INCENTIVES AND EQUITY UNDER STANDARDS-BASED REFORM

Julian R. Betts[1] and Robert M. Costrell[2]

Forthcoming in Diane Ravitch (ed.), *Brookings Papers on Education Policy 2001, (Washington, D.C: Brookings Institution)*.

The authors thank Meredith Phillips, Diane Ravitch and Herbert Walberg for helpful suggestions.

[1] Associate Professor, Department of Economics, UCSD and Senior Fellow, Public Policy Institute of California, San Francisco, California. Mailing address: Department of Economics, University of California, San Diego, La Jolla, CA 92093-0508, jbetts@ucsd.edu.

[2] Professor of Economics, Department of Economics, University of Massachusetts, Amherst and Director of Research and Development, Executive Office for Administration & Finance, Commonwealth of Massachusetts. Mailing address: Executive Office for Administration & Finance, Commonwealth of Massachusetts, State House, Room 373, Boston, MA 02133, costrell@econs.umass.edu.

Institutional affiliations are provided for identification only.

Abstract

The paper considers theoretical and empirical evidence on the impact of standards-based school reform. Our theoretical synthesis distinguishes between sorting and incentive effects of high standards, and spells out the potential tradeoffs and complementarities between enhancing efficiency and equity in student achievement. Differentiated credentials can be helpful in ameliorating tradeoffs, provided that distinct signals are clearly understood, especially between cognitive and non-cognitive skills. The paper reviews trends in state-level school accountability systems, and examines empirical evidence on the impact of increased standards and expectations on student achievement. Finally, the paper reviews some of the practical challenges facing the standards movement.

JEL Codes: I2 (Education), J24 (Human Capital Formation; Occupational Choice; Labor Productivity)

INCENTIVES AND EQUITY UNDER STANDARDS-BASED REFORM

Julian R. Betts and Robert M. Costrell

Introduction

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 (e.g. 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 this strategy, according to national polling data,[1] 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% of the general population favors graduation exams. Support is slightly more emphatic from urban than suburban respondents, and somewhat broader (75%) from those with income under $25,000. When respondents are asked if they would still support the exams should 25% of students in their communities fail on the first try, support remains unchanged overall at 70%, and actually rises to 81% among those with income under $25,000.[2]

Nonetheless, vocal, if not yet necessarily wide, opposition has emerged in several states, in the runup 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-90s over national content standards in history and English, and more recently in science and math, have had their counterparts in the states.[3] Despite continuing conflicts, it does seem that certain broad (if not universal) agreement can be obtained in basic content areas (at least math and English). In this paper, we shall not focus on content disagreements, but rather on disputes over testing and cutoffs. However, it is worth bearing in mind that at least some of the more vocal opposition to testing is itself based (if not always explicitly so) on unresolved disagreements over content standards, since it is the tests that 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. To take a key 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 more generally. Indeed, it is the technologically caused losses of those with low cognitive skills over the last two decades that drives much of the standards-based reform movement. So, too, may standards-based reform create its own 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). Rather, it forces us to examine the nature of the losses and craft an appropriate set of policies 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, as we shall explain, it makes a great deal of difference whether the

failure rate rises as a result of sorting or whether it also reflects adverse incentive effects. The distinction is important both for evaluating the costs of increased standards and also for focusing policies to mitigate costs. Similarly, it is important to distinguish sorting and incentive effects among the winners from various points on the educational spectrum.

Clearly, standards generate a mix of sorting and incentive effects, and we need to understand them both. How are incentives altered by standards-based reform, for better or for worse, to encourage or discourage achievement? What are the tradeoffs between some students' losses and others' gains, in learning and/or income? Do these tradeoffs adversely affect equity, as opponents to standards-based reform often claim? Or is equity enhanced by raising standards in schools attended by disadvantaged students? How can we explain the fact that some of the most vocal opposition often comes from the most advantaged districts? Finally, and most importantly, what steps can and should be taken to minimize the losses and spread the gains most broadly from standards-based reform?

Our analysis below begins by reviewing the economic theory of educational standards in order to sort out the several effects of standards on incentives and equity. Although this literature cannot quantify these effects, it can help us understand the forces at play, as well as the dilemmas we face. A key issue that comes out of this analysis is the structure of information. It matters a great deal whether we have a coarse pass/fail signal or more finely grained information, such as multiple credentials.

We then review the current array of state educational standards, and provide some evidence regarding the factors that help explain the variation across states. Next, we turn to the rather limited statistical evidence that currently exists regarding the effects of standards, and also provide some new evidence. Ideally, we would like to know how

strong a response high-stakes testing will elicit from schools, parents, and students in order to bring the failure rate down. We report briefly on the standards controversies in Massachusetts and California. We summarize four key obstacles confronting the movement to raise educational standards, and suggest partial solutions. We conclude by drawing a few lessons for policy-makers that seem justified by the theory and evidence at hand, which might ameliorate some of the potentially harsh tradeoffs.

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. The reason we apply economic theory to the subject of standards is precisely 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 tradeoffs (a familiar phenomenon in economics). To be sure, there are also limitations to the economic analysis of standards, as we discuss below.

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 world, 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, of course, is silent on the magnitude of the rise in the failure rate (which is critical in comparing with the beneficial effects), but it

does help us distinguish between more and less compelling reasons for concern. Specifically, a pair of papers by us brings out the critical distinction between the sorting and incentive effects of standards.[4]

Sorting Effects of Graduation Cutoffs

Consider first a simple sorting model, where behavior and thus learning are held constant, independent of the standard. Then a rise in the cutoff merely re-labels some students as failers who would otherwise be considered passers. There is, by assumption, no effect on learning or productivity, so 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 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 a point stressed by Betts, namely that higher standards also raise the average quality of the pool of non-graduates, insofar as some students who would previously have passed now fail. Since non-graduates (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 non-graduates typically fare so poorly under the existing system is that the ease of social promotion exacerbates the stigma attached to non-graduation.[5] Thus, it is a logical fallacy to argue, as many do, that higher standards will reduce more students to the current economic level of non-graduates; the stigma on non-graduates depends on their average quality, and that depends

critically on the standard itself.

To summarize this very simple model, a rise in standards 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. Such an individual now suffers from being pooled with a group that includes those less skilled than himself (those without the diploma) rather than with those more skilled than himself. There has been no efficiency loss in this pure sorting model, only a distributional effect due to the individual's re-labelling. Do these losses constitute a compelling case against higher standards? We believe not, for two reasons.

First, in terms of the narrow choice between high and low cutoffs, it is not immediately clear that a high cutoff leads 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.[6] Again, the point is not academic: the equity implications of higher standards are not limited to those who are at increased risk of failing, but include also those who would fail in any case, and whose stigma stands to be *reduced*.

The second, and more fundamental reason that we do not find the losses from adverse pooling to constitute a compelling case against higher standards is that it is not the standards themselves that are at the heart of this issue. Rather, the crux of the matter is the imperfect information that underlies such pooling. How concerned should we be if someone loses from no longer being confused with those of greater skill? We should

indeed be concerned about those able students who are now pooled with those of lesser talent, but the answer is not necessarily to reverse the rise in standards and re-classify them with those of greater talent. Perhaps, instead, the analysis suggests that 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 has long argued. We return to this question -- full information vs. binary credentials -- at greater length below, since it arises not only in the context of sorting, but also of incentives.

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/or attitudes toward study.[7] Suppose the cutoff is raised from a level at which 10% fail to one at which 20% would fail *under existing behavior*. Under a pure sorting model, where behavior is held constant, this rise in standards would of course lead to a doubling of the failure rate. Under a more realistic model, students (and their parents) respond to higher standards by re-evaluating the costs and benefits of student effort.[8]

How are the incentives for student effort affected at different parts of the distribution? Consider first those students at or near the 20th percentile under the original

distribution of achievement. In this example, these are students who passed under the old standard by a margin of 10 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 rather than 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 pre-existing distribution of student achievement,[9] because students in this part of the distribution will rise to the challenge.

It is important to emphasize that 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 20th percentile in this example). Unlike the elite, who will easily pass the higher standard with unchanged effort, these are students who 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.[10]

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 10th 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. Indeed, they may simply drop out of school, as critics of standards-based reform warn. This effect is more troubling than the sorting effect discussed above, 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] As Figure 1 illustrates, we can distinguish four groups of students who are at risk of failing under the higher standard:[13]

- 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.
- Other students who would have passed under low standards will not change their behavior and will now fail. (In Figure 1, these are the students remaining between the old and new standards, on the dashed distribution.) These students lose, but only by virtue of being re-labelled.
- 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.
- 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 discussed previously.

Policy-makers 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. What those policies might be is considered below, but the point here is to be clear on what segment of the population is at issue, both for potential losses and gains.

Curiously, though, much of the most vocal opposition to standards-based reform 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."[14] Similarly, efforts in Massachusetts to boycott the state-wide 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.[15]

Perhaps the most plausible claim that suburban critics have to offer is that higher-order skills may be de-emphasized 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.[16] The more elite

students, aiming for selective college admissions, are more likely focused on SATs, 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.[17] If so, it is important to understand that these objections to standards-based reform are not based on equity concerns, but quite the opposite.[18]

Again, 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. Rather, the challenge is to meet these objections by accompanying the standards with policies addressed toward the high achievers as well. In our view, discussed below, this is a rather easier and less pressing challenge than the one concerning lower achievers, who might be discouraged from continuing academic effort.

Centralized vs. Decentralized Standards

What is the proper locus of standard-setting -- Federal, state, or local? Over the last two decades, 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.[19] 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 very simplest case, where all districts are alike, decentralization would likely lead to inefficiently low standards.[20] To see this, 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.[21] 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.[22] 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. Since 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.[23] This logic is independent of the weights attached to winners and losers; even the most egalitarian collection of standard-setters would prefer standards set centrally, rather than each of them riding free in a standard-cutting race to the bottom.[24]

Heterogeneity across districts makes things more complicated, but is also an important factor in understanding current controversies.[25] For example, centralization typically raises standards in low-achieving districts, but *may* lower it in high-achieving ones. To the extent that diplomas reflect *some* degree of district reputation (i.e. 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.[26] Thus, there may be a conflict of interest between those high-striving urban black students whose diploma is enhanced in value and those suburban students whose diploma could be depreciated from that which obtained under decentralized standards.

With heterogeneity across districts, centralization need not always outperform decentralization.[27] However, if we take the analysis one step further, a rather general result obtains. 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. We get 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.[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. We suspect that political considerations beyond the model 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 we have considered assumes there is 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, as we have discussed, 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 policy-makers are able to reduce the degree of cross-district pooling to reduce the need for centralization, then why not reduce intra-district pooling as well, so that high-achievers in any district can be evaluated by their individual accomplishments? It is to this question that we now turn.

Binary Credentials vs. 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 has quite a bit to say about 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 we have no totally satisfactory answer, is important in designing policies to ameliorate the tradeoffs carried by a system of binary credentials.

In understanding the effects of improving information flows over that of binary credentials, it is again important to distinguish sorting effects from incentive effects. Consider the simplest case, where 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, where 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, we do not find the case for redistribution by blurring of differences to be compelling, unlike a case based on improved incentives.

Even before considering incentive effects, however, there is another aspect of sorting that bears examination, and that is 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, i.e. it is more important to get the very best people into the

very top jobs 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 reason is that the wage earned by the least-skilled depends very much 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, due to 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, since 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 to a coarse pass-fail signal, better information about high achievement is surely a stimulus to those at the top of the distribution, who would otherwise find no payoff in exceeding the cutoff. This, it seems to us, provides much of the answer to the criticism that high-achieving districts are "dragged down" by standards-based reform. Clearly, high-achieving students are already motivated to excel by an array of credentials over and above high school graduation exams (e.g. SAT's and AP exams). If these are insufficient, it is a relatively simple matter to differentiate diplomas based on the level of performance on the graduation exams, as a number of states do.

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 multi-dimensional, for those who find traditional graduation requirements overly narrow. For example, there are many credentials based on artistic and musical talent that students place 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.

It might be argued that schools will 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, rather than pass rates only (e.g. the real estate market may tend to do this). However, this effect should not be exaggerated, since districts will surely 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,[31] fuller information at this end of the distribution advances egalitarian goals by giving students less extreme alternatives to dropping out.

This is the rationale 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 non-cognitive skills such as persistence, punctuality, and discipline that are also important and rewarded in the labor market.[32] The 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 non-cognitive skills as even a diploma based on "seat time" alone.[33] So there remains room for a credential to certify such non-cognitive skills (which may be particularly important for some special education children).

The challenge is to make sure that such a non-cognitive credential is properly differentiated from a standards-based credential that signifies both cognitive and non-cognitive 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 only allow a local "certificate of completion." For reasons perhaps better understood by psychologists than economists, such terminological distinctions seem to be empirically quite 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" vs. "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. Of course, there will remain 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 we do not find such sorting arguments persuasive.

Finally, we turn from those near the top and those near the bottom to our final group of students, 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 are students who would choose to meet a lesser level of achievement that has a lesser payoff, but not as dramatically so 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. That is because the labor market signals to students are somewhat remote, and also because many students are notoriously present-oriented.[34] It

is also likely that schools 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.

This brings us to one of the key policy dilemmas that comes out of our theoretical analysis: how much differentiation should there be 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.

We have reached the limits of our theoretical analysis. We believe it 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 tradeoff 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 tradeoff 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 non-cognitive skills. The optimal degree of differentiation among these credentials can probably only be worked out in practice over time, by trial and error, since it depends very much on the way employers

will treat different credentials, which is not something that is easily foretold.

A Description of Current State Educational Standards

We now turn our attention to how in the United States educational standards have been defined in practice, with a focus on the variations among states. Effective educational standards require the following three components:

- Content or curriculum standards that clearly delineate what students should learn in each grade.
- An assessment system that measures student progress toward mastery of the content standards.
- An accountability system that stipulates a set of rewards and/or interventions based on student progress. Such a system should hold not only students but also teachers, principals, and entire school systems accountable 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 the fact that even though standards in practice typically resemble the binary "pass-fail" model discussed earlier, these standards have taken many forms. Some states have implemented high school exit exams. Other states have left the task of assessment to 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.

Consider first graduation standards. Throughout the 1990's states' graduation

requirements varied radically. For instance, in 1993, the number of courses states required students to complete before graduating with a standard diploma varied from 13 in California and Wisconsin to 24 in Florida and Utah. (U.S. Department of Education, 1996) By 1996, California still required only 13 courses to graduate, but Wisconsin had increased its graduation requirements from 13 to 21.5. At the top end, three states – Alabama, South Carolina and Texas – had either joined or were about to join Florida and Utah in requiring 24 courses for high school graduation. (U.S. Department of Education, 1999) [35]

These variations in course requirements become stronger once one examines the specific courses required to graduate across states. For instance, in 1996, over half of states required that high school students take at least two math courses in order to graduate. Another 15 states required 3, and two states (Alabama, South Carolina) required four courses. A number of states' requirements defy a simple categorization. Colorado, Iowa, Massachusetts[36], Michigan, Minnesota and Nebraska rely mainly on local boards to set graduation requirements. 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, in order to recognize advanced achievement. The AFT (1999) reports that currently 20 states offer advanced diplomas, up from only 8 in 1996.[37] [38] 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 earlier theoretical section of this paper suggests that 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.

Notably, in the late 1990's, New York decided to begin phasing out local diplomas in favor of requiring all students to acquire a Regents' diploma. This transition process has not yet finished. 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 relabelled. But it will be important to ascertain how employers and institutions of higher education respond to such relabelling, 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. Table 1 shows 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". 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 of the AFT's variables measures the quality of states' assessment systems, while the third measure partially describes the state's 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 reveals some fascinating patterns. By all three measures – content standards, assessments, and student accountability - the national trend is clearly toward more stringent requirements. Second, the table indicates large variation across states in these three components of educational standards and accountability.

Third, and 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 22 states have implemented content standards that the AFT deems clear and specific. Lack of clarity in standards will obviously 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 write 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.

As Table 1 shows, in 1996 almost no states based decisions to promote students to the next grade on standards, but by 1999 13 states had such policies in place. This number clearly 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, 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 (1999), 28 states currently have or plan to implement graduation exams that are aligned with state's curriculum standards.

It appears that the most difficult aspect of implementing a graduation or exit exam is to design the exam so that it links well to curriculum standards. For instance, California published science and 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 2003-2004. The strong desire among California's policymakers to implement a school-leaving exam that is

well articulated with content standards has 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 (1999) reports that 20 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. It is probably fair to say that a 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 than those who face grade retention. By the start of high school, some students are likely to view college attendance as a somewhat dim prospect. It remains to be seen what positive incentives can be created for such students, 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 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 categorized 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.) Based on the above analysis, Table 2 presents our calculations of the number of states that fit into each of eight possible categories. 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.[39] 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 , it becomes important to know what causes these variations. 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. On the other hand, the existing federally mandated National Assessment of Educational Progress data, which beginning in the 1990's began to release 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. Costrell's (1994) work suggests that 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.

The degree of socioeconomic homogeneity, and the overall socioeconomic status of the state population, may also influence standards. 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 listed in that table. 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 well-linked to the content standards. Third, in order 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 and/or exit exams aimed at grade 10 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-1990's 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 Grade 4 reading assessment on the National Assessment of Educational Progress (NAEP), the analogous percentage in the 1996 Grade 4 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 to be discussed below.[40]

The results are best conveyed graphically. Figure 2 plots the states' scores on our measure of overall quality of standards 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 quite strongly. States that in the mid-1990's 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.

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

Table 3 shows the correlation coefficients for the relationships depicted in Figures 2-3, and also for more disaggregated relationships. The table gives the correlations between the three components of our overall measure of standards, as well as their composite, on the one hand, and, on the other hand, 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. Obviously, initial student achievement and population in the state do not determine all of the variation across states in the standards that they have set, but these variables do seem to matter in important ways.

Table 3 also shows 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 aged 25 and higher who hold at least a high school diploma, and the percentage of the population living above the poverty line. [41] 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. We also note that the level of standards is more strongly related to initial student achievement than it is to our three measures of socioeconomic status. It seems that low student achievement rather than socioeconomic disadvantage has been the more important factor driving the move to higher standards.

In summary, we have documented a rise in courses required for graduation in many states in the 1990's, 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. Clearly, 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 lagged behind in the middle of the 1990's tend to have done more to implement content standards, testing and accountability. 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 that studies what happens to studentoutcomes under different sets of academic standards is small but growing. This section summarizes several unpublished and forthcoming papers that use rigorous statistical analysis, reviews a fairly large literature on the effects of grade retention, and then examines in some detail the sweeping reforms to student testing and accountability in the Chicago schools.

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. Lillard and DeCicca (forthcoming) compare high school dropout rates and

attrition rates among states in 1980 and 1990, and individual-level data from about the same times. [42] 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% 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 towards 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 (hopefully few) who will drop out in any event.

The companion paper in this volume by John 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.

A number of papers have examined the correlation between homework and test scores. Cooper (1989) provides a detailed review of earlier research on the link between homework and student achievement. [43] He cites a number of experiments, some but not

all of which suggest a positive link. However, the sample sizes in these studies are very small (39 to 400 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 non-experimental cross-sectional framework. Cooper reports the results of 11 studies that model student achievement as a function of homework while controlling for background variables. Most of the studies indicate a positive link between homework and achievement. But in some cases the research used small samples which 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. Two notable exceptions are Keith et al. (1986) and Walberg et al. (1986), 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 which the student reported doing per time period, while controlling well for prior achievement and characteristics of the school environment. [44]

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 which a school administrator or teacher can directly control. In particular, much of the variation in homework performed by students in a school 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.

Betts (1997) attempts to get around these problems by analyzing a nationally

representative sample of students attending grades 7-12. [45] Because teachers indicate the amount of homework they assign per week, it reduces the chance 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 quite robust to the addition of a dummy variable for each student to control for omitted ability or motivation among students.

The paper by 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 8 hours per week. Within this range, no 'tailing off' of the effectiveness of math homework emerged. Of course, this study, focused on math homework, cannot indicate the optimal amount of homework that schools should assign across all subjects. Second, the paper finds 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.

A separate paper by Betts (1997) examines variations across schools in math and science grading standards. [46] 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 a decided 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

The theoretical analysis in the earlier part of the paper 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 our earlier 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 additional classes after school, on weekends, or in summer school. Notably, 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, Holmes (1989) reports that grade retention is typically associated with poorer student performance after the student is held back a year. Only nine of 63 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 quite intensive remediation. It appears that 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.[47] [48] The Chicago Public Schools system has received national attention for a bold program called Summer Bridge. As reported by Betts (1998), beginning in the 1996-97 school year, students in Grades 3, 6, 8 and 9, students 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 3, 5.2 for Grade 6, 6.8 for Grade 8 and 7.9 for Grade 9. (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 Grade 3 school year.) At the end of 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, fully 27.1-62.2% 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 actually wrote the summer tests, the success rate at the end of summer ranged from 38.4% to 49.6%, with the highest success rate among Grade 8 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.

Roderick and others (1999) present the results of a two-year study of Chicago students. [49] 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 re-emerged during the1997-98 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 3 were fairly muted. However, the percentage of students making the grade cutoffs during spring testing increased considerably between 1995 and 1997 in Grades 6 and 8. 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 and/or teacher effort, at least in Grades 6 and 8. Table 4 reproduces results for the reading test in Grade 6. It shows the percentage of students in various categories who met the Grade 6 reading cutoff at stated times. Students were divided into groups based on how many grade equivalents they would need to gain during Grade 6 in order to reach the stipulated cutoff. We show the results for

students who needed to gain at least some positive fraction of a grade equivalent by May of their year in Grade 6 to be promoted to Grade 7. 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 table shows 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% met the cutoff by May of the following year in 1995, compared to 31% in 1997. Because these two groups of students had similar initial achievement, the 11% gain suggests that the replacement of "social promotion" with strict grade promotion policy in the 1996-97 school year induced very strong incentive effects. Weaker incentive effects are apparent among students whose initial grade equivalents were higher, as shown in the table.

Table 4 also makes 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 re-emerged over time.

We cannot be sure whether the apparent incentive effects derive from greater

effort among students, teachers, or parents of at-risk children, or all three. 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 rather strong incentive effects related to the raising of standards, as posited in the theoretical review section of this paper. Our theoretical analysis suggested that we need to consider four groups of students who are at risk of failing. In order of increasing achievement, these groups are: first, those at the very bottom who exerted no effort with or without the new standard; second, slightly more highly achieving students who reduce their effort after the standard is raised because they believe that they can't meet the new cutoff; third, students who do not change their effort, and fail under the new system, and fourth, 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 to 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.[50] Indeed, students who had to improve by more than 1.5 grade equivalents showed the strongest improvement relative to similarly weak achievers who entered Grade 6 before the standard was raised.

Surely, we must exercise caution 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 (1997) 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

Background

The 1993 Massachusetts Education Reform Act (MERA) established two prongs in a 7-10 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 in order to bring all localities up to a newly formulated foundation budget by 2000.[51] Real state aid more than doubled over this period.[52] The annual growth rate of state aid in current dollars averaged 12.4%, exceeding inflation plus enrollment growth by 7.7%.

As a result, all districts were successfully brought up to foundation budget, and the gaps in spending were markedly narrowed. 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 10th percentile (i.e. low-spending districts) rose $862 (in 1999 dollars) from 1993-98, and by $449 at the 90th percentile, due to a combination of local and state funding.[53] 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 10th-grade proficiency in the core subjects.

Both prongs of MERA were essential to the broad, bipartisan consensus among the Democratic Legislature, Republican Governor, and the press and public, in an otherwise rather politicized state. It is important to note 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 wherewithal to localities, but also strengthens the backbone of public officials for phase two: they are now committed to follow through on accountability measures in order 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 due to 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.[54] This delayed the development of some of the exams in the Massachusetts Comprehensive Assessment System (MCAS), since they are specifically aligned with the state frameworks. Unlike some states, which have taken 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 4, 8, and 10, 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) will 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, rather than the originally intended cutoff at "Proficient," since the initial 10th grade failure rates exceeded 50%.[55] Students will have at least four opportunities to retake the tests before the end of 12th grade.

Both math and ELA exams include sizeable open-response and/or 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), 4 open-response questions and 32 multiple choice. The Spring 1999 4th and 10th 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 actual 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 not used again. This greatly reduces the problem of artificial test-inflation

over time as the questions on existing forms become more widely known.[57] This raises the cost of testing, but at about $15/head, it is still cheaper than AP and SAT exams.

Early Test Results

The 1998 and 1999 failure rates were quite high on math in grade 8 (over 40%) and grade 10 (over 50%), as well as grade 10 ELA (about 30%). The failure rates are much higher in most of the urban districts (over 75% in Boston and over 80% in Springfield). Moreover, the 10th grade scores did not improve in the 2nd year of the test. Two math examples illustrate some of the range in level of performance:

(1998, grade 8) According to the 1990 census, the population of Massachusetts was 6,016,425. Approximately what percent of those people lived in Boston?

Population of Cities in Massachusetts	
City	**Population**
Boston	574,283
Cambridge	95,802
Fall River	92,703

A. 10%

B. 20%

C. 30%

D. 40%

Only 28% of Massachusetts' 8th-graders answered correctly, barely more than the 24% that would obtain if those who answered the question guessed randomly.[58] This was a particularly low-scoring question, but performance on the following question was slightly better than most:

(1999, Grade 10) 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% got it right. Other questions were harder, primarily because they demand students know how to apply mathematical concepts, including multi-step 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 non-urban students who failed one or both 10th grade exams.[59] 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 quite independent of the MCAS. It seems unlikely that MCAS would have negative incentive effects on such students once it starts to count, and may well 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% of the failing urban students in this sample who answered no multiple choice questions at all, and 20-23% 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% 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 11 and 12.[61] Also, about 20% of the students who failed the math exam are special

education students, some of whom will be eligible for test-taking accommodations and/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 10th grade math exam were enrolled in remedial/basic math or algebra 1, so they have not been taught much of the 10th-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 9th grade at the latest.

In short, there is good reason to believe that the failure rates will be substantially lower once the exam starts to count, but they still threaten to be quite high on the math portion. Consequently a full array of remedial measures are 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 due to 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 4th grade MCAS exams have already shown improvement in the second year of testing. We now turn to some econometric evidence on ELA-4, which suggests that these improvements were larger than the raw data indicate, and appear to reach back into 3rd grade as well.

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

In the second year of the MCAS, 1999, the mean score on ELA-4 rose approximately 3.5 percentiles, and the median score rose 4.2 percentiles over the scores of the previous cohort. As always, the question arises as to how much of this improvement was due to 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 4. Fortunately, the Massachusetts DOE has assembled a very useful micro data set that allows one to answer this question for the ELA-4. The state required all school districts to administer the 3rd-grade 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 3rd-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 byproduct 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 4th-grade learning.

More rigorous statistical analysis bears this out. The DOE has linked the 3rd-grade reading scores with the 4th-grade MCAS ELA scores for over 2/3 of the state's 75,000 4th-graders, in order to validate the MCAS exam. The ITBS score accounts for 56% 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 1,000 schools in the sample, for MCAS scores of 1998 and also for 1999. This allows us to decompose the mean gain in MCAS scores into that part which is due to 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 about 5 percentiles (vs. 3.5 in the raw data).

We take the analysis a few steps further, in order to shed some light on whether the improvement in MCAS scores represented a superficial test-specific improvement, 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, LEP, and free lunch status (but without a prior test score to control), we find that ITBS scores improved quite dramatically from 1998 to 1999, despite an adverse cohort effect. The underlying improvement in mean ITBS scores was over 8 percentiles, after correcting for the cohort effect.[63]

Was it merely a coincidence that 3rd-grade ITBS scores rose dramatically the same year that 4th-grade MCAS scores rose by 5 percentiles? If both events reflect improved practices and/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 we have found some suggestive circumstantial econometric evidence. Roughly speaking, schools that added more to 1999 student performance on their 3rd 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 4th grade students' MCAS scores than would have been predicted.[64] This is consistent with, though it does not prove, the hypothesis that those schools which were stimulated most to action by the introduction of MCAS were likely to have made improvements in 3rd grade reading instruction as well as 4th 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 rather small (about 200-300 students). Students, of course, are by tradition adverse to exams[65], so the more important question is why some adults are encouraging them.

Objections fall into several categories. The protestors (and groups such as FairTest and the ACLU) 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. As has been widely noted, the opposition is "mostly in the affluent suburbs west of Boston and in pockets of progressivism like Cambridge."[66] With a few exceptions (such as the local NAACP), 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, since they already knew the general message MCAS was bearing.

A disproportionate number of the teacher opponents to MCAS come 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 their 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 1-7 will spend more than 5 hours/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). It seems that the objection is not really so much to a single test, but rather to a set of external common assessments vs. a set of local and possibly idiosyncratic criteria.

The Massachusetts Teachers Association (state affiliate of the NEA) has also joined in opposition to the MCAS. The MTA recently announced its intention to file legislation to eliminate the MCAS graduation requirement.[67] Two months later the MTA 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.[68]

What seems to be at issue here is that the MCAS is the key component in the accountability phase of Massachusetts 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 leading up to this juncture 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.[69]

The ways in which these 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 re-organization of the school day, revamped professional development, etc.[70]

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 30 parent and community groups, such as the Urban League and the Black Ministerial Alliance, have 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."[71]

One cannot help but noting 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 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, in this section we outline four key obstacles that can stand in the way of higher educational standards. These obstacles are: 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. The theoretical section of this paper demonstrates that any change in standards typically leaves some students worse off. This makes the politics of higher standards inherently divisive. As an earlier section made clear, 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 actual performance of their children, opposition could swell.

Indeed, many parents and legislators might be surprised to learn just how much variation there is in student performance at present. Figure 4 shows the 25th through 75th 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). 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 to 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. [72] In other words, what percentage of Grade 9 students would be held back if their math scores were below the national average for students in grade 8 or even grade 7? The predicted percentage of students who would be held back if their achievement lagged by a year ranged from 37-46%, depending on grade, in grades 8 to 12. If instead students were retained only if their scores lagged national norms by two years, then 26-40% of students would have been retained. These are very large shares of the student population.

Of course, these 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 and teaching practices. The evidence cited earlier from the Chicago Public Schools suggests that the development of standards, testing and accountability can indeed 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.[73]

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 rather from families in areas served by good schools. (Recall our earlier evidence that in Massachusetts and Wisconsin, at least, the most vocal opposition to tighter standards has come from rather affluent communities.) Parents in more successful schools may fear that districts will shift resources from their schools to under-performing schools in the district. Clearly, 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, it makes sense to implement higher standards at a time when state budgets make higher funding a real possibility. Massachusetts appears to have followed this policy prescription quite closely.

Some affluent parents might worry that higher standards will make it more difficult to "stand out from the crowd" 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, etc.) 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, to create incentives for a wider range of students to excel. If a multi-tiered 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 multi-tiered 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 1990's. 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.[74] Similarly, when California first attempted to develop science standards, two rival groups, one led by Nobel

Prizewinning 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. [75]

Clearly, the care and attention to detail that is 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 to the test.

Second, writing tests that provide both in-depth and sufficiently wide coverage of a subject creates challenges.[76] The solution would appear to be to lengthen existing test instruments in order 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. On the other hand, broadening the test then evokes the complaint that it is too long, diverting student time from other learning activities. The fact that it is often the same critics who object to a "single test" being used for high stakes and also object to

the length of a multi-faceted set of exams indicates the objections are not being quite accurately framed; it seems likely that it is the external nature of the assessments that 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. Koretz (1996) summarizes earlier work he conducted with co-authors 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. [77] 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: 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 co-authors 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 subsequent '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 towards 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). [78] 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 re-assigned. 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. It seems clear that 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 aforementioned 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 rather weak compared to 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 Murnane and others (1991) 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. [79] 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 perhaps most notably the city of Denver, are beginning to experiment with rewards for teachers based on the rate of progress of their students. [80]

Clearly, 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. Indeed, during the 1990's a movement for what became known as "opportunity to learn standards" argued forcefully for equalization of school spending before implementing student accountability.[81]

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.[82] But we think it important that the public 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. [83] Similarly, the link between school resources and longer-term measures of student outcomes, such as educational attainment and wages, is modest. [84]

Consider for example Betts, Rueben and Danenberg (2000), who analyze the distribution of school resources and test scores on a school-by-school basis in California. The authors find strong inequalities in teacher preparation among schools (even within the same district), with lower socio-economic status (SES) 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% of teachers hold no more education than a Bachelor's degree, compared to only 8.8% 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 rather small. Figure 5 shows the predicted effects on the percentage of students scoring at or above national norms in reading when a school moves from the 25th to the 50th and then the 75th percentile in a number of school resources. 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 quite 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. Is seems plausible that the creation of uniform educational standards could provide the incentive to improve student performance in a way that spending hikes alone cannot. Indeed, 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 together produce meaningful gains in achievement.

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

The preceding 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, providing sanctuary for students from these schools in other schools or in new charter schools.
- Potentially harsh tradeoffs 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 remediated to reach stipulated levels of cognitive skills, credentials need to be developed to signal important non-cognitive skills. These credentials, such as certificates of completion, should be sufficiently differentiated from cognitive credentials to maintain the incentive to acquire cognitive skills.

Of course, no such list of recommendations can fully anticipate what will work and what will not work as we move to full-blown standards-based reform. 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 12. For others, alternative settings will be developed, such as the 9th-grade remedial schools in Chicago. Proposals have been made in Massachusetts for the community colleges to admit students into special non-degree 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 precisely known, of one thing we can be sure: it would be a disservice to all too many high school graduates to continue granting diplomas that provide no guarantee of minimal literacy and numeracy skills. Amid all the rising controversy it is a remarkable 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 districts whose graduates are known to often lack basic skills, and whose communities have been notably absent from the protests against standards-based reform.

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

Year	Clear Specific Standards	Assessments Aligned with Standards	Promotion Policies Based on Standards
1995	13	33	Not available
1996	15	42	3
1997	17	46	7
1998	19	47	7
1999	22	49	13

Notes: Source: American Federation of Teachers, 1999.
The counts include the District of Columbia and Puerto Rico.

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

Clear Standards All Core Subjects	Assessments Aligned with Standards in All Core Subjects	Promotion or Exit Policies	Number of States
yes	yes	yes	5
no	yes	yes	12
yes	no	yes	1
yes	yes	no	9
no	no	yes	4
yes	no	no	3
no	yes	no	11
no	no	no	7

Notes: Source: Calculated from data in American Federation of Teachers, 1999. The counts include the District of Columbia and Puerto Rico.

Table 3 Correlation Coefficients between Measures of Quality of State Standards, and Measures of Student Achievement and State Population

	Content Standards	Assessments	Accountability	Overall Standards
Math % at Basic, 1996	-0.31	-0.22	-0.49	-0.46
Reading % at Basic, 1994	-0.32	-0.19	-0.55	-0.47
Average % at Basic	-0.32	-0.19	-0.54	-0.47
Natural Log Population	0.28	0.27	0.31	0.42
% Population White Non-Hispanic, 1997	-0.16	-0.04	-0.49	-0.31
% with High School Diploma or Higher, Aged 25 and Above	-0.19	-0.32	-0.22	-0.37
% of Population above Poverty Level	-0.20	-0.12	-0.21	-0.25

Note: See text for definition of variables. Source: Authors' calculations based on AFT data on standards, NAEP test scores, and Bureau of the Census demographic estimates. In a small number of cases, only one test score was available, in which case the average % of students at or above basic levels was set using the one available test score.

Table 4 The Percentage of Grade 6 Students Meeting Reading Test Score Cutoff in 1995 and 1997 in Chicago Public Schools by Number of Grade Equivalents Behind in Previous Year

Initial Number of of Grade Equivalents (G.E.) Behind	May 1995	May 1997	After Summer Bridge, August 1997
> 1.5	20	31	52
1.5 to 1	36	43	65
1 to 0.5	50	57	79
0 to 0.5	65	71	88

Source: Roderick and others, "Ending Social Promotion: Results from the First Two Years," page 27.

Figure 1: Incentive Effects of a Rise in Standards, Across Productivity Levels

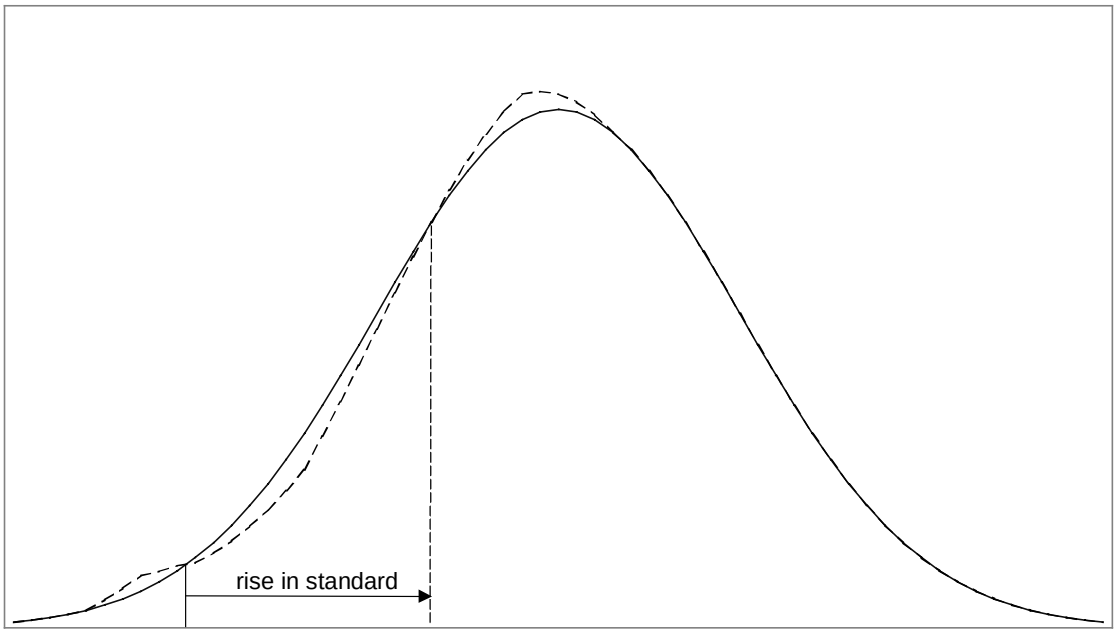


Figure 2

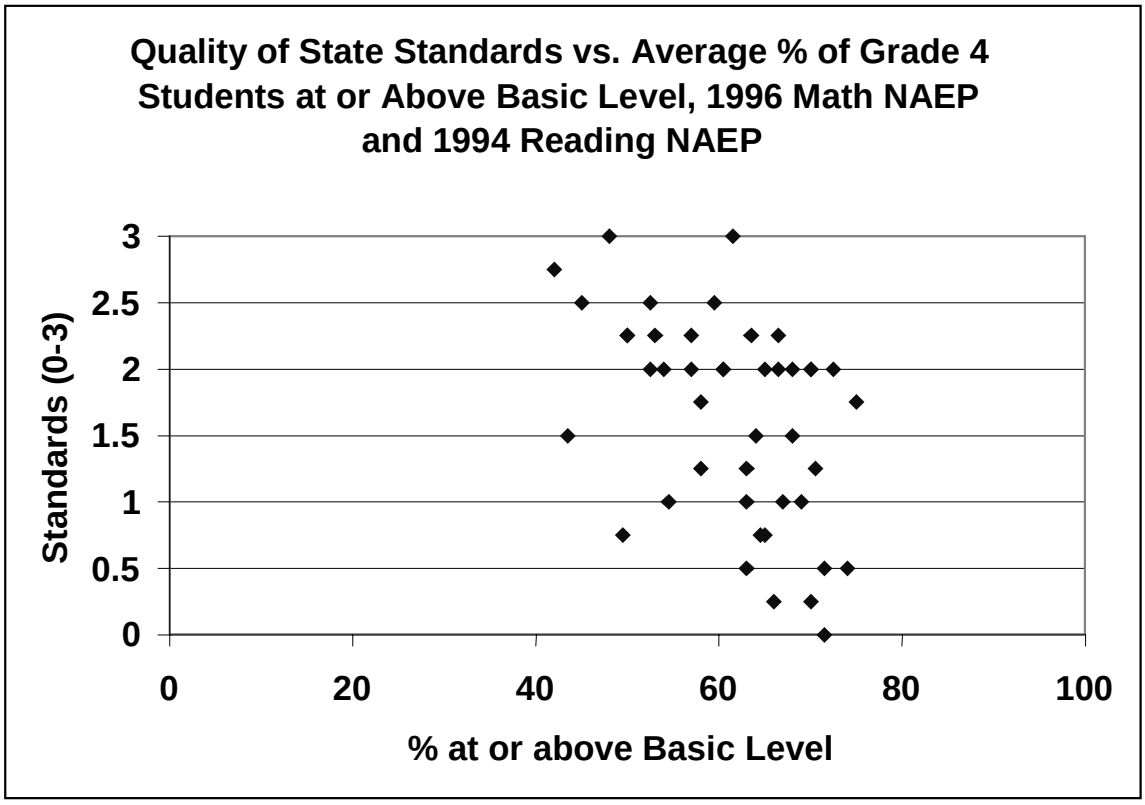

Quality of State Standards vs. Average % of Grade 4 Students at or Above Basic Level, 1996 Math NAEP and 1994 Reading NAEP
Standards (0-3)
3
2.5
2
1.5
1
0.5
0
0
20
40
60
80
100
% at or above Basic Level

Figure 3

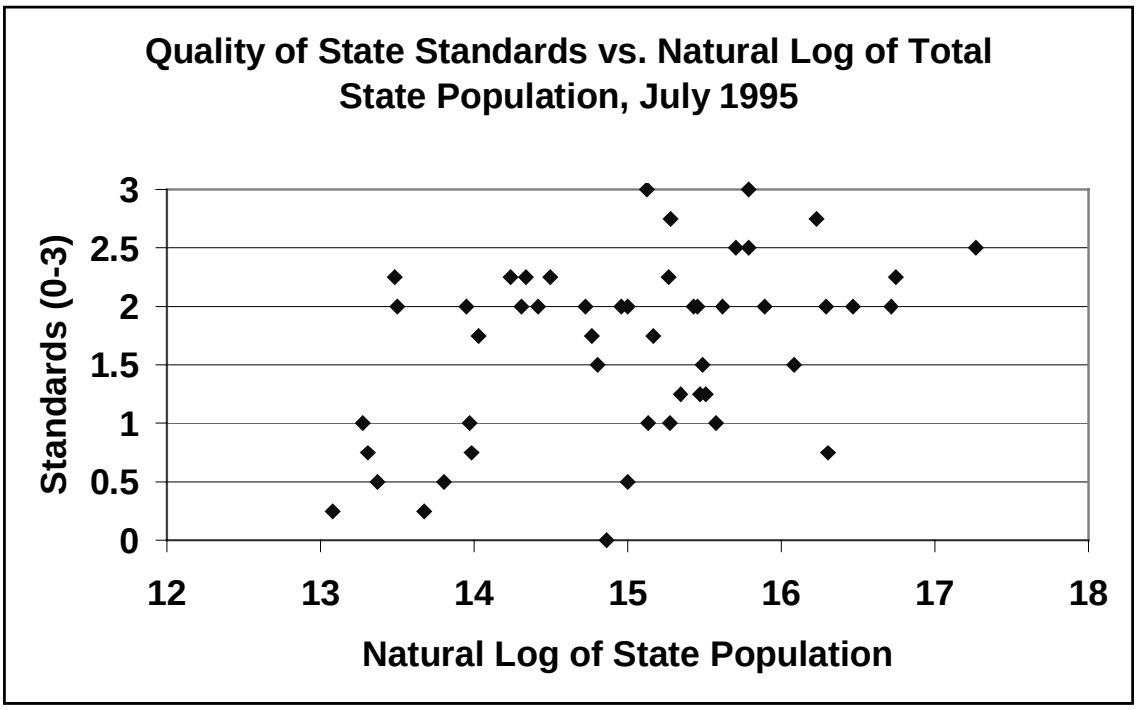
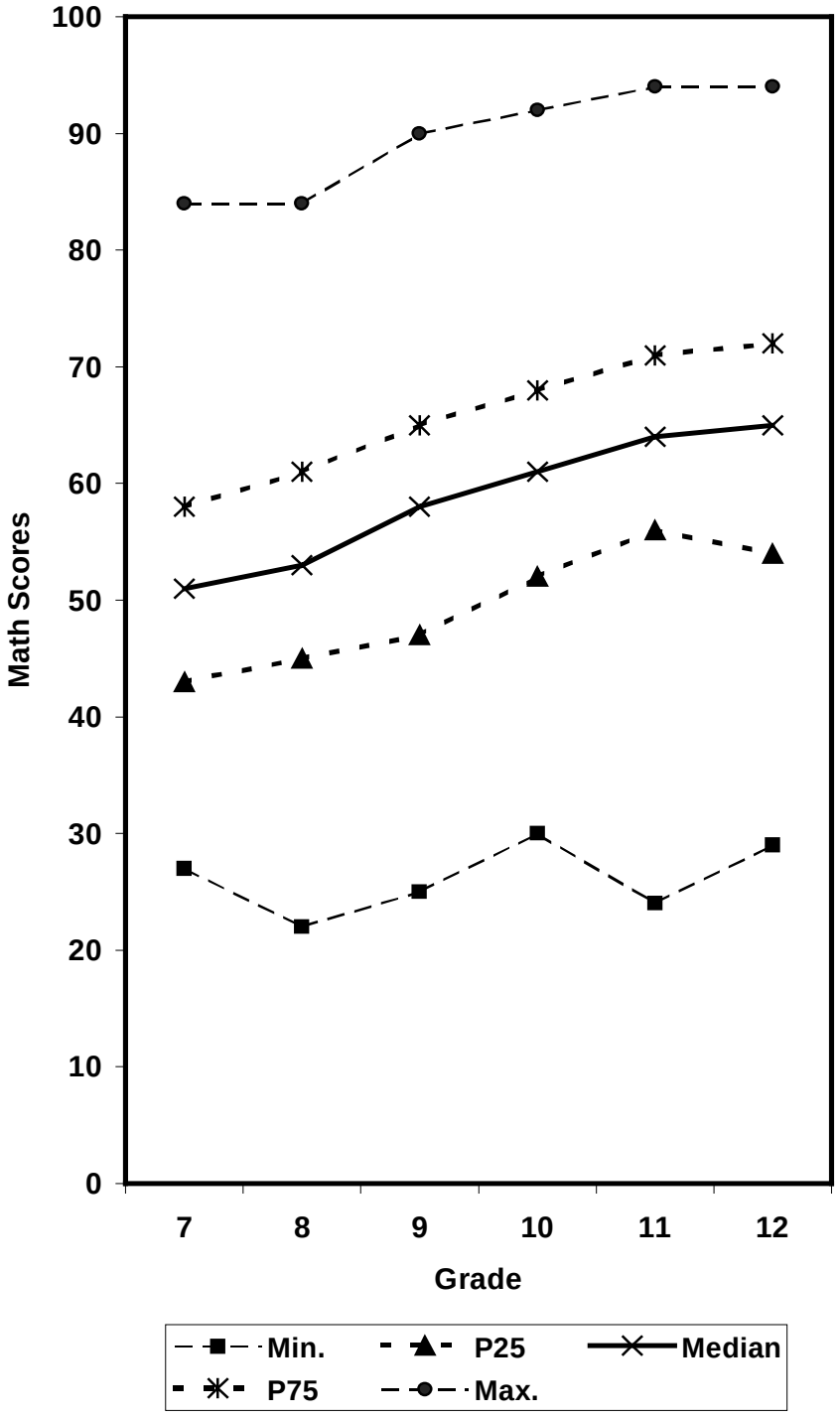
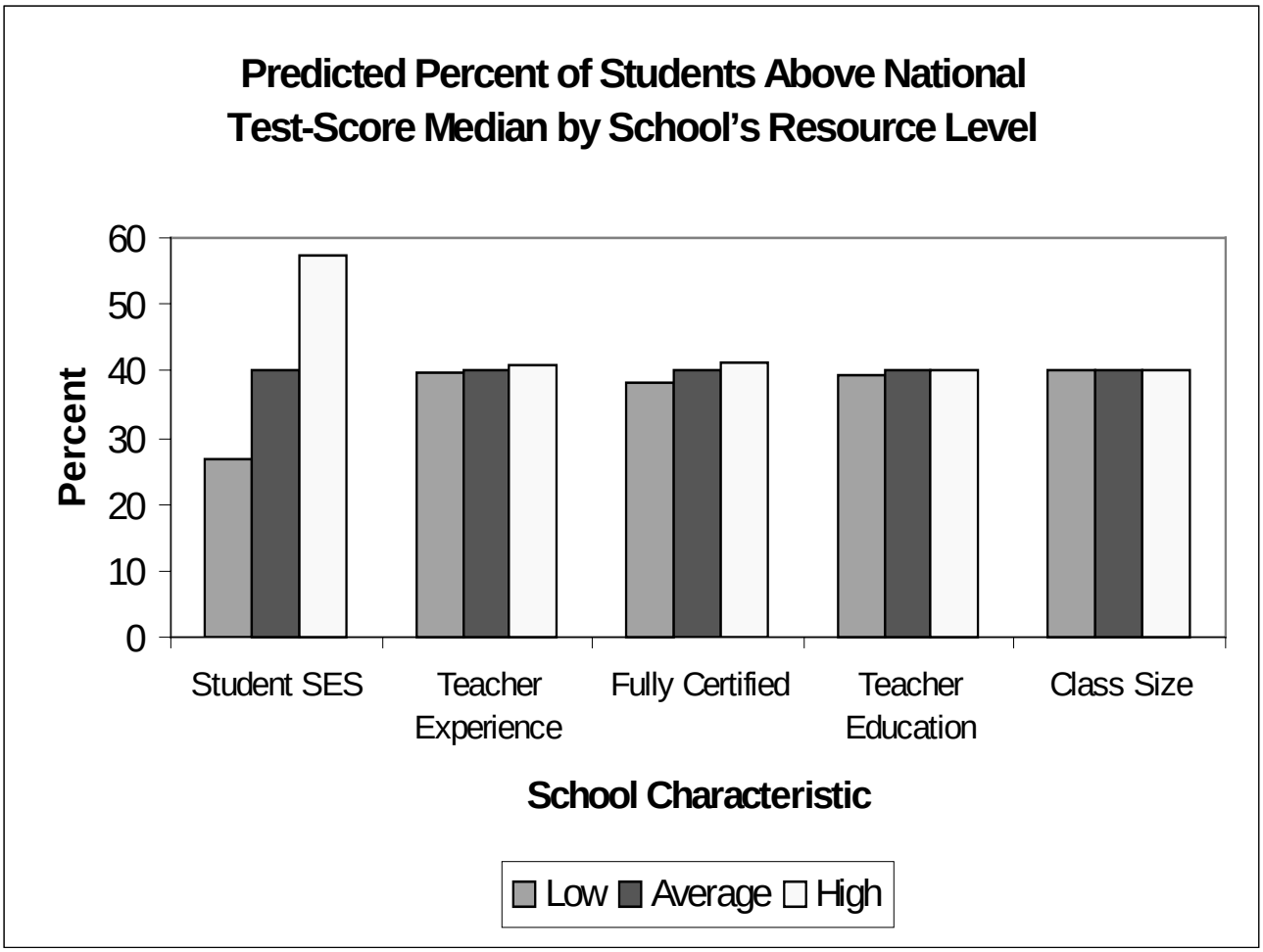

Quality of State Standards vs. Natural Log of Total State Population, July 1995
Standards (0-3)
3
2.5
2
1.5
1
0.5
0
12
13
14
15
16
17
18
Natural Log of State Population

Figure 4 The Distribution of Test Scores by Grade, 1987-1992.



Source: Julian R. Betts, "The Two-Legged Stool: The Neglected Role of Educational Standards in Improving America's Public Schools". Data: The Longitudinal Study of American Youth.

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



Source: Betts, Rueben and Danenberg, *Equal Resources, Equal Outcomes? The Distribution of School Resources and Student Achievement in California*.

Endnotes

[1] See Public Agenda polls in recent years.

[2] Mass Insight Education poll, November 1999.

[3] See Sandra Stotsky (ed.), **What's At Stake in the K-12 Standards Wars: A Primer for Educational Policy-Makers**, Peter Lang Publishers, New York, 2000. Authors such as Stephen Arons (**Short Road to Chaos**, University of Massachusetts Press, 1997) 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 thorough-going system of school choice and vouchers. 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.

[4] Robert M. Costrell, "A Simple Model of Educational Standards," *American Economic Review* 84 (4), 1994, 956-971; Julian R. Betts, "The Impact of Educational Standards on the Level and Distribution of Earnings," *American Economic Review* 88 (1), 1998, 266-275. These papers, and others cited below, provide the formal models underlying the summaries given in the text.

[5] A century ago, when a high school diploma was held by a small minority of the population, there was far less stigma attached, economically or otherwise, to being a non-graduate. Similarly, under the traditional British system that prevailed until very recently, many students left school at age 16. Far more students left school at this age than occurs in the U.S., and the stigma was presumably much less, since their numbers included more capable workers.

[6] Indeed, 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 100% pass rate, the strategy of a very low standard will surely lead to the least egalitarian outcome, by the Rawlsian standard, since the rare failure is most highly stigmatized. In short, the wage of failers rises monotonically with the standard in this simple model. See Betts (1998).

Costrell ("Are High Standards Good or Bad for those who Fail?" University of Massachusetts at Amherst, Department of Economics, 1999) 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 Costrell and Glenn C. Loury ("Distribution of Ability and Earnings in a Job Assignment Model," University of Massachusetts at Amherst, and Boston University, 2000). Then it can be shown that there is another effect of raising standards which works in the opposite direction from the pooling effect discussed above. 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 (1999) finds that in a benchmark case, the relationship between the wage of failers and the standard is U-shaped, and, moreover, the standard which *minimizes* the failers' wage actually *maximizes* output. The paper 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 where a rise in the standard reduces the wage of failers are also the cases where equity is most likely advanced by moving away from pass-fail systems altogether, toward fuller information. This is discussed further, below in the text.

[7] See Costrell (1994), and other literature cited there.

[8] In addition, schools facing the prospect of higher failure rates would also respond with interventions to assist "at-risk" students. We will discuss such interventions later in the paper.

[9] It is, of course, an empirical matter of some importance *how much* less the rise would be, whether it would be closer to the full 10-point rise, or closer to zero. The empirical section will use nationally representative data to document how many students are likely to fail under a number of scenarios for grade retention, under the naïve assumption that student effort does not respond to the change in standards for grade promotion. We will also present limited evidence from Chicago about how student and school effort responds to a rise in standards.

[10] See Costrell ("An Economic Analysis of College Admission Standards," *Education Economics* 1 (3),1993, 227-241) 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 actually 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 it may be the case that the optimal dropout rate is not zero. Disruptive students provide an obvious example that is unfortunately not as rare as one might hope. Of course, the best solution in such cases is not necessarily to encourage dropouts, but rather to create alternative educational settings for such students, such as those under creation by many systems such as 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 Bishop, et. al. They find that among C/C- students, minimum competency exams raise both the number of non-completers and the number of college attendees.

[13] Although the general points discussed here and depicted in Figure 1 derive from the theoretical literature cited above, 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.

[14] "Lawmakers seek to limit standard tests," Anjetta McQueen, Associate Press, April 5, 2000, as published in the *Boston Globe*, p. A16.

[15] Andreae Downs, "Parents, educators debate MCAS," *Boston Globe*, February 13, 2000.

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

[17] J.E. Jacobsen found some evidence of this as a result of state "minimum competency" tests in the last 70s and early 80s ("Mandatory Testing Requirements and Pupil Achievement," 1992, mimeo, M.I.T.) For classroom-based evidence that teachers devote more attention to the lowest-achieving students in class, see Brown, B. W., and D. H. Saks, "The Microeconomics of the Allocation of Teachers' Time and Student Learning," *Economics of Education Review* (1987) 6:319-32, and Julian R. Betts and Jamie L. Shkolnik, "The Behavioral Effects of Variations in Class Size: The Case of Math Teachers", *Educational Evaluation and Policy Analysis*, (Summer, 1999) (**20**:2), pp. 193-213, who show that reductions in class size lead teachers to spend more time on review and individual instruction, ostensibly directed toward the lowest-achieving students.

[18] To be sure, this does not prevent some of the critics in these "progressive" 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.

[19] For a contemporary account of the national standard-setting movement, see Chapters 2 and 5 of Diane Ravitch *National Standards in American Education: A Citizen's Guide.* (Washington, D.C.: The Brookings Institution, 1995), and for more of a retrospective, see Robert B. Schwartz and Marian A. Robinson, "Goals 2000 and the Standards Movement," *Brookings Papers on Education Policy 2000*, The Brookings Institution, 2000.

[20] See Costrell (1994), Section IV.

[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] Note that 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 there is no systematic difference between local and central authorities regarding the weights attached to winners and losers (i.e. they hold the same "social welfare function").

[24] Indeed, 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 non-egalitarians. 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 non-egalitarians favor centralization if all districts are alike, but under cross-district heterogeneity, egalitarians may favor centralization in some cases that non-egalitarians do not.

[25] Robert M. Costrell, "Can Centralized Educational Standards Raise Welfare?" *Journal of Public Economics* 65, September 1997, 271-293.

[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 to any alternative.

[27] Indeed, we cannot even be sure 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 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 in fact the law in Massachusetts: no district will be able to award a diploma to students who fail the MCAS, but districts can impose additional graduation requirements, including a higher MCAS score.

[29] Costrell and Loury (2000), applied to the issue of standards by Costrell (1999).

[30] For a formal analysis, see Costrell (1994), Section VI.

[31] John D. Owen, *Why Our Kids Don't Study: An Economist's Perspective*, Johns Hopkins University Press, Baltimore, 1995.

[32] See James Heckman, "Doing it right: job training and education," *The Public Interest*, (135) Spring 1999, 86-107.

[33] There is a considerable econometric literature on this point, beginning with Stephen V. Cameron and James J. Heckman, "The Nonequivalence of High School Equivalents," *Journal of Labor Economics* 11 (1), 1993, 1-47.

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

[35] U.S. Department of Education. Digest of Education Statistics 1996. (Washington, D.C.: National Center for Education Statistics, 1996), and U.S. Department of Education. Digest of Education Statistics 1998. (Washington, D.C.: National Center for Education Statistics, 1999).

[36] This will change dramatically, beginning with the class of 2003, as discussed below.

[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 added credibility. The following summary will draw heavily from these AFT analyses.

[38] Data for 1999 and 1996 are from American Federation of Teachers, *Making Standards Matter 1999* and *Making Standards Matter 1996* (Washington, D.C.: American Federation of Teachers, 1996) respectively.

[39] 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] Our sources for the math, reading and population data are, respectively, Clyde M. Reese and others, *NAEP 1996 Mathematics Report Card for the Nation and the States: Findings from the National Assessment of Educational Progress* (National Center for Education Statistics, 1997), Jay R. Campbell and others, *NAEP 1994 Reading Report Card for the Nation and the States: Findings from the National Assessment of Educational Progress and Trial State Assessment* (National Center for Education Statistics, 1996), and *State Population Estimates: Annual Time Series, July 1, 1980 to July 1, 1999* (U.S. Bureau of the Census, ST-99-3, 1999).

[41] These three variables were obtained from pages 34, 169 and 479 respectively of U.S Bureau of the Census, *Statistical Abstract of the United States: 1998, 118th Edition* (Government Printing Office, 1998).

[42] Dean R. Lillard and Philip P. DeCicca, " Higher Standards, More Dropouts? Evidence Within and Across Time," Economics of Education Review, (forthcoming).

[43] Harris Cooper, *Homework*. (New York, NY: Longman, 1989).

[44] See Timothy Z. Keith and others, "Parental Involvement, Homework, and TV Time: Direct and Indirect Effects on High School Achievement." *Journal of Educational Psychology*, vol. 78 (October 1986), pp.

373-80, and Herbert J. Walberg, Barry J. Fraser, and Wayne W. Welch, "A Test of a Model of Educational Productivity among Senior High School Students." *Journal of Educational Research*, vol. 79 (January/February 1986): pp. 133-39.

[45] Julian R. Betts, "The Role of Homework in Improving School Quality," University of California, San Diego, Department of Economics, 1997.

[46] Julian R. Betts, "Do Grading Standards Affect the Incentive to Learn?," University of California, San Diego, Department of Economics, 1997.

[47] For a review of national trends toward increased use of summer school, see Catherine Gewertz, "More Districts Add Summer Coursework," *Education Week*, June 7, 2000.

[48] The analysis in this paragraph is based on Julian R. Betts, "The Two-Legged Stool: The Neglected Role of Educational Standards in Improving America's Public Schools," *Economic Policy Review*, vol. 4 (1998), pp. 97-116.

[49] Melissa Roderick and others, *Ending Social Promotion: Results from the First Two Years* (Chicago, IL: Consortium on Chicago School Research, 1999).

[50] To be sure, we do not know what is happening among the individuals within any G.E. category. There may well 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.

[51] The other important piece of background, aside from the court case, was a large drop in state aid that occurred in Massachusetts' very 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.

[52] Compared to the pre-recession figure, real state aid grew by about one-half.

[53] Thomas J. Kane, "An Update on School Reform in Massachusetts," John F. Kennedy School of Government, Harvard University, 2000. Compared to the pre-recession year of 1989, the corresponding rise by 1998 was $513 at the 10th percentile and $165 at the 90th.

[54] See Costrell's chapter in Sandra Stotsky's edited volume (*op. cit.*), for an account of the tortuous development of the economics strand in the history and social sciences framework.

[55] John Silber, who was Chairman of the Massachusetts Board of Education up 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 National Council on Measurement in Education Annual Meetings, New Orleans, 2000.) For the 10th grade math exam, this required 24 out of 60 possible points.

[56] See http://www.doe.mass.edu/mcas/student/grade10/g10comp.html.

[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 ELA essay questions, however,

are not matrix-sampled the previous year.

[58] Fewer than 4% 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 over 100.

[60] Twenty-five of 28 superintendents interviewed for the Mass Insight 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 10th graders, for whom it does not count, and those in 9th grade, for whom it will.

[61] In Indiana, 54% of 10th graders passed the math and English exit exams on the first try that counted, in Fall of 1997, but by time that class was due to graduate, in 2000, 86% had passed both exams. The exam is pitched at a 9th 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% of students in grades 2–9 (except 4) now face mandatory summer school, to avoid grade retention (*Boston Herald*, June 7, 2000). 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."

[63] These are Massachusetts percentiles, not U.S. percentiles. Since 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 ITBS scores bypassed the lower third of the Massachusetts distribution. One possible interpretation is that efforts of 3rd grade teachers to prepare students with skills useful for 4th grade MCAS paid off only for those 3rd 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 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."

[64] 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.

[65] Not all students are, however. The Massachusetts Student Advisory Councel 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)

[66] *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.

[67] *Boston Globe*, February 18, 2000.

[68] Interestingly enough, the ads feature students cheering their SAT scores, which have risen of late in Massachusetts. The 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.)

[69] *Commonwealth Magazine*, Spring 2000, p. 46. See also, "Urban parents support MCAS tests," *Boston Herald*, April 16, 2000.

[70] Mass Insight Education, *Over the Bar*, pp. 14-15.

[71] "Coalition Targets Teacher Seniority," *Boston Globe*, May 2, 2000; see also, "Seniority at Issue in Boston Teacher Contract Talks," *Boston Globe*, May 3, 2000, which specifically mentions MCAS in this regard.

[72] Julian R. Betts, "The Two-Legged Stool: The Neglected Role of Educational Standards in Improving America's Public Schools".

[73] As an earlier note reports, a comparable percentage of Boston students in grades 2-9 are now being held back, contingent on successful summer school remediation.

[74] See for instance Anemona Hartocollis, "The New, Flexible Math Meets Parental Rebellion," *New York Times*, April 27, 2000, p A1. There was also a heated dispute in Massachusetts over these issues.

[75] See Tanya Schevitz, "State Panel Drafts Set of Tough Educational Standards," *San Francisco Chronicle*, April 3, 1998.

[76] Koretz, Daniel, "Using Student Assessments for Educational Accountability," in Eric A. Hanushek and Dale W. Jorgenson, eds., *Improving America's Schools: The Role of Incentives* (Washington, D.C.: National Academy Press, 1996).

[77] Koretz, Daniel, "Using Student Assessments for Educational Accountability."

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

[79] Richard J. Murnane and others, *Who Will Teach? Policies that Matter.* (Cambridge, MA: Harvard University Press, 1991).

[80] See "Pay-for-Performance: An Issue Brief for Business Leaders," (The Business Roundtable, 2000).

[81] See Chapter 5 of Ravitch, "National Standards in American Education," for a summary and critique of this movement.

[82] But see Caroline Minter Hoxby, "Are Efficiency and Equity in School Finance Substitutes or Complements?," **Journal-of-Economic-Perspectives** 10(4), Fall 1996, pages 51-72, for an analysis of the pitfalls in moving too far toward state finance of local education, as a result of equalization suits.

[83] See Eric A. Hanushek, "School Resources and Student Performance," in Gary Burtless, ed., *Does Money Matter? The Effect of School Resources on Student Achievement and Adult Success*, (Brookings Institution, 1996), and Gary Burtless, "Introduction and Summary," in the same volume.

[84] See Julian R. Betts, "Does School Quality Matter? Evidence from the National Longitudinal Survey of Youth," *Review of Economics and Statistics*, vol. 77 (1995), pp. 231-50, and for a review of the literature Julian R. Betts, Is There a Link between School Inputs and Earnings? Fresh Scrutiny of an Old Literature," in Gary Burtless, ed., *Does Money Matter? The Effect of School Resources on Student Achievement and Adult Success*, (Brookings Institution, 1996).