PART I

1

The Pending Reauthorization of NCLB

An Opportunity to Rethink the Basic Strategy

Daniel Koretz

The pending reauthorization of No Child Left Behind (NCLB) is generating intense debate about possible modifications of many of its provisions such as the requirements for disaggregated reporting, adequate yearly progress (AYP), the draconian requirements for the assessment of students with disabilities, and the provisions for testing students with limited proficiency in English.

But as important as it is, the debate about the specifics of NCLB obscures three more important problems that we cannot afford to ignore:

- First, we know far too little about how to hold schools accountable for improving student performance. NCLB and its state-level forebears—dating back to the first minimum-competency testing programs more than three decades ago—have been based on a shifting combination of common sense and professional judgment, not on hard evidence.

- Second, some important aspects of NCLB (and its antecedent state programs) are inconsistent with the evidence we do already have.
- Third, much of the apparent progress generated by NCLB and similar programs is spurious, a comforting illusion that we maintain for ourselves—at a great cost to children—by failing to perform appropriate evaluations.

In this chapter, I will elaborate on these three points and will briefly sketch a few of the most important things we know—and don't know—about test-based educational accountability and its effects. I will end with a plea that we use the coming reauthorization as an opportunity to belatedly ramp up the hard work of research, development, and evaluation needed to create effective accountability systems—not as a substitute for alterations to the requirements for AYP, disaggregated reporting, and the like, but as an essential complement to them.

As a former public school teacher, parent, and educational researcher for more than a quarter of a century, I remain convinced that the educational system needs more effective accountability systems and that achievement testing has to be one element of them. But research has shown that we are making a hash of it. It is our obligation to children—particularly to those faring poorly in the current system—to do better.

WHAT THE EVIDENCE DOES AND DOES NOT TELL US

Clues to more productive approaches to educational accountability—in particular, approaches that are most likely to benefit the students in historically low-performing groups—lie both in what research has found and in the questions it has not yet answered.

Does High-Stakes Testing Work?

A modest number of studies argue that high-stakes testing does or doesn't improve student performance in tested subjects. This research tells us little. Much of it is of very low quality, and even the careful studies are hobbled by data that are inadequate for the task. Moreover, this research asks too simple a question. Asking whether test-based accountability *works* is a bit like asking whether medicine works. What medicines? For what medical conditions? Similarly, test-based accountability takes many forms that are likely to have different effects. Its impact is likely to vary among types of schools and students.

Test-based accountability also has diverse effects that go beyond the test scores that serve as outcomes in these studies. A program that succeeds in raising mathematics scores may reduce achievement in other important subjects, for example, if teachers rob Peter to pay Paul, taking time away

from science or history to give more time to mathematics. And education has important goals that are not easily measured with standardized tests and remain unevaluated (Rothstein & Jacobsen, 2006).

Thus, the debate about whether high-stakes testing works is a red herring, distracting us from the question we ought to be asking: what *types* of accountability systems will most improve opportunities for students while minimizing the inevitable negative side effects? We need research and evaluation to address this question, because we still lack a well-grounded answer. We need to look at a range of outcomes beyond scores on the tests used for accountability. We need to create opportunities for designing these programs and for rigorously evaluating their positive and negative effects.

Can Score Increases Be Trusted?

Although research does not tell us whether high-stakes testing works, evidence does show us that high-stakes testing does not work nearly as well as it seems to. Just as economic work on incentives predicts, people try—often successfully—to game the system. As a consequence, scores on high-stakes tests can become dramatically inflated. That is, gains in scores can be much larger than the true improvements in student achievement that the scores are intended to signal. This creates an illusion of progress that is comforting to policymakers and educators but of no help whatever to children.

The issue of score inflation remains oddly controversial. Many in the policy world ignore it altogether or treat it as something that we really need not worry about. One superintendent of a large urban district dismissed the entire issue with a single sentence: "That's just a matter of opinion." He was wrong. Score inflation is a matter of evidence, not merely of opinion, and the problem is severe.

The inflation of test scores should not be surprising, since similar corruption of measures occurs in many other fields. Over the years, the press has documented corruption of measures of postal delivery times, airline on-time statistics, computer chip speeds, diesel engine emissions, television program viewership, and cardiac surgery outcomes, as well as scores on achievement tests (e.g., Cushman, 1998; Farhi, 1996; Hickman, Levin, Rupley, & Willmott, 1997; Lewis, 1998; Markoff, 2002; McAllister, 1998; Zuckerman, 2000). If many cardiac surgeons avoid doing procedures on high-risk patients who may benefit for fear of worsening their numbers, as the majority of respondents to a recent survey admitted (Narins, Dozier, Ling, & Zareba, 2005), it is hardly remarkable that some teachers and students will take shortcuts that inflate test scores.

The few relevant studies are of two types: detailed evaluations of scores in specific jurisdictions, and a few broad comparisons of trends on state tests and the National Assessment of Educational Progress (NAEP).

We have far fewer of the former than we should. The reason is not hard to fathom. Imagine yourself as superintendent of a district or state with rapidly increasing test scores. A researcher asks you for permission to evaluate the validity of these gains, to explore whether they are inflated and, if so, whether there are any useful patterns in the amount of inflation. This is not a politically appealing prospect.

The logic of both types of study is the same. The goal of education is to teach students skills and knowledge. A test score, which reflects performance on a very small sample of this material, is valuable only to the extent that it accurately represents students' overall mastery. In this respect, a test is much like a political poll. For example, two months before the 2004 election, a Zogby International poll of 1,018 likely voters showed George W. Bush with a 4 percent lead over John Kerry. This was a good prediction, as two months later Bush's margin was about 2.5 percent. But should we have cared how the specific 1,018 respondents actually voted? In general, the answer is "no." The voters sampled are just a drop in the bucket of millions of voters, and we worry about their opinions only because of what they suggest about the inclinations of the electorate as a whole. Analogously, we should not be too concerned about performance on the few specific items on a given test. Instead, we need to worry about the much larger domain of knowledge and skill that these few items are designed to represent.

For that reason, gains in scores on a high-stakes test, if they represent real gains in achievement, should *generalize*. Higher scores should predict better performance in the real world outside of the students' current schools—whether that is further education or later work. By the same token, score increases should generalize to better performance on other tests designed to measure similar bundles of knowledge and skills. Gains will not be exactly the same from one test to the next, but when tests are designed to support similar inferences about performance, gains ought to generalize reasonably well.

The results of the relatively few relevant studies are both striking and consistent: gains on high-stakes tests often do not generalize well to other measures, and the gap is frequently huge. When students do show improvements on other, lower stakes measures used to audit gains (most often NAEP), the gains on the audit test have generally been one-third to one-fifth the size of the gains shown on the high-stakes test. And in several cases, large gains on high-stakes tests have been accompanied by no improvement whatever on an audit test. For example, when Kentucky instituted a high-stakes testing program in the early 1990s—in several respects, a precursor of NCLB—fourth graders showed an increase of about three-fourths of a standard deviation on the state's high-stakes reading test in the space of only two years. This was a remarkably large increase for such a short time. By way of comparison, the pervasive decline in scores during the 1960s and 1970s that did much to provoke the last several decades of education reform averaged roughly three-tenths of a standard deviation

over a span of ten to fifteen years. During the two years that scores on the Kentucky reading test skyrocketed, the scores of Kentucky students on the NAEP reading test showed no increase at all (Hambleton et al., 1995). Other studies have found similar results in Chicago, in Houston, in Texas as a whole, and in an anonymous district I studied (Jacob, 2002; Klein, Hamilton, McCaffrey, & Stecher, 2000; Koretz, Linn, Dunbar, & Shepard, 1991; Schemo & Fessenden, 2003).

These few studies are complemented by a second group that provides a broad overview of the comparability of trends on state tests and NAEP. Studies in this second group show that in many—but not all—states, gains on state tests are larger, often markedly larger, than the same states' gains on NAEP (Fuller, Gesicki, Kang, & Wright, 2006; Fuller, Wright, Gesicki, & Kang, 2007; Ho & Haertel, 2006; Lee, 2006; Linn & Dunbar, 1990).

The implication of this research is inescapable: much of the apparent progress shown by increasing scores on high-stakes tests is simply spurious, an illusion that allows us to proclaim success while students continue to be deprived of opportunity.

Research indicates that score inflation varies markedly from school to school, but it does not yet provide any guidance about which types of schools are usually most affected. Given the current state of our knowledge, we cannot accurately predict which schools have sizable inflation and which do not, and we usually have no means of determining this from available data. This has two unfortunate consequences.

First, it vitiates conclusions about the *relative* effectiveness of schools. If inflation were uniform, overall gains would be exaggerated, but one could still identify the schools with relatively large or relatively small improvements in learning. But given our inability to pin down school-level variations in score inflation, conclusions about relative effectiveness are entirely untrustworthy if they are based only on scores on high-stakes tests, and we can expect to reward or sanction the wrong schools a good bit of the time.

Second, we cannot ascertain the relative impact of test-based accountability programs on the groups of students who are the focus of reform—in particular, groups that currently show low average achievement. Several other researchers and I have hypothesized that score inflation will often be worse in low-achieving schools. Our logic is simple: Systems such as NCLB require teachers in high-achieving schools to make relatively modest gains. (This depends on states' performance standards, of course, but it is also built into the AYP system and the *straight-line* systems many states used before NCLB.) Moreover, many high-achieving schools are in communities that offer relatively substantial out-of-school supports for student achievement, such as well-educated parents who press for high grades and can reteach material at home and buy afterschool tutoring. Teachers in low-achieving schools are required to generate far larger gains, and in many cases, they must do it with weaker community support. Faced with the need to do more with less, teachers in low-achieving schools will face

stronger incentives to cut corners in ways that inflate scores. But this remains only a hypothesis, not yet tested by much empirical evidence. Ho and Haertel (2006) found evidence that the disparity between trends in scores on state tests and those on NAEP tended to be larger for students eligible for free and reduced-price lunch, but with few data points, the difference was not statistically significant. We urgently need finer grained studies of this issue.

Some researchers have argued that unrealistically high performance standards are analogous to auto emissions controls: if you require more improvement than manufacturers can provide, you end up with some fraction of what you demand and, thus, are better off than you were before. Whether or not this is true of emissions controls, it is not likely to be true of test-based accountability. Under NCLB, one gets no credit for getting partway to AYP, and the tools for inflating scores are ready at hand. Therefore, one might get *less* real improvement by requiring too much gain, because teachers will have incentives to abandon legitimate instructional improvements that generate slower gains in favor of shortcuts—inappropriate test preparation or simple cheating—that generate faster gains. After more than three decades of high-stakes testing in the United States, we ought to have some hard evidence on this point, but we do not.

When I recently gave a talk on test preparation to a large group of principals, many from inner-city schools, I encountered educators' responses to excessive expectations for score gains. I explained the principle that tests represent very small samples from larger domains of knowledge and skills. Therefore, the good way to prepare students for high-stakes tests is to focus on the knowledge and skills the tests are supposed to represent so that students will have better capabilities when they leave school. The bad way to prepare them is to focus narrowly on the specifics of their own test—that is, to focus on raising scores on that specific test *as an end in itself*—which can lead to spurious gains limited to that one measure. By analogy, they should try to persuade the entire electorate in order to win the election, rather than trying to persuade Zogby's 1,018 respondents to change their votes.

I then gave the principals a dozen real examples of test-preparation activities, ranging from egregiously bad to reasonable by this criterion. I asked them to decide whether each one would teach the underlying knowledge and skills and therefore produce real gains that would generalize to more than one test.

A minority of the principals identified the particularly bad examples, and a few added examples of their own. One said that they are told what parts of the state's standards will be emphasized on the test so that teachers need not spend much time on the others—a sure recipe for score inflation. (There is now a term for this that makes it seem innocuous: *power standards.*)

But many of the principals steadfastly defended every single example of test preparation, even those that were unarguably bad. The most extreme was a case in which a district provided the actual test item in advance, changing only three trivial details, which is no more than simple cheating. That too was fine with many of the administrators. Many of them became hostile as the discussion continued.

In retrospect, these responses may not be surprising, given the incentives and sanctions these principals face under NCLB. For several years, they have been struggling to make AYP, which requires many of them to make far more rapid gains than any of us can tell them how to do by legitimate means. And the consequences of failure are dire. Then I explained to them that many of the methods they have been using in their desperate fight to keep their noses above water are simply inflating test scores. Upton Sinclair's principle applies: "It's difficult to get a man to understand something when his salary depends on his not understanding it." Until we impose a system that creates the right incentives, it is not reasonable to expect educators to ignore the perverse incentives we have already put into place.

How Do Educators Respond to High-Stakes Testing?

A substantial number of studies over the past few decades have investigated teachers' responses to high-stakes testing. These studies show a mix of desirable and undesirable responses (Stecher, 2002), and they help explain the inflation of scores found in the previously noted studies.

On the positive side, research suggests that in some cases, high-stakes testing has motivated teachers to work harder and more effectively. It leads many teachers to align their instruction more closely with the tested content, which—as we will see—can be both good and bad. Some teachers report that the results of high-stakes tests are useful for diagnosis. (However, it is the test—not the high-stakes attached to it—that is useful in this respect; tests designed for diagnostic purposes were widely used in American schools for decades before high-stakes testing became common.) Some studies have found specific instructional effects consistent with the goals of the accountability systems of which they are a part such as an increase in writing instruction when tests require substantial writing.

At the same time, research has shown a variety of negative effects of high-stakes testing on educational practice. Many of these can inflate test scores, and some are undesirable for other reasons as well. It is helpful to distinguish among different types of test preparation in terms of their potential to generate meaningful gains in achievement, score inflation, or both (Koretz & Hamilton, 2006; Koretz, McCaffrey, & Hamilton, 2001). I use "test preparation" to refer to all techniques used to prepare students for tests, whether good or bad, and deliberately avoid terms such as *teaching the test* and *teaching to the test*, which come freighted with inconsistent and

often poorly reasoned connotations. The different types of test preparation are as follows:

- Teaching more
- Working harder
- Working more effectively
- Reallocation
- Alignment
- Coaching
- Cheating

The first three are what most proponents of high-stakes testing programs—including NCLB—want and expect. "Working more effectively" (for example, finding better methods for presenting material to students) can only be for the good. "Teaching more" (allocating more time to instruction) and "working harder" (for example, implementing more demanding lessons) can both be carried to excess, to a point at which the marginal effects on learning are negative or they have other negative effects (such as an aversion to schooling or to learning) that offset short-term gains in achievement. But within reason, all three of these forms of test preparation can be expected to lead to meaningful gains in scores that signal higher achievement.

Cheating is the other extreme: it can only produce spurious gains in scores. There are limited systematic data about cheating, but there are enough accounts in newspapers and elsewhere to make it clear that it is hardly rare (see, for example, www.caveon.com/resources_news.htm). It takes all manner of forms: providing inappropriate hints during test administration, changing answer sheets after tests are completed, circulating actual test items (or items that are nearly identical) before a test, and so on. It is not clear that all instances of cheating are intentional, but they inflate scores regardless. The incentive to cheat is strongest in the schools that must make the largest gains—that is, low-scoring schools—but we have no systematic data showing whether it is in fact more common in such schools.

The controversial types of test preparation are the remaining three: reallocation, alignment, and coaching. All three can produce real gains, inflation, or both. The general principle is clear: these forms of test preparation are desirable when they improve students' mastery of the broad domains of achievement—say, eighth-grade mathematics—that the tests are designed to represent. They are undesirable and inflate test scores when they focus unduly on the particulars of the specific test chosen and therefore produce greater gains on that particular test than true improvements in learning warrant. In practice, however, the dividing line between the good and bad forms of reallocation, alignment, and coaching is sufficiently indistinct; keeping educators on the right side will be very hard until we do a better job of creating incentives for them.

Reallocation refers simply to shifting resources—instructional time, students' study time, and so on—to better fit the particulars of a testing program. Research has found that educators report reallocating their instruction in response to high-stakes tests. Reallocation occurs across subject areas, as shown by the reports of districts and schools reducing or eliminating time allocated to untested subject areas to make more time for the subjects that count in the accountability system (e.g., Rothstein & Jacobsen, 2006; Sunderman, Tracey, Kim, & Orfield, 2004). Educators can also reallocate instructional time within subject areas by emphasizing the particular portions that are emphasized by the test. Reallocation *within* subject areas is a key piece of the score-inflation puzzle.

Some amount of reallocation within subjects is desirable and is one of the intended effects of test-based accountability. If a testing program shows that students in a given school are not learning topic A, and topic A is important, the school's teachers should put more effort into teaching topic A.

The problem is that instruction is very nearly a zero-sum game: more resources for topic A necessarily mean fewer for topic B. If topic B is also important for the inference about performance, then taking resources away from it can inflate test scores.

Remember that a test is a small sample of a large domain of achievement, just as a poll is small sample of voters. The key to the success of both is that the small sample has to *represent* the larger domain. If teachers take resources away from relatively unimportant material to make way for emphasizing topic A, then all is fine. But if the material that gets *less* emphasis is an important part of the domain—if it is an important part of what users of the scores think they are measuring—then performance on the tested sample will show improvements when mastery of these other important parts of the domain is stagnant or even declining. This is precisely what studies of score inflation have found.

The more predictable a test is, the easier it becomes for teachers to reallocate in a way that inflates scores. For any number of reasons—the pressure of time, costs, a desire to keep test forms similar to facilitate linking of scores from year to year, the creativity needed to avoid similarities—most tests show a considerable resemblance from year to year. In many programs, much of the specific content is replaced each year, but the types of content and the style and format of test items show noticeable similarities from year to year. Some educators try hard to discern these recurrences, but they need not do it on their own; there is a vibrant industry of test-prep firms that will do it for them, and many districts and states provide help with this as well.

Alignment is a cornerstone of current education policy and is noted repeatedly in NCLB. Instruction is to be aligned with content and performance standards, and assessments must be aligned with both. Up to a point, alignment is clearly a good thing: we want teachers to focus on important

material, and no one would want to judge teachers or schools by testing students on content that schools are not expected to teach.

Alignment is often cast as an unmitigated good and is frequently presented as a means of preventing score inflation. Not long ago, for example, a principal well known for achieving high scores in a poor, mostly minority school angrily told a crowd of college students that concerns about teaching to the test are completely unwarranted in her state. We don't have to worry about teaching to the test, she maintained, because her state's test covers important knowledge and skills that the students need.

This is nonsense, even assuming that her state's test does focus on important knowledge and skills. She was mistaking the test for the domain it represents—confusing Zogby's 1,018 respondents with the electorate. Alignment is nothing more than reallocation by another name, albeit with the constraint that the material emphasized must be consistent with standards. But whether alignment or other reallocation inflates scores depends on more than the quality of the material that is given additional emphasis. It also critically depends on the material that is given *less* emphasis. Because tests are such small samples from large domains, it is entirely practical to give more emphasis to some important material while taking it away from other equally important material. There is ample room to take it away from other material aligned with standards (hence test preparation focusing on "power standards") Research confirms this. Studies of Kentucky's assessment program of the 1990s, which was an archetypal standards-based system, found severe score inflation in every comparison examined (Koretz & Barron, 1998).

The final form of test preparation is *coaching*, a term that I use to refer to focusing instruction on fine details of the test such as the format of test items, the particular scoring rubrics, or minor details of content. Encouraging students to use format-dependent, test-taking strategies such as plugging in and process of elimination, is a form of coaching, and these can generate gains that evaporate when students are presented with different tasks—for example, constructed-response tasks that have no choices to eliminate.

Inflation of scores does *not* require that teachers or students focus on unimportant material. It can arise that way, for example, if teachers focus on test-taking tricks rather than important content. But this is not necessary. Inflation can occur from excessive narrowing of instruction, even if the material taught is valuable. One secondary-school mathematics teacher told me that her state's test presented only regular polygons, and therefore, she asked, why would she bother teaching about irregular polygons? What she meant was, "Because my goal is to raise scores, why would I. . .?" If her question had been, "Because my goal is to teach my students plane geometry, why would I. . .?" the answer would have been different and equally obvious.

The lesson is that the incentives we currently give teachers are too crude and simply don't work as advertised. The goal has become raising scores as an end in itself—persuading Zogby's 1,018 respondents—rather than improving learning. The incentives teachers face do not favor the good forms of reallocation, alignment, and coaching over the bad. Many educators take the path of least resistance; by doing so, they inflate scores. The system cheats kids of the education they deserve.

A common but mistaken response is that inappropriate reallocation and coaching arise because we use "bad" tests. The argument is that if we just built better tests, these problems would be solved. This was an argument made for moving from multiple-choice to performance assessments nearly twenty years ago, and for moving from those to today's standards-referenced tests. Neither change solved the problems of inappropriate test preparation or score inflation, and better tests will not solve them now. With enough creativity, time, resources, and evaluation, we could improve tests to *lessen* these problems, for example, by deliberately avoiding unneeded recurrences over time and by building in novel content and forms of presentation for purposes of auditing score gains. But there are numerous factors that limit how much we ameliorate the problem—for example, the need to keep tests sufficiently similar from year to year to allow meaningful linking of scores; resource limitations; the limited and already strained capacity of the testing industry; and the requirement that, when students are given scores, those within a cohort are administered the same or comparable sets of items. Moreover, many important outcomes of education are difficult or impossible to measure with standardized testing.

Finally, there is the problem of incentives for chief state school officers. Under the provisions of NCLB, what would motivate one to spend considerably more money to buy a somewhat inflation-resistant test that would generate smaller observed gains in scores? Better tests—by which I mean tests designed with an eye to the problems caused by test-based accountability—might indeed be an important step, but it will not suffice, and it is no substitute for putting in place a more reasonable set of incentives for teachers.

How Much Gain Is Feasible?

One of the most remarkable and dysfunctional aspects of the test-based accountability systems now in place under NCLB is that performance targets are usually made up from whole cloth, with no basis in experience, historical evidence, or evaluations of previous programs. And for political rather than empirical reasons, the targets are uniform for all schools in a state, regardless of their initial levels of performance and the particular impediments they face to improving scores.

Proponents of standards-based reporting of test scores will bristle at the word *arbitrary*, but that is a reasonable label for performance targets