Discussion

What Is the Appropriate Role for Student Achievement Standards?

David N. Figlio*

Politicians and policymakers across the political spectrum advocate high standards as a means of evaluating students. However, there exists very little published evidence that student achievement standards, such as the curriculum-based external exit exam (CBEEE) systems described by Bishop, lead to the substantial performance gains that advocates argue should occur. His paper takes an ambitious step in this direction. Bishop describes compellingly how a system of standards could change the culture of a school and its student body, then presents a series of empirical exercises in which he shows that, in cross-section, countries with CBEEE systems have higher performance than do countries without these systems. Canadian provinces with CBEEEs also do better than those without, and New York State, with its Regents exams, has a higher performance than might be predicted in the absence of its CBEEEs. Of these analyses, the New York analysis is the least plausible, because it relies on the presumption that New York is observationally equivalent to the rest of the country (holding observables equal) save for the Regents Exam, and the Canadian analyses are by far the most believable because the assumption of *ceteris paribus* is most likely met with detailed background characteristics controlled for and a single national educational system and infrastructure controlled for.

Each of these analyses points in the direction of standards leading to substantial improvements in student test scores in the tested subjects, and the Canadian analysis presents evidence of a series of mechanisms through which these standards might work. For instance, Bishop shows that schools in Canadian provinces with CBEEEs tend to focus instruc-

*Walter Matherly Professor of Economics, University of Florida.

tional resources and homework time on the evaluated subjects (suggesting that schools respond to the standards) and that students and families in provinces with CBEEEs apparently alter their behavior, watching less television (but more science shows), talking more about coursework, reading for fun, and changing attitudes about mathematics and science, as examples of suggested behavioral modifications resulting from standards.

This research is consistent with the small amount of research in existence on the related topic of grading standards. Many of the mechanisms put forward by Bishop that might lead to improved student outcomes with CBEEEs also would lead to improved student outcomes when grading standards are elevated. Betts (1995) and Betts and Grogger (2000) present national cross-sectional evidence on the effects of school-level grading standards, measured by the difference between grade-point averages of students in the school and the same students' examination scores. These studies find that, on average, students perform better on examinations, attain more education, and earn more in the post-school early labor market when they attend high schools with high grading standards.

In my own research with Lucas (Figlio and Lucas 2000), we follow the same elementary school students over time and find that they learn more (measured by improvements in mathematics and reading scores) and behave better (measured by fewer serious disciplinary incidents) in a year in which they have a teacher with high grading standards than in a year in which they have a teacher with low grading standards. Both the papers by Betts and Grogger (2000) and by Figlio and Lucas (2000) demonstrate that there exist differential effects of grading standards on different types of students. Bishop's findings complement the results of this other literature nicely. Moreover, my paper on grading standards suggests that parents actually view high-standards teachers less favorably than they view teachers with softer grading standards, a finding consistent with Bishop's assertion that parents may view high standards (in the absence of a systemwide set of standards) unfavorably.

Taken together with these other results, Bishop's findings of very large, positive effects of CBEEEs on student outcomes might suggest that CBEEEs are a "silver bullet"—an educational intervention that dramatically improves performance at low cost. However, there are reasons to be skeptical of the magnitudes of the findings, if not the signs of the general relationships reported in this paper. The principal reason for concern is Bishop's identification strategy—that is, the way in which he empirically uncovers the relationship between CBEEEs and student outcomes. As it stands, it is difficult to be certain that these standards are driving the estimated results. All of Bishop's analyses rely exclusively on cross-sectional variation. This type of variation is fine if there exist no omitted variables that might be correlated with both standards and student

outcomes. But it is possible that some third variable could explain both standards and outcomes. One such case of this is that causality may be reversed, and countries (or provinces) where it is easier to meet standards, for unobserved reasons, are the governments that are most likely to impose them.

While Bishop alludes to this possibility, the issue is more substantial than its presentation in the paper. In his Canadian analysis, Bishop contends that there is little evidence that provinces with CBEEEs have higher tastes for education than those without CBEEEs. He provides evidence from Bishop (1996) suggesting that provinces with CBEEEs do not demonstrate increased tastes for education, as measured by improved discipline, attendance, or computer availability. I cannot speak to the computer access issue, but I find the results regarding discipline and attendance to be somewhat suspect. Using Florida data in the past, I have compared principal-reported measures of perceived discipline and attendance problems to actual discipline and attendance problems (as measured by administrative records). In these analyses, I have found that at best there is no correlation between principal perceptions of discipline and attendance problems and actual levels of discipline and attendance problems, and in most settings there is actually an inverse relationship between these measures. Principals in affluent schools may be more sensitive to these types of problems, and perceive even mild problems as severe, while principals in poor schools may perceive even serious problems as acceptable. (The same patterns are evident for drug problems, tardiness, teen pregnancy, and juvenile delinquency.) While Canada is obviously different from Florida, the conclusion drawn is that there is, at best, weak evidence against the presumption that provinces with CBEEE systems value education more.

But there is evidence, presented in the present paper, that seems to support the reverse causation argument. Some of the outcome variables discussed by Bishop may easily be thought of as causes of CBEEEs. For example, Bishop finds that parents in provinces with CBEEEs talk to their children more about math and science classes, and children in these provinces watch less television (but more science programming) and read for fun more. The conclusion drawn by Bishop is that these are outcomes of CBEEEs. This may certainly be the case. But it is just as likely, in my view, that these are attributes of the communities that impose CBEEEs, and thereby reflect tastes for education. While it is true that the provinces that imposed CBEEES run the gamut from the affluent west (Alberta and British Columbia) to more moderate Quebec to the poor provinces of New Brunswick (Francophone portion only) and Newfoundland, the population distribution of these provinces is such that the sample is dominated by Alberta, British Columbia, and metropolitan Quebec. In the 2001 Census, nearly three-quarters of the population of these provinces resided in Alberta, British Columbia, and the Montreal metropolitan area,

implying that the population of the CBEEE provinces is not as diverse as one might expect given their numbers.

Bishop also presents cross-sectional evidence suggesting that provinces with CBEEE systems dedicate more resources to the topics covered by CBEEEs. For instance, schools in these provinces have more math and science specialist teachers, more math and science class hours, and teachers with more math and science experience. These results suggest that CBEEEs lead to institutional behavioral changes. However, in a cross-section, it is impossible to be certain of the direction of the causality. It may be that these variables indicate that the provinces ultimately imposing CBEEEs have a greater taste for mathematics and science instruction, tastes that are reflected both in curricular emphasis and in standards-setting. If this latter explanation is true, then it may be the case that the differences in resource use (and, presumably, in outcomes) would have existed in the absence of CBEEEs. We have no way of distinguishing these two explanations, and, therefore, the paper would be strengthened considerably if some within-school, over-time variation could be exploited. While Bishop's results are plausible and compelling, they are not fully convincing, and will never be so unless one can be more certain that the identification problem is solved. This will not occur in a cross-sectional setting. Ultimately, while I believe that CBEEEs lead to higher average performance, I do not know whether the magnitudes put forward by Bishop are accurate.

Bishop's analysis looks only at the mean effects of CBEEEs. This is a necessity in his cross-national analyses, but is not necessary in the case of his Canadian research. More research needs to be done to look at the distributional consequences of CBEEEs. This is important for several reasons. First, the existing theoretical research on standards, including work by Betts (1998) and Costrell (1994), suggests that they might have differential effects on students at varying parts of the ability distribution. One can tell stories in which high achievers and low achievers could either be helped or harmed by a CBEEE. Bishop presents arguments for how these students could be helped. But high achievers who are likely to have exceeded the standard without additional effort may work less, and low achievers who are unlikely to make the standard under most circumstances may give up and work less as well. (For instance, Lillard and DeCicca, in their forthcoming article, find that high graduation coursework standards induce greater dropout rates.) I have no way of ruling out these possibilities, so the conflicting stories that can reasonably be told make the question of the distributional consequences of CBEEEs an empirical one. While the research on grading standards mentioned above suggests that few, if any, students are harmed by high grading standards, there is still evidence that high achievers may benefit more. I hope that Bishop, in his future work, will investigate whether CBEEEs

help certain types of students, or students in certain types of settings, more than other types of students.

Regardless of the magnitude of the effect of CBEEEs, the present policy environment may present challenges for their implementation. Dozens of states currently have test-based systems of school accountability—effectively, high- or medium-stakes standards for schools—and with the federal No Child Left Behind Act of 2001 these stakes are elevated nationwide. On July 1, 2002, the U.S. Department of Education deemed 8,652 Title I schools nationwide sufficiently in need of improvement that students attending them are eligible for enhanced public school choice. In most states, schools and students are evaluated on the same curriculum-based test. For instance, the Florida Comprehensive Assessment Test determines not only school rewards and sanctions, including eligibility for private school vouchers, but also student promotion in Florida. High student performance on the FCAT both ensures student promotion and helps schools earn a higher performance grade, with financial and governance ramifications for the school.

School accountability systems that evaluate schools using the same curriculum-based examination used to evaluate students may have the effect of setting student standards lower than what might ordinarily have been set. This may be even more the case under the No Child Left Behind law, where school "failure" is tied to removal of federal dollars, and hence states may prefer to sanction fewer schools than they might have in the absence of the federal law. Even before the passage of No Child Left Behind, the state of Florida postponed its planned increases in student standards in a move that was publicly speculated at the time of decision (though there is no definitive evidence of this) to be caused in part by the implications for Florida's own school accountability system. The incentives are much clearer toward setting low proficiency standards under the new federal law.

If student standards are set very low, however, one might ask whether low standards are better than no standards at all. An analogy might be made with teacher merit pay. In a recent working paper, Lawrence Kenny and I (Figlio and Kenny 2001) suggest that student performance is lower in schools that give merit pay to all or most teachers, regardless of teacher productivity, than if no merit pay is offered at all. On the other hand, offering merit pay to a small fraction of teachers tends to increase student test scores substantially in our U.S. national sample of students in schools. While this is by no means definitive, it is suggestive that low standards might be less productive than no standards. This leads one to ask whether students and schools should be evaluated on the same standard. The twin goals of student and school accountability may be met more easily if the two are uncoupled.

Bishop is to be commended for the work that he has done in assembling evidence on the effects of CBEEEs from so many different

sources and in so many different settings. His work is provocative and extremely interesting, empirical identification issues notwithstanding. His title, "What Is the Appropriate Role for Student Achievement Standards?" is a relevant question to ask with respect to standards for schools as well. Is it appropriate to use student achievement standards to evaluate schools? And if not, one must ask, If schools are evaluated on a low standard, and student standards are multi-level, to whom will the schools pay attention? The answers to these questions are difficult to know right now with the current research, but the questions must be asked as the nation embarks on its new experiments with student and school accountability.

References

Betts, Julian. 1995. "Do Grading Standards Affect the Incentive to Learn?" University of California, San Diego, working paper.

———. 1998. "The Impact of Educational Standards on the Level and Distribution of Earnings." *American Economic Review* 88 (1): 266–75.

Betts, Julian and Jeff Grogger. 2000. "The Impact of Grading Standards on Student Achievement, Educational Attainment, and Entry-Level Earnings." NBER Working Paper No. 7875 (September).

Bishop, John. 1996. "The Impact of Curriculum-Based External Examinations on School Priorities and Student Learning." *International Journal of Education Research* 23 (8): 653–752.

Costrell, Robert. 1994. "A Simple Model of Educational Standards." *American Economic Review* 84 (4): 956–71.

Figlio, David and Lawrence Kenny. 2001. "Does Teacher Merit Pay Improve Student Test Scores?" University of Florida, working paper.

Figlio, David and Maurice Lucas. 2000. "Do High Grading Standards Affect Student Performance?" NBER Working Paper No. 7985 (October).

Lillard, Dean and Philip DeCicca. Forthcoming. "Higher Standards, More Dropouts? Evidence Within and Across Time." *Economics of Education Review.*

Discussion

What Is the Appropriate Role for Student Achievement Standards?

Ellen Guiney*

The Boston Plan for Excellence, a private, nonprofit organization, is a local education fund whose mission is to improve instruction and student performance in the Boston Public Schools. As such, we often work with and learn from education reform taking place in the nation's largest cities so as to bring lessons to Boston from other districts. Our knowledge base is derived primarily from studying what is happening to the students in the 35 largest cities. These students comprise 15 percent of the country's students. Generally, these students are poor, and they often begin school without preschool or other advantages enjoyed by middle-class children. A majority of them are children of color. A few statistics from our experiences are relevant:

- In a study of three Boston kindergarten classes that tested students' skills upon entering, Voices of Love and Freedom found that 60 percent of students knew fewer than 10 capital letters, 70 percent knew fewer than 10 lower-case letters, and 90 percent could make fewer than 10 letter-sound correlations.
- Nationwide, only 68 percent of all students complete high school in four years; in the 35 largest cities, fewer than 50 percent do so.
- Nationwide, half of ninth graders entering high school read at a sixth grade level.

Response to Bishop

Our experience in Boston coincides with Bishop's conclusions. Our on-site observations in 50 Boston and other schools is that curriculum-

*Executive Director, Boston Plan for Excellence.

based exit examinations do what he suggests. Tests that are aligned with and carefully measure high standards do affect a school's priorities, teachers' decisions, and students' decisions; they also influence the redirection of resources within schools to core subjects.

If we are to meet our civic and moral responsibilities as a country, however, setting standards, aligning assessments to measure whether students are learning them, and creating an aligned accountability system are only the foundation. Without the creation of a coherent system of improving instruction in classrooms—which will involve extensive professional development for principals and teachers and a deliberate reorganization of schools—urban students will not meet standards.

Examination of Underlying Assumptions

Standards-based reform rests on certain assumptions that do not hold in the 35 largest cities. These assumptions include the following:

1. The system described will give students and teachers information about what students are not learning in a timely, usable way;
2. Students will be motivated to invest more in their learning because they face consequences and because they realize how much they need to know;
3. We have the right type of classroom instruction for these students;
4. There is an adult accountability system that creates the right information, sanctions, and incentives that lead to instructional improvement;
5. There are other professionals who are better equipped, prepared, and willing to take the places of those let go because they do not succeed with students; and finally,
6. Schools are coherently organized, at scale, to respond to the standards-based foundation laid out by Bishop.

Such an accountability system rests on the idea that educators already possess all the knowledge and skills they need to bring about substantial improvements in instruction that will lead to greater learning for urban students. Further, it assumes that teachers and principals (a) need more political and civic pressure to do what is effective; and (b) are not rewarded enough by the present system to be motivated to do what they know they should.

Let's look more closely at each of these assumptions.

Assumption 1. The present system gives good information that helps teachers know the extent and depth of student learning in a timely way. Virtually no large city district has a data system that puts into the hands of teachers and principals fine-grained, user-friendly information about individual students that teachers can use on an

ongoing basis. For the most part, large-scale testing directed by the state takes place once a year, and considerable time passes before individual student results are reported. Students have usually moved on to a different grade and teacher, and sometimes to a different school. Few districts have a "formative" system to supplement these summative tests, and even if they do, management of these data is a challenge at the school level. Further, principals and teachers have not been trained in data analysis, so what might be useful lies unused.

Assumption 2. Students will be motivated to invest more time and energy because they face consequences, and because they are aware of how much they need to learn. Most teachers can teach the students easiest to teach (those who come into their classes with the store of basic knowledge about literacy and numeracy that readies them to learn). These students also come with the understanding that leads them to value what the teachers tell them is important, or at least they are compliant enough to suspend their disbelief.

Many of those far behind, however, have had previous academic experiences that have led them to believe school has little value for them. They may know abstractly that it has value but are not convinced that they will benefit from it because they never have. Many do not read well enough to learn or to enjoy reading, for instance. They know that they face consequences, and are disappointed in themselves, but have no sense of how they might turn the situation around.

Assumption 3. We have the right type of instruction for these students. Most teachers have not been prepared to teach students with differing levels of preparedness and knowledge, nor do they work in schools organized to make differentiated instruction reasonably possible. They do not know how to assess accurately where each student is, design a course of study for each depending on need, and then manage all these different levels. Most do not know how to teach in a sophisticated, highly intellectual way to build students' knowledge and skills, which is what a standards-based reform system requires.

There is a further problem with the instruction urban students receive. The crucial relationship in teaching is the one between the teacher and students, and their mutual engagement in the content. In most urban classrooms, however, teachers have an uneasy sense of the unknown and unknowable lives of their students and fear losing control. This leads teachers to minimize interactions with students and to make the exploration of content, ideas, and differences rare, even though these are essential to the higher learning demanded by standards. There is little "talk" or discourse in average urban classrooms. All of this results in a lacking sense of efficacy on either teachers' or students' parts, and little overall engagement.

Finally, the organization and use of resources and time in most schools is not conducive to change or improvement. Adults have little time to learn or to interact with each other, nor a means to reorganize themselves.

Assumption 4. Adults are held accountable for the learning of students. In virtually all states and urban districts, the unit of analysis for accountability is the district, the school, or the student. It is not the teacher. Principals are evaluated, and are sometimes held accountable and let go if there are available replacements. But because replacements are not often available, districts stick with mediocre principals. Furthermore, the current teacher evaluation system rarely includes student performance results, teacher knowledge of the material on the state's standards, or the practice of effective pedagogy. When teachers are evaluated, the evaluation seldom includes an analysis of the effectiveness of their instructional practices in a deep way that leads to improvements in their classrooms. Many teachers report being visited by their principal rarely.

Intensive teacher evaluations are usually centered on the worst teachers, not the average ones, and rarely do evaluations highlight and elevate the superb practice of the best, who are obtaining wonderful performance with their students. Many critics blame unions for protecting teachers, but teacher evaluation problems go well beyond teachers' union issues. Although few districts have the contractual relationships right yet, there could be steps taken within existing contracts that would begin including student performance as part of evaluations. This would lead to a more robust adult accountability system.

Assumption 5. There is a supply of well-prepared and interested individuals ready and willing to step into urban classrooms, were we ready to terminate the mediocre ones. This is demonstrably untrue, as a look at California and cities elsewhere makes clear. Reports by the Education Trust, the National Commission on Teaching for America's Future, and others have documented the supply problem well. Beyond the numbers, even when states are tightening up qualifications, they tend to be assessing only low-level skills of future teachers. Teacher preparation institutions receive accreditation routinely without making any substantive changes in how teachers are prepared and trained.

Assumption 6. Schools and systems are coherent, and we have, at scale, examples of how to organize time, money, people, and support to get instructional improvement. We do not have the examples of high-performing districts that we need. The knowledge base about large-scale improvement is shallow. The Annenberg Institute for the Redesign of Urban School Districts has created a task force to find good models to

inform and improve support for schools so that instruction improves, but their work is incomplete. To date, they have found some high spots, but overall, the research in this field is weak.

Looking to the Future: Is the Situation Hopeless? Not Necessarily

The situation can be changed if we collectively take several important steps.

Step 1: Recognize and accept that upping the stakes and consequences for schools and students, as the new federal legislation No Child Left Behind does, will not by itself cause instructional improvement, school coherence, or improved student performance.

Step 2: Conduct much more research on instructional improvement and then highlight the visible models of how it takes place. Many cities, like Boston, have parts of the answer, but San Diego, several New York community districts, Cincinnati, Long Beach, Houston, Denver, and many others have other pieces to the puzzle.

Step 3: Start making greater investments in the right things: improving teacher and principal knowledge about content, pedagogy, and the relationship between them. School staff cannot do what they do not know how to do at a high level: teach urban students to master challenging content no matter where they begin and how far behind they are.

Step 4: Get the data systems right so that they yield useful and fine-grained information for students, parents, and teachers. Technology has a greatly underdeveloped role in helping to solve this problem, but the knowledge and skill of teachers and principals to reflect on and use data about students' performance also must be addressed.

Step 5: States and the media should stop misusing assessments so we can build public understanding of the true problem and the solutions. Tests are important and useful, but not a good instrument to pinpoint the problem to be solved.