and fairly test the students’ knowledge of the given content standards are developed. With the creation of excellent tests, teaching to the test should eventually become a good thing.
Creating Incentives for Students, Teachers, and Administrators

Many states now hold students accountable for performance, through policies of grade retention, summer school, and exit exams. However, most states lag behind considerably in creating incentives for teachers and school administrators to work toward student success in mastering content standards.

California’s Public School Accountability Act of 1999 provides one example of the limited incentives that states have put in place to date. California schools that lag furthest behind in the Academic Performance Index (a non-linear average of student achievement) are eligible to participate in the Immediate Intervention/Underperforming Schools Program (II/USP). Initially, schools in this program receive money to speed improvement in student achievement. However, any school that does not meet its growth target must hold a public hearing and is subject to intervention by the local district board. If, after two years, the school still shows few signs of improvement, then the state superintendent can take over the school. The principal can be reassigned. In addition, the state superintendent can take a number of other actions, including allowing parents to send their children to other schools or to create a charter school, reassigning certified administrators or teachers, or even closing the school. The threat that a principal could be removed from a school creates incentives for the principal to improve student achievement quickly. As the legislation behind these accountability measures was passed only in 1999, it will take some time to observe how often and how effectively the measures come into play in California.

The II/USP program and similar programs in other states create incentives for teachers and principals, but they seem weak compared with the incentives already facing students, such as the threat of grade retention. For instance, outright firing of teachers or principals seems unlikely given the collective bargaining agreements that typically apply. Similarly, large merit bonuses for teachers, in groups or individually, to reflect gains in student achievement are by no means a widespread phenomenon. Merit pay for teachers has been attempted many times in the past. But as Richard J. Murnane and others show, such programs have typically collapsed because of legitimate teacher concerns that principals were setting merit pay based on unverifiable information, opening up the possibility of cronyism. One reason for hope in this regard is that current attempts to improve student assessment might provide mutually agreeable and objective ways of gauging the overall performance of teachers in a school or the performance of individual teachers. A number of states, and per-
haps most notably the city of Denver, are beginning to experiment with rewards for teachers based on the rate of progress of their students.81

Much remains to be done to increase the incentives of all participants in public education, especially teachers, principals, and administrators, to work toward fulfillment of content standards by all students.

Gaps in School Spending and Opportunity-to-Learn Standards

Inequities in school spending among districts can threaten to derail the movement to impose uniform educational standards. During the 1990s, a movement for what became known as opportunity-to-learn standards argued forcefully for equalization of school spending before implementing student accountability.82

The call to partly or fully level the playing field in terms of school spending before holding all schools equally accountable makes sense and is sometimes required to meet a state constitutional provision for adequacy or equity.83 But the public should not overestimate the achievement disparities that are attributable to existing inequalities in school finance per se. The reason is simple: Existing research suggests that school resources such as class size, and to a lesser extent teacher education and experience, have fairly limited effects on student achievement.84 Similarly, the link between school resources and longer-term measures of student outcomes, such as educational attainment and wages, is modest.85

Consider, for example, Julian R. Betts, Kim Rueben, and Anne Danenberg, who analyze the distribution of school resources and test scores on a school-by-school basis in California.86 The authors find strong inequalities in teacher preparation among schools (even within the same district), with lower socio-economic status students receiving teachers who are considerably less well prepared, whether measured by teacher certification, experience, or education. (SES is measured by the percentage of students receiving full or partial lunch assistance.) For example, in elementary schools in California, in the lowest SES quintile of schools, on average 32.6 percent of teachers hold no more education than a bachelor’s degree, compared with only 8.8 percent in the highest SES quintile of schools. Low SES schools also have much lower test scores, raising the question of whether low achievement in these schools is caused by a lack of resources or by the direct effects of poverty.

Regression analysis suggests that school resources do affect achievement, but the effects are small. The predicted effects on the percentage of students
scoring at or above national norms in reading when a school moves from the twenty-fifth to the fiftieth and then the seventy-fifth percentile in a number of school resources are represented in figure 5. All variables in the figure except for class size have a statistically significant impact on student achievement. But the figure demonstrates that variations in poverty can account for a far higher share of variations in student performance than can variations in school resources, in spite of the large variations in teacher resources that currently exist in California.

Thus, equalization of resources among all schools might reduce inequalities in student outcomes, but only modestly. Looking at the data another way, existing inequalities in resources bear only a small part of the blame for variations in achievement in California. The creation of uniform educational standards might provide the incentive to improve student performance in a way that spending hikes alone cannot. The results on the Summer Bridge program in Chicago imply that reasonably small interventions such as several weeks of summer school can bring impressive and lasting improvements in student performance. The lesson from Chicago seems to be that higher standards, accompanied by judicious new expenditures aimed at the truly needy students, can produce meaningful gains in achievement.

Figure 5. Predicted Effect of Changing School Characteristics on the Percentage of California Fifth-Grade Students Scoring at or above National Median in Reading Test, Spring 1998

Source: Julian R. Betts, Kim Rueben, and Anne Danenberg, *Equal Resources, Equal Outcome? The Distribution of School Resources and Student Achievement in California* (San Francisco, Calif.: Public Policy Institute of California, 2000).
A similar finding emerges from analysis of the effects of grade retention. Grade retention appears to work only when schools try to do something different, possibly with additional resources, for students as they attempt to complete a grade for a second time.

States that reduce historical inequalities in school spending before creating content standards reduce the risk of political opposition based on opportunity-to-learn lines. States that implement rigorous standards while targeting programs of demonstrated effectiveness to the students most at risk do even better.

**Conclusions and Policy Implications**

Our theoretical and empirical analysis and review of standards in practice suggests a number of conclusions and policy implications:

— Standards and accountability systems do affect incentives of students, parents, and schools. Limited, but growing, empirical evidence establishes that significant numbers of students rise to greater levels of achievement than when little was expected of them and their schools.

— Assessments should be aligned to standards; they should include open-ended questions and essays worth teaching to; and new forms should be introduced annually to avoid artificial inflation of test scores.

— Localities should retain the option to set higher standards than those set by the state.

— School financing systems should meet state constitutional requirements for adequacy or equity across districts before high-stakes standards take hold (as in Massachusetts).

— Judicious additional spending targeted at students who are likely to fail to reach standards without help makes sense. For example, programs of demonstrated effectiveness, such as Chicago’s mandatory summer school at early grade levels for those who fail to meet standards, should be replicated.

— Incentives should be strengthened for schools, especially school leaders, to ensure that students meet standards. Examples include reconstituting failing schools, reassigning teachers and administrators in these schools, and providing sanctuary for students from these schools in other schools or in new charter schools.

— Potentially harsh trade-offs can be minimized by multiple credentials, signaling different levels of achievement. Such signals already exist for high
levels of achievement. At the other end, for those students who cannot be remedi-
ated to reach stipulated levels of cognitive skills, credentials need to be
developed to signal important noncognitive skills. These credentials, such as
certificates of completion, should be sufficiently differentiated from cognitive
credentials to maintain the incentive to acquire cognitive skills.

No such list of recommendations can fully anticipate what will work and
what will not work as full-blown standards-based reform takes effect. Not
everyone will meet the new standards, just as not everyone met old standards
in the past, before social promotion became the norm. New answers will
evolve to the question of what shall be done for those who fail to meet the
new standards. In the past, the GED arose to meet the needs of those who
wished to convey some level of cognitive achievement without attending
school through grade twelve. For others, alternative settings will be devel-
oped, such as the ninth-grade remedial schools in Chicago. Proposals have
been made in Massachusetts for the community colleges to admit students
into special nondegree remedial programs, for those who fail the MCAS but
receive a certificate of school completion. After-school programs analogous
to the Japanese jukus will also arise, whether by public or private initiative.

Although the optimal configuration of credentials is not yet known, one
thing is certain: It would be a disservice to all too many high school gradu-
ates to continue granting diplomas that provide no guarantee of minimal
literacy and numeracy skills. Amid all the rising controversy, it is a remark-
able fact that not even the most vocal critics of standards-based reform claim
that a diploma currently guarantees these skills. The only logical conclusion
is that those who would go back to the old system believe students should
receive a diploma even if they have not been taught basic cognitive skills, so
that they may continue to be pooled with those who have. This may seem to
be a convenient arrangement for those schools that graduate mostly high
achievers, while waving through their lagging students with a wink and a nod.
But it is no longer a credible option for those schools in disadvantaged dis-
tricts whose graduates are known to often lack basic skills and whose
communities have been notably absent from the protests against standards-
based reform.
Comment by Herbert J. Walberg

Julian R. Betts and Robert M. Costrell have contributed a splendid paper with useful policy implications. Two of their major arguments are the focus of my comments: (1) multiple cut points are better than single cut points for standards, and (2) standards appear to be working.

Multiple Cut Points

When I chaired the Design and Analysis Committee of the National Assessment Governing Board (NAGB), the committee considered the possibility of a dichotomous criterion of proficient versus nonproficient. But the committee members and I felt more information would be useful. In the end, we proposed three levels—Basic, Proficient, and Advanced. NAGB maintained these distinctions, and other groups have since found them useful. (In fact, there is a fourth level; that is, Below Basic.)

A disadvantage of several cut points, however, stems from the technical limitation of psychometric tests, particularly non-multiple-choice varieties. Conventional tests are optimized for discriminating best at a single cut point, often near an average. Other things being equal, they cannot discriminate equally well at other cut points. Computer-adaptive testing and longer tests can overcome this problem. However, these solutions may be innovative, costly, and difficult to implement.

Betts and Costrell emphasize that multiple cut points and differential standards give individual students across the ability spectrum reasonable targets to aim for. I would add that such differentiated standards can serve as goals for schools, local school districts, and states. If a dichotomous criterion of proficiency were employed, schools, for example, might concentrate all their energies on proficiency and neglect students near NAGB’s Basic and Advanced levels. Still, it is not clear how many levels are needed, and several precedents may be useful in planning the best choice, given several variations in education policy.

The most long-lasting professional precedents for standards are the M.D. and the J.D.—both dichotomies. The medieval craft distinction was apprentice and journeyman, another dichotomy, which still reigns in many apprentice and other induction systems. The value to consumers of such dichotomous distinctions may be their simplicity; they economize on mental effort, a surprisingly precious resource. People want a proficient plumber or physician.
Nonproficient practitioners such as apprentice plumbers and medical students may require extra negotiation, supervision, and uncertainty. Nonetheless, advanced specializations signaled by board certification and extensive experience may be useful indicators when nonroutine decisions are in order, as in the case of choosing a surgeon. Michael de Bakey, for example, as a consequence of carrying out thousands of heart operations, has dealt with rare complications that naturally arose in his specialized practice.

Occupational and professional practices and education systems in other countries may suggest tying standards to examination performance, years of schooling, or both. Many countries, for example, employ a trichotomous system of education—six years of primary school and three years of lower secondary school, both of which are compulsory. Three years of upper secondary education are elective. This third level may require matriculation examinations, and it may provide a terminal technical diploma or stepping-stone to managerial and professional careers requiring a university education.87

Americans would like all of their students to be excellent and equal, too. But the country has not been successful at either of these contradictory criteria. U.S. test scores, especially valued-added test scores, compare unfavorably with those in other countries. At the same time, the United States has fallen behind other countries in the egalitarian ideal of providing a secondary education for all students.88

For these reasons, it may be useful to consider standards-based school-leaving examinations required for a diploma. These might even be dichotomous standards required for diplomas, say, after nine and the usual twelve years of schooling. The lower secondary diploma might guarantee general or basic proficiency for many nontechnical, nonmanagerial, and nonprofessional jobs. Several kinds of upper secondary diplomas could be created, one of which would be college preparatory; others might qualify students for specialized technical occupations such as computer programming.

Some may want to avoid combining standards with the length of schooling. If so, I would still maintain as a psychologist that students will vary in their learning rates and their knowledge and skills at age eighteen (and even when they start school). At the end of twelve years of schooling, some high school students can exceed the performance of college graduates; others score at primary school achievement levels. For this reason, the possibility should be considered of granting diplomas for achieving standards instead of spending time in school. Such a possibility would allow some students to graduate from high school with as little as nine instead of twelve years of schooling.
Betts and Costrell have reasonably argued for more finely graded information on standards attainment. The next step is to formulate precisely how such information, however finely divided, may be employed in specific policies, especially those that would boost achievement and save costs—including the often forgotten costs of student time. A fascinating menu of possibilities is on offer, and the most promising should be tried and evaluated.

*Standards Work*

The theory and evidence presented by Betts and Costrell and others suggest that standards elicit academic effort and improve achievement. This conclusion accords well with common experience. The assumption, for example, which economists share, is that people usually arrange their affairs to accomplish what they value while minimizing their time, costs, and risks. Most managers in the workplace take pains to specify performance standards and related incentives. Not only might standards improve students’ performance in school but they would also better prepare them for the real accountability that is customary in workplaces.

Even though some evidence suggests that standards would improve achievement, Betts and Costrell are cautious about reaching this conclusion. In addition to their concerns, other worries are teaching to tests, possibly higher dropout rates, intense disciplinary battles over content such as in history, and so on. There may be further problems that have not been emphasized, such as breaches of test security engendered by high stakes. These problems can be overcome as policymakers and administrators gain more experience.

Chester E. Finn Jr. and Marci Kanstoroom show that states are only slowly and fitfully enacting standards, and few have powerful carrots and sticks. Because theory, evidence in hand, and experience in other fields support the causal efficacy of standards, faster and more forceful implementation and experience in remedying obstacles could produce substantial results. This would seem especially likely if such efforts included desirable implementation features.

*Incentives*

Although it seems their implicit assumption, Betts, Costrell, and others deal less directly with the added effect that incentives bring to standards. Evidence suggests that the assumption is reasonable, and greater incentives are needed given that they appear to have substantial large effects. A 1997 Public Agenda national survey of high school students showed that three-fourths believe
stiffer examinations and graduation requirements would make students pay more attention to their studies. Three-fourths also said students should not graduate who have not mastered English, and a similar percentage said schools should promote only students who master the material. Almost two-thirds reported they could do much better in school if they tried. Nearly 80 percent said students would learn more if schools made sure they were on time and did their homework. More than 70 percent said schools should require after-school classes for those earning D’s and F’s.

In these respects, however, teacher educators differ sharply from students and the public. A 1997 Public Agenda survey of education professors showed that 64 percent think schools should avoid competition. More favored giving grades for team efforts than did those who favored grading individual accomplishments.

Teacher educators also differ from employers and other professions on measuring standards or even employing them at all. Employers use standardized multiple-choice examinations for hiring. And selective colleges and graduate and professional schools use them for admission decisions. Such examinations are required in law, medicine, and other fields for licensing because they are objective and reliable. Yet 78 percent of teacher educators wanted less reliance on them.

Nearly two-thirds of teacher educators admitted that education programs often fail to prepare candidates for teaching in the real world, and only 4 percent reported that their programs typically dismiss students found unsuitable for teaching. Thus, even starting with their undergraduate education, many prospective educators are laden with anticompetitive ideas against standards and incentives.

Seventy-nine percent of the teacher educators agreed that “the general public has outmoded and mistaken beliefs about what good teaching means.” They apparently have forgotten that citizens, who pay for schools, constitute their ultimate clients. Perhaps the public and students are right. It seems a good time to raise the question of whether standards associated with incentives and disincentives can work in schools as they do in much of the rest of society.

Educators hope that long-range and somewhat vague incentives will motivate students, but graduating from high school, going to a good college, and making money may seem hopelessly distant. Instructional psychologists like clear and immediate feedback, preferably within seconds. This is a potential aspect of the new computer and Internet technologies.

I recently evaluated an Advanced Placement (AP) incentive program for the O’Donnell Foundation in Dallas, Texas. This program is to my knowledge
the first clear-cut, large-scale trial of monetary incentives for public school students. The program made use of the Advanced Placement examinations, the only national tests that provide external, objective, and rigorous standards for high school students. More than half a million high school students take AP exams on the content of more than twenty-five college-level courses. More than twenty-five hundred colleges grant course credits for passing grades, allowing students to graduate early or take more advanced college courses.

Beginning with the 1990–91 school year, the incentive program paid students $100 for each passing AP examination score in English, calculus, statistics, computer science, biology, chemistry, and physics plus a reimbursement for the cost of taking the exam. The program also provided a $2,500 stipend to each teacher undergoing training to teach advanced courses in these subjects. The teachers also received $100 for each passing AP examination score of their students.

In the nine participating Dallas schools, sharply increasing numbers of boys and girls of all major ethnic groups took and passed the AP exams. The number rose more than twelvefold from 41 the year before the program began to 521 when it ended in 1994–95. After terminating, the program continued to have carry-over effects: In the 1996–97 school year, two years after the program ended, 442 students passed, about eleven times more than the number in the year before the program began.

Though these numbers speak for themselves, interviews with students, teachers, and college admission officers revealed high regard for the incentive program. They felt that even students who failed AP exams learned better study habits and the importance of hard work to meet high standards.

In addition, the program had other benefits: Students could take courses that are more advanced in college. Those who passed a sufficient number of AP courses could graduate from college early, which saves their families the cost of tuition and saves taxpayers the cost of subsidies. Those who passed AP courses also had a better chance for merit scholarships and entry into selective colleges.

Conclusion

The Advanced Placement incentive program shows that standards with incentives work in schools as they do in many spheres of life. The lack of school incentives may be the reason that students find academics so boring and sports so exciting. Social promotion and graduating students for mere atten-
dance are insufficient. Nor can paying teachers for their degrees and years of experience bring out their best. Though not all incentives are monetary, rational people require reasons to work hard.

Comment by Meredith Phillips and Tiffani Chin

Julian R. Betts and Robert M. Costrell make several contributions to policymakers’ understanding of standards-based reform. First, Betts and Costrell’s economic analysis helps identify the likely costs and benefits of policies that require higher standards for high school graduation. Their analysis reveals where policymakers should target resources to mitigate the probable costs of raising standards. They point out, for instance, that policymakers should focus their energies on the group of students who may reduce their effort or drop out of school because they do not expect to be able to meet the higher standard.

Second, the authors try to determine how higher standards will affect the achievement of high school students. Because accountability policies are so new, direct evidence of their effectiveness is thin. Therefore, the authors review data about the effects of more homework, harder grading standards, and summer school on student achievement. Although more research is needed on these topics, Betts and Costrell’s review of this literature suggests that asking more of students generally raises their achievement. This conclusion is consistent with claims from the Effective Schools movement of the 1980s about the importance of “academic press” for student achievement. And although Betts and Costrell correctly worry about possible negative effects of raising standards for lower-achieving students, most of the evidence they review points to the benefits of high standards for all students. Only time will tell whether this conclusion continues to hold after states implement higher standards for retention and high school graduation.

Because microeconomics is the study of how rational actors respond to incentives, microeconomic theory provides a good starting place for an analysis of the likely costs and benefits of standards-based reform. However, Betts and Costrell’s economic perspective ignores some insights from sociology and psychology that may be useful for constructing policies that maximize the benefits (and minimize the costs) of standards-based reform. Moreover, much of Betts and Costrell’s analysis focuses too narrowly on how high school students will respond to high-stakes assessment. The effects of standards-based reform on younger students and parents, as well as the people and institutions
implementing the reform such as teachers, schools, and districts, deserve as much, if not greater, attention.

Much of Betts and Costrell’s discussion, and much political rhetoric, assumes that high-stakes testing will improve student achievement by encouraging students to work harder. This assumption has several flaws. To begin with, it relies on a reinforcement theory of motivation. Scholars of motivation have identified a number of problems with trying to improve student achievement by using extrinsic rewards and sanctions. First, several studies have shown that external rewards may make students less willing to take on challenging tasks. When students expect to be rewarded for some discrete accomplishment, such as reading a certain number of books or earning straight A’s, they typically choose the path of least resistance so as not to jeopardize their chances of receiving the reward. Second, if students are encouraged to engage in a behavior, such as learning, simply to receive a reward, such as a high school diploma, this may convey the message that learning is not worth doing for its own sake. Emphasizing external incentives for learning may not be an optimal long-term strategy for encouraging students to become life-long learners. Third, worries about punishment and failure can cause excessive anxiety, which may hinder learning. In our study of fourth graders, we found that some students whose report cards were marked “may be retained” because of their low performance on Stanford 9 exams would raise their hands in the middle of lessons and go up to teachers and aides in the middle of activities to ask, often for the third or fourth time that week, “Am I going to pass?” Clearly, for these students, the possibility of failure interfered with the learning process.

Finally, and most relevant to the testing-with-consequences debate, if punishment is inappropriately applied, it can have disastrous consequences for students’ future effort and learning. Consider, for example, children who study hard and pay attention in class but fail a test (perhaps because they know little English because they recently immigrated to the United States, or because they often miss important lessons because their parents consistently bring them late to school). If these children are punished (for example, retained) because they did not meet a particular academic standard, they are effectively being punished rather than rewarded for their effort. As the behaviorists would predict, if punishment follows effort, these children will exert less effort the next time around. While teachers’ grades typically try to reward effort, even when a child does not perform as well as other children, high-stakes tests are less subjective and thus cannot distinguish poor performance because of low effort.
from poor performance despite high effort. Any system that punishes students when they have done their best will, in effect, encourage such students to decrease their effort in the future. Policymakers want to avoid this unintended negative consequence, especially among young children. If these children become discouraged and their efforts flag, they will have many years to disrupt not only their own learning, but also that of their classmates.

A second problem with Betts and Costrell’s assumption that high-stakes testing will improve student achievement by encouraging students to work harder is that it ignores differences between elementary school-age children and adolescents. While adolescents may be expected to rise to challenges by exerting more effort, younger children usually try their best from the outset. And, although higher standards may push adolescents to direct more time and energy into academic rather than social endeavors, when younger children divert their energy from school-related tasks, they tend to do so for irrational reasons (for example, they are unintentionally distracted by peers or emotional issues). Thus, for younger children, even the strongest incentives may not be enough to improve academic performance, a conclusion that the study by Melissa Roderick and her colleagues seems to support. Roderick and her colleagues found strong incentive effects of the threat of retention for sixth and eighth graders, but little effect for third graders. Although this may just be an anomaly of the data, it may also signal developmental differences in how children of different ages respond to the same incentives.

A third problem with assuming that high-stakes testing will boost students’ effort and thus improve their academic achievement is that many factors beyond student effort affect how much math students can do and how well they can read and write. Although a certain level of student effort may be necessary for passing high-stakes assessments, it is unlikely to be sufficient. For students to meet high standards, parents and teachers also have to exert more effort. But, ultimately, effort alone cannot improve students’ skills. Suppose policymakers created the ideal package of incentives that encouraged students to pay attention and complete their schoolwork, parents to help with their children’s homework, and teachers to spend extra time preparing lessons and helping students. All this effort would probably not raise students’ math scores much unless the students also had high-quality textbooks that they could take home, parents who had learned and still remembered some geometry and algebra, and teachers who both knew and could teach math. In other words, imposing consequences on students is unlikely to pay off in higher achievement unless it comes packaged with other standards-based reforms that encourage parents,
teachers, schools, school districts, states, and the federal government to provide both the effort and the resources needed to raise achievement.

Providing sufficient resources for students to meet standards will also help make standards-based reform fairer for the students it impacts the most. When schools raise the bar for promotion to the next grade or for high school graduation, the consequences (both positive and negative) will be greater for disadvantaged children than for advantaged children. Under higher standards, African American, Latino, and poor children will probably work harder, be taught more, and thus learn more. Because standards-based reform is targeted at the bottom of the achievement distribution, it has the potential to reduce the achievement gap, which would be an important accomplishment. However, raising standards will also mean that a greater percentage of African American, Latino, and poor children will be retained or denied a high school diploma.

Although Betts and Costrell downplay the negative effects of this relabeling of some students as nongraduates, the persistent salience of race in the United States may make these sorting effects more negative than the authors assume. Retained students or the nongraduates will, at least in the short run, be disproportionately African American and Latino. This pooling of ethnically identifiable students into the category of “failures” will perpetuate the already prevalent stereotype that non-Asian minorities are intellectually inferior to whites and Asians.

The continuing salience of race may also make the negative incentive effects of high-stakes testing more egregious for some groups than for others. For example, if students suspect that tests are culturally biased and thus do not measure useful and important skills, students may decrease instead of increase their effort in the face of higher standards. Moreover, Claude M. Steele and Joshua Aronson have shown in randomized experiments that African American students tend to underperform relative to whites in settings where they risk confirming stereotypes about their group. This “stereotype threat” may not only negatively influence African American students’ scores on high-stakes exams, but it may also interfere with their assessments of their likelihood of passing high-stakes tests. If African American students routinely underestimate their probability of passing the test, they may become more discouraged (and possibly drop out earlier) than students from other ethnic groups with the same level of academic skills.

These race-related, negative effects should not discourage policymakers and educators from raising academic standards for all students. They should, however, highlight the importance of making standards-based reform fair.
This can be done by developing a general consensus about the skills and knowledge that high-stakes tests should measure and by making prototypical test items available for public scrutiny. Giving students a number of opportunities to take and retake high-stakes tests will also improve fairness. And because grades are better measures of students’ effort and enthusiasm than tests are, consequential decisions should incorporate both tests and teachers’ evaluations.

Perhaps most important, standards-based reforms can improve fairness by ensuring that disadvantaged students have sufficient access to high-quality extra instruction (in summers, in pullout programs, and after school) so that they can boost their test performance. Parents also need accurate information about how they can help prevent their child’s failure. In one community we have been studying, when affluent parents learn that their child may be retained, they pull out all the stops. They buy educational computer software and hire private tutors at fifty dollars per session to teach the child how to take a test apart, avoid the “tricks,” and navigate the easiest parts of the test to gain confidence.

But the lower class children are relegated to district-run Saturday school where teachers who were never trained in test preparation ineffectively coach the students for the high-stakes test. Not only do the teachers repeatedly drill the children with tests that are too difficult for them (making the actual test seem only more daunting) but the children are also taught poor test-taking strategies. In one instance, a teacher explained to the children that the best way to find the “main idea” in a reading comprehension passage is to look at the first sentence. She then required the children to repeat this technique over and over again. While this technique generally works on the fourth-grade tests, a quick glance at fifth grade (and higher) tests shows that passages often start with an anecdote and the first sentence quickly becomes the standardized test’s favorite “trick” answer. Thus, even though these lower-class children commit to spending their Saturdays in a classroom, they leave feeling more discouraged, with their heads drilled full of incorrect techniques.

While disadvantaged children are the most likely to fail under higher standards, Betts and Costrell note that most of the emerging opposition to standards-based reform comes from affluent communities. Highly educated and affluent parents (and, in fact, any parent of a high-achieving child) should worry that teaching to a test that is targeted at the bottom of the achievement distribution will short-change their high-achieving child. This is a big concern in heterogeneous classrooms (such as most elementary classrooms) because high-stakes tests give teachers incentives (and often a moral imperative) to
focus time and effort on lower-achieving students. In the fourth-grade classroom that we have been studying, the teacher devoted almost a month to preparing her students for a performance assessment of the concepts of area and perimeter. The high-achieving students learned the new skills eagerly but mastered them within the first couple of lessons. They then spent the next four weeks repeating the skill over and over again while the rest of the class caught on. That same year, these children found themselves stuck with a science curriculum that never covered anything but volcanoes, largely because so much class time was devoted to test preparation. Even in districts where high-achieving and low-achieving students attend different schools, high-achieving students may still be shortchanged if administrators are forced to take resources away from high-performing schools to help low-performing schools meet the new standards.

Although we support the general principles of raising academic standards and creating a high school diploma that “means something,” we are less optimistic than Betts and Costrell about the positive consequences of standards-based reform. Assuming that children and teens will respond strongly to the positive incentives of higher standards glosses over motivational differences among children and teens who are moving through a range of developmental stages and come from a variety of ethnic and social class backgrounds, home and community support structures, and past experiences with academic success and failure. Further, placing the burden of meeting high-stakes standards on these students, without providing generous institutional support, will doom too many of them to failure. Finally, dismissing middle-class parents’ concerns as purely self-serving seems shortsighted. To maximize every child’s academic potential and maintain sufficient political support for standards-based reform, policymakers should avoid reforms that impede the academic progress of bright students from any background. In the end, standards-based reform cannot be a quick fix to the problem of low academic performance. But if policymakers consider all these complexities, it may turn out to be a step in the right direction.

Notes

1. See Public Agenda polls in recent years.
3. See Sandra Stotsky, ed., What’s at Stake in the K-12 Standards Wars: A Primer for Educational Policy-Makers (New York: Peter Lang Publishers, 2000). Authors such as Stephen
Arons have argued that such battles over content are a permanent feature of the public (or common) school system and can only be fully resolved by a thoroughgoing system of school choice and vouchers. Stephen Arons, *Short Road to Chaos* (Amherst: University of Massachusetts Press, 1997).

However, with or without vouchers, the demand for educational accountability in the use of public funds seems likely to rise, particularly in states where the share of funding is shifting from the localities toward the state. The specification of content standards and measurable outcomes is central to these accountability efforts.


5. A century ago, when a high school diploma was held by a small minority of the population, far less stigma was attached, economically or otherwise, to being a nongraduate. Similarly, under the traditional British system that prevailed until recently, many students left school at age sixteen. Far more students left school at this age than occurs in the United States, and the stigma was presumably much less, because their numbers included more capable workers.

6. Under this model, they should favor standards that are so high that everyone fails, so that the lowest achievers are pooled with the very best. This may seem indistinguishable from the opposite extreme, where the standard is set so low that everyone passes and is similarly pooled together. However, unless the results are perfect, with a 100 percent pass rate, the strategy of a very low standard will lead to the least egalitarian outcome, by the Rawlsian standard, because the rare failure is most highly stigmatized. In short, the wage of failers rises monotonically with the standard in this simple model. See Betts, “The Impact of Educational Standards on the Level and Distribution of Earnings.”

Robert M. Costrell relaxes a key technological assumption of this model—that the productivity of any individual is independent of other individuals (perfect substitutability, to use the technical term from economics). Suppose, instead, workers operate in teams, providing complementary services in the production of output, as in the job assignment model of Robert M. Costrell and Glenn C. Loury. Then it can be shown that another effect of raising standards works in the opposite direction from the pooling effect. High standards reduce the number of workers supported by those of lesser skill, which tends to reduce the wage of failers. Taken together with the pooling effect, raising standards need not have a monotonic effect on the wage of failers. Costrell finds that, in a benchmark case, the relationship between the wage of failers and the standard is U-shaped, and, moreover, the standard that minimizes the failers’ wage maximizes output. Costrell also analyzes the effect on this relationship of varying technology, cost of acquiring skill, and test accuracy. An important finding, however, is that those cases in which a rise in the standard reduces the wage of failers are also the cases in which equity is most likely advanced by moving away from pass-fail systems altogether, toward fuller information. Robert M. Costrell, “Are High Standards Good or Bad for Those Who Fail?” University of Massachusetts at Amherst, Department of Economics, 1999; and Robert M. Costrell and Glenn C. Loury, “Distribution of Ability and Earnings in a Job Assignment Model,” University of Massachusetts at Amherst and Boston University, 2000.


8. In addition, schools facing the prospect of higher failure rates would also respond with interventions to assist at-risk students.

9. It is an empirical matter of some importance how much less the rise would be, whether it would be closer to the full ten-point rise or closer to zero.
10. See Robert M. Costrell, “An Economic Analysis of College Admission Standards,” *Education Economics*, vol. 1, no. 3 (1993), pp. 227–41, for a formal analysis of the effect of standards in the context of college attendance, where students are uncertain how difficult college will be until they get there. A rise in admission standards forces applicants to be better prepared and can raise the resulting number of graduates, even though the number of attendees declines.

11. The analysis here excludes consideration of possible externalities created by peer effects. If there are adverse peer effects generated by some of those who are unwilling or unable to exert extra effort to pass, and if the potential benefit for some of staying in school is low, then the optimal dropout rate may not be zero. Disruptive students provide an obvious example that is unfortunately not as rare as one might hope. The best solution in such cases is not necessarily to encourage dropouts, but to create alternative educational settings for such students, such as those under creation by systems in Boston and Chicago, as long advocated by the American Federation of Teachers, among others.

12. Evidence consistent with the bifurcation in this part of the distribution is found in the contribution to this volume by John H. Bishop, Ferran Mane, Michael Bishop, and Joan Moriarty. They find that among C/C- students, minimum competency exams raise both the number of noncompleters and the number of college attendees.

13. Although the general points discussed here and depicted in figure 1 derive from the theoretical literature cited, figure 1’s continuous distribution is not strictly consistent with that literature’s simplest theoretical models. Those models generate distributions with discrete segments and a discontinuity in the vicinity of the standard.


16. It seems more likely that there could be some redistributive effect on learning in the lower grades, where heterogeneous grouping prevails.


18. This does not prevent some of the critics in these communities (both parents and educators) from couching their objections in egalitarian terms, as the defenders of those children in less-advantaged areas whose parents have chosen not to object.


21. John Bishop has provided evidence in a number of papers over the years that is consistent with this behavior of employers. See, for instance, John Bishop, “Incentives for Learning: Why American High School Students Compare So Poorly to Their Counterparts Overseas,” *Research in Labor Economics*, vol. 11 (1990), pp. 17–52.
22. The extent of this problem is inversely related to the strength of local reputation, which in turn depends on the size of the entities in question.

23. This assumes that no systematic difference exists between local and central authorities regarding the weights attached to winners and losers (that is, they hold the same social welfare function).

24. With cross-district heterogeneity, it can be the case that egalitarian societies—those that assign greatest weight to preventing dropouts—should prefer centralization even more than nonegalitarians. The problem of free-riding under decentralization is more pronounced for egalitarians because they tend to cut standards further below the optimal level. That is, egalitarians may like low standards in their own district, but they face particularly high losses from the free-riding of their fellow egalitarians in other districts choosing particularly low standards. Both egalitarians and nonegalitarians favor centralization if all districts are alike, but under cross-district heterogeneity, egalitarians may favor centralization in some cases that nonegalitarians do not.


26. Different patterns can emerge, depending on the degree of pooling. But the general point remains: There are winners and losers in any system of standard-setting, compared with any alternative.

27. It is not even certain that a centralized standard-setter would choose a higher standard than any of the localities. If the optimal central standard is tailored to the weakest districts (as it will be under some circumstances), then the central standard could end up even lower than those weak districts would choose on their own. The reason is that under decentralization, the stronger districts would choose high standards, raising the wage of non-college-bound graduates everywhere, including those in the weaker districts, to the extent they are pooled together. This would enhance the incentive for students in the weaker districts to graduate, which, in turn, allows those districts to set higher standards than otherwise without deterring too many students from graduating. In this way, it is possible that under cross-district heterogeneity central standards could be lower than under decentralization. Even if standards rise for some or all districts under centralization, the constraint that all districts face the same standard may still lead to lower social welfare than under decentralization.

28. This is the law in Massachusetts: No district will be able to award a diploma to students who fail the Massachusetts Comprehensive Assessment System (MCAS), but districts can impose additional graduation requirements, including a higher MCAS score.

29. Costrell and Loury, “Distribution of Ability and Earnings in a Job Assignment Model,” applied to the issue of standards by Costrell, “Are High Standards Good or Bad for Those Who Fail?”

30. For a formal analysis, see Costrell, “A Simple Model of Educational Standards,” section VI.


34. Economists have documented that they have a generally high rate of time preference.

36. For Massachusetts, this will change dramatically, beginning with the class of 2003.

37. The American Federation of Teachers (AFT) has published an annual review of the educational standards in each state, Puerto Rico, and the District of Columbia. These publications provide a succinct overview of progress, and because the AFT gives each state an opportunity to respond to the annual synopses, the synopses gain credibility. The summary draws heavily from these AFT analyses. Data for 1999 and 1996, respectively, are from American Federation of Teachers, Making Standards Matter 1999 (Washington, 1999), and American Federation of Teachers, Making Standards Matter 1996 (Washington, 1996).

38. Massachusetts meets the criteria for clear standards and aligned assessments, but its exit exams for the class of 2003, which were established by law in 1993, were not formally voted upon by the Board of Education until the fall of 1999 (and only for math and English), too late for inclusion in the AFT tables.


40. These three variables were obtained from Bureau of the Census, Statistical Abstract of the United States: 1998 (Government Printing Office, 1998), pp. 34, 169, 479.


48. Melissa Roderick and others, Ending Social Promotion: Results from the First Two Years (Chicago, Ill.: Consortium on Chicago School Research, 1999).

49. We do not know what is happening among the individuals within any grade equivalent (G.E.) category. There may be individuals in the lower G.E. categories whose effort is at the same low level that it would have been without the standards, among those who still fail to pass the test after summer school. However, we are struck by the fact that the strongest average response is in the lowest G.E. category.

50. The other important piece of background, aside from the court case, was a large drop in state aid that occurred in Massachusetts’s deep recession of 1989–92, which followed the
unsustainably rapid growth in spending in the latter part of the 1980s. During that recession the income tax rate was temporarily raised to balance the budget, but it remained high after the recession ended and has not yet been returned to its 1989 rate.

51. Compared with the prerecession figure, real state aid grew by about one-half.
52. Calculations by the Executive Office for Administration and Finance.
53. Thomas J. Kane, “An Update on School Reform in Massachusetts,” John F. Kennedy School of Government, Harvard University, 2000. Compared with the prerecession year of 1989, the corresponding rise by 1998 was $513 at the tenth percentile and $165 at the ninetieth.
55. John Silber, who was chairman of the Massachusetts Board of Education until early 1999, was a vociferous opponent of this weakening of the standard. Exams in the Needs Improvement category were judged by standard-setting panels to meet the description that “Students at this level demonstrate a partial understanding of subject matter, and solve some simple problems.” Neil M. Kingston, “The Body of Work (BoW) Standard Setting Method: Massachusetts Comprehensive Assessment System,” presented at the annual meeting of the National Council on Measurement in Education, New Orleans, La., 2000. For the tenth-grade math exam, this required twenty-four out of sixty possible points.
57. Students also take a few matrix-sampled questions each year, which do not count toward their scores, but from which future core questions are drawn. That means that each year core questions have been seen by a few students the previous year but have not been made public. The English language arts essay questions, however, are not matrix-sampled the previous year.
58. Less than 4 percent left this question blank.
59. *Up and Over the Bar* (Mass Insight Education, April 2000). The factors isolated in this study are not mutually exclusive, so percentages sum to more than one hundred.
60. Twenty-five of twenty-eight superintendents interviewed for the Mass Insight Education study report that “motivation on the test” was one of the primary factors. Many of them report a significant difference in attitude toward the test between current tenth graders, for whom it does not count, and those in ninth grade, for whom it will.
61. In Indiana, 54 percent of tenth graders passed the math and English exit exams on the first try that counted, in fall of 1997, but by time that class was to graduate, in 2000, 86 percent had passed both exams. The exam is pitched at a ninth-grade level. Students also have two alternative routes to a diploma. Lynn Olson, “Indiana Out in Front on Giving Students Extra Help,” *Education Week*, May 31, 2000.
62. Boston superintendent Thomas Payzant, an advocate for MCAS, uses the MCAS and other exams to identify students most at risk. Over 30 percent of students in grades two through nine (except four) now face mandatory summer school, to avoid grade retention. Ed Hayward, “12,000 Hub Kids Face Summer School,” *Boston Herald*, June 7, 2000, p. 1. During the regular school year many of these students will receive doubled instruction in literacy and math, and customized lessons. A *Boston Globe* editorial (April 24, 2000) opined, “This is the kind of urgent remedial attention that many students need and should have been getting for years. They are getting it now because the state Education Reform Act of 1993 is supplying money and MCAS is applying pressure.” Payzant observed, “We wouldn’t be as far along with our reform efforts if there weren’t high stakes along the way.”

64. These are Massachusetts percentiles, not U.S. percentiles. Because Massachusetts performs above the national average, at a thinner portion of the U.S. distribution, the shift in U.S. percentiles would be smaller. It should also be noted that the improvement in Iowa Test of Basic Skills (ITBS) scores bypassed the lower third of the Massachusetts distribution. One possible interpretation is that efforts of third-grade teachers to prepare students with skills useful for fourth-grade MCAS paid off only for those third graders who were academically more prepared for the challenge. Another possible factor was a 1999 change in the Massachusetts regulations, which expanded the number of limited-English proficient (LEP) students required to take the test. The ITBS scores considered here are reading total, which is an average of reading comprehension and vocabulary. The rise in reading comprehension was larger than that of vocabulary and reading total.

65. Specifically, we analyze the ITBS school effects for 1997, 1998, and 1999, and the MCAS school effects for 1998 and 1999. The 1998 ITBS school effect is fitted to an equation with the 1997 ITBS school effect. Using that equation to predict the 1999 ITBS school effect, we calculate, school-by-school, how much better the ITBS school effects turned out than predicted. These second-year-effects for ITBS are then added to MCAS 1998 school effects in a regression for MCAS 1999 school effects and found to be highly statistically significant.

66. Not all students are averse to the exams. The Massachusetts Student Advisory Council defends the MCAS. “No one would argue that passing the MCAS is all education is about,” said the student representative to the Massachusetts Board of Education. “But the idea of sending someone out without ascertaining that they can write a coherent paragraph or do algebra or geometry is unthinkable.” Jules Crittenden, “Some Students Call MCAS Boycott Counterproductive,” *Boston Herald*, April 14, 2000.

67. Millicent Lawton, “The Acid Test,” *Commonwealth Magazine*, Spring 2000, p. 46. Cambridge spends about $12,000 per pupil, among the highest in the state, but nonetheless scores below the state average on MCAS.


69. The ads feature students cheering their Scholastic Assessment Test (SAT) scores, which have risen of late in Massachusetts. The Massachusetts Teachers Association (MTA) had not previously been known to advocate for the SAT, but its spokesman now says, “The SAT scores are one of the most reliable indicators in the public’s mind about how the schools are doing. On the MCAS, the jury is still out.” Steve Leblanc, “Teachers Union Touts SAT Scores in Television Ads,” Associated Press, April 18, 2000. More recently, a $600,000 MTA ad campaign against MCAS portrayed students suffering from anxiety during that exam.


74. A comparable percentage of Boston students in grades two through nine is now being held back, contingent on successful summer school remediation.


79. For a summary of this program, see Julian R. Betts, Kim Rueben, and Anne Danenberg, *Equal Resources, Equal Outcomes? The Distribution of School Resources and Student Achievement in California* (San Francisco: Public Policy Institute of California, 2000), chapter 1.


82. See chapter 5 of Ravitch, *National Standards in American Education*, for a summary and critique of this movement.

83. But see Caroline Minter Hoxby, “Are Efficiency and Equity in School Finance Substitutes or Complements?” *Journal of Economic Perspectives*, vol. 10, no. 4 (Fall 1996), pp. 51–72, for an analysis of the pitfalls in moving too far toward state finance of local education, as a result of equalization suits.


86. Betts, Rueben, and Danenberg, *Equal Resources, Equal Outcomes?*


90. Steve Farkas and Jean Johnson, *Different Drummers: How Teachers of Teachers View Public Education* (New York: Public Agenda, 1997).


92. The following discussion is drawn from Deborah J. Stipek, *Motivation to Learn: From Theory to Practice*, 2d ed. (Boston: Allyn and Bacon, 1993).


