The Pending Reauthorization of NCLB: An Opportunity to Rethink the Basic Strategy

October 10, 2006

Daniel Koretz
Harvard Graduate School of Education

Invited Paper for Civil Rights Project/ Earl Warren Institute
Roundtable Discussion on the Reauthorization of NCLB
Washington, D.C.
November 16, 2006

The pending reauthorization of NCLB is generating intense debate about possible modifications of many of its provisions, such as the requirements for disaggregated reporting, the structure of AYP, the safe-harbor provision, the draconian requirements for the assessment of students with disabilities, and the provisions for testing students with limited proficiency in English. Much of this meeting, I expect, will be devoted to these issues.

I would ask that you look beyond them. I don't mean to belittle their significance. To pick just one example: most of us believe that the requirement for disaggregated reporting has been highly valuable and should be retained in some form, but the current requirement poses serious technical problems, and devising a better alternative would be an important improvement. The current AYP provisions are also seriously problematic.

But as important as it is, the debate about the specifics of NCLB obscures three more important problems that the civil rights community 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 – in fact, the entire run of test-based accountability systems 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. They were not based on hard evidence. Einstein is reputed to have said that "common sense is the collection of prejudices acquired by age eighteen." It is in any case an inadequate basis for designing educational accountability programs.
- 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 students – by failing to perform appropriate evaluations.

In the main part of this paper, I will briefly sketch a few of the most important things we do – and don't – know about 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 and disaggregated reporting and the like, but as an essential complement to them.

Before starting, I think it is important to put my own views on the table. Often, I and others who criticize NCLB and other test-based accountability systems are branded as 'anti-accountability' or 'anti-testing.' Some of the critics are, but I am not. My brief experience as a public school teacher convinced me of the need for more effective methods for holding educators accountable. My experiences as a parent of children in public schools made this need only more apparent. Nothing in my career in educational research, which extends back more than a quarter of a century, has led me to question this opinion. And I remain convinced that achievement testing has to be one element of a successful accountability system. 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 than we have.

WHAT THE EVIDENCE DOES AND DOES NOT TELL US

Clues to potentially more productive approaches to educational accountability – in particular, approaches that are most likely to benefit the students whose well-being is the focus of the civil rights community – lie both in what research has found and in the questions it has not yet answered.

Does High-Stakes Testing Work?

The research community has produced a modest number of studies in recent years arguing that high-stakes testing does or doesn't improve student performance in tested subjects. In my opinion, this research tells us little. Much of it is of very low quality, and even the careful studies are hobbled by data that in several ways are not up to the task – for example, that do not allow analysts to disentangle the effects of accountability from other influences on student performance.

But the bigger lesson to be drawn from this research, I believe, it that it asks the wrong question, one that is in two respects too simple.

First, asking whether test-based accountability 'works' is a bit like asking whether medicine works. The answer obviously depends on which medicines and for which medical conditions. Test-based accountability takes many forms that are likely to have

different effects, and the impact of many of them is likely to vary among types of schools and students.

Second, test-based accountability has diverse effects that go beyond the tested outcomes measured in many of these studies. For example, a program that succeeds in raising mathematics scores may well reduce achievement in science, if teachers struggling to improve performance in mathematics rob Peter to pay Paul, taking time away from other important subjects. And education has important goals that go beyond those that are easily measured with standardized tests, so adding tests in more subjects is only a partial solution to this problem.[i]

Thus, the debate about whether high-stakes testing 'works' is a red herring, distracting us from the questions we ought to be asking. We need to be asking what *types* of accountability systems will most improve opportunities for the students about whose welfare the civil rights community is particularly concerned, while minimizing the inevitable negative side-effects. To be more precise, we need to be conducting the research and evaluation needed to address this question, because we still lack a well-founded answer. We need to look at a wide range of outcomes in addressing this question. And we need to create the 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 that it works far less well than it seems. 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, creating 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 treating it as a minor fly in the ointment, something that we really need not worry about. In a meeting not long ago, the 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 opinion, and the problem is severe.

The inflation of test scores should not be surprising, since similar corruption of measures occurs in all manner of other domains and should be familiar to any careful reader of a good newspaper. For example, over the years, the press has documented corruption of measures of postal delivery times, airline on-time statistics, computer chip speeds, diesel engine emissions, TV program viewership, and cardiac surgery outcomes, as well as scores on achievement tests.[ii] If many cardiac surgeons avoid doing procedures on patients who may benefit for fear of worsening their numbers, as the majority of respondents to a recent survey admitted,[iii] it is hardly remarkable that some teachers and students will take shortcuts that inflate test scores.

But the issue of score inflation remains contentious nonetheless, so I want to be very clear about the nature and extent of this research. The relevant studies are of two types: detailed evaluations of scores in specific jurisdictions, and a few broad comparisons of trends on state tests and NAEP. The former are few in number, far fewer than we would ideally have. 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. Not a politically appealing prospect. The first study of this sort, which I conducted with Bob Linn, Lorrie Shepard, and Steve Dunbar almost 20 years ago, is illustrative. We first proposed it in response to a state's request for evaluations of its new, and at the time innovative, high-stakes testing program. The state responded to our proposal by withdrawing and revising its request for proposals to make sure that nothing like our study would be done on its tab. We then obtained independent funding and permission to conduct the study in a large district, but after we had expended considerable effort and resources, we were thrown out. The Superintendent explained to me that the results of the study could get him in trouble with his legislature. On the third try, we succeeded, but only by promising extreme steps to protect the anonymity of the district.

The logic of both types of study is the same. Our goal is to teach kids 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. A test is in this respect 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 percentage point lead over John Kerry. Not too bad a prediction: Bush's margin two months later was about 2.5 percent. But should we have cared how the specific 1,018 respondents themselves actually voted? In general, no; the specific 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 worried 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.

And for that reason, gains in scores on a high-stakes test, if they represent real gains in achievement, should *generalize*. They should predict better performance in the real world outside of the students' current schools – whether that be later studies or the world of work. And by the same token, score increases should generalize to better performance on other tests designed to measure similar bundles of knowledge and skills. Of course, tests differ in their content, and one would not expect gains to 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 from one to the next.

The results of the relatively few detailed studies are both striking and remarkably consistent: gains on high-stakes tests typically do not generalize well to other measures, and the gap is usually huge. When students do show improvements on other the lower-stakes measures used to audit gains (most often the National Assessment of Educational Progress), 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, during the first two years of the high-stakes testing program Kentucky instituted in the early 1990s – in several respects, a precursor of NCLB – fourth-graders showed a staggering increase of about 3/4 of a standard deviation on the state's high-stakes reading test. NAEP, however, showed no increase at all.[iv] Other studies have found similar results in Chicago, Houston, Texas as a whole, and the unnamed district I mentioned earlier.[v]

The second group of studies, which provide a broad overview of the comparability of trends on state tests and NAEP, complement the small number of detailed studies. They are consistent, showing that in many but not all states, gains on state tests are substantially, sometimes dramatically, larger than the same states' gains on NAEP.[vi]

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

The extant research indicates that this inflation of scores is highly variable from school to school, but it does not provide any general guidance about which schools have the most severe inflation. That is, with the current state of our knowledge, we cannot accurately predict – or determine from reported scores – which schools have sizable inflation and which do not. This has two unfortunate consequences.

First, it vitiates conclusions about the *relative* effectiveness of schools. If inflation were fairly uniform, one could argue that even though overall gains are exaggerated, one could at least identify and reward the schools with relatively large improvements in learning and find and sanction those with relatively small gains. But given our lack of knowledge about school-level variations in score inflation, such conclusions 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 the civil rights community's concerns. For two decades, several of us who have toiled in this vineyard have hypothesized that under many high-stakes regimes, 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 in 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 re-teach material at home and buy after-school tutoring. Teachers in a low-achieving school

must generate far larger gains, and in many cases must do it with weaker community support. Our hypothesis is that 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 it remains a hypothesis: we have reports of variations in test preparation strategies, but we have no strong evidence of differences among types of schools in the severity of score inflation.

Chris Edley and I have argued about this issue for several years. Chris has suggested a model analogous to auto emissions controls, arguing that 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. As the diesel engine emissions example noted above suggests, this optimism may not always be warranted in the case of emission controls, but for the sake of argument, say that it generally is. I have countered that educational testing under NCLB and some other accountability programs is different: one gets no credit getting part way 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 short-cuts – inappropriate test preparation, or simple cheating – that generate faster gains. After more than three decades of high-stakes testing in the U.S., we ought to have some hard evidence on this point, but we do not.

I recently gave a talk on test preparation to a group of principals attending an institute to help them learn how to use data more effectively to improve instruction. They came from various backgrounds, but a disproportionately large number were drawn from inner-city schools with overwhelmingly poor and minority enrollments. I began by explaining the basic principle of testing I noted above: that tests represent very small samples from the larger domains of knowledge and skills with which we should be concerned. Therefore, the good ways to prepare students for the high-stakes tests these principals confront 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 so narrowly on the specifics of their own test so that their students' gains will be largely limited to that one measure. That is, if they focus too narrowly on the particulars of their own test, worrying about raising scores on that

specific test *as an end in itself*, they will inflate their students' scores on the test and leave them without stronger skills that would be of use to them in later education or employment. 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 voters to change their votes.

I then gave the principals a dozen real examples of test preparation activities, ranging from egregiously bad to quite reasonable by this criterion. I asked them to decide whether each one was bad, good, or both, and to explain why, focusing their explanation on whether students would learn the underlying skills and therefore show gains that would generalize to more than one test.

Their responses were distressing. Some principles got the idea, explained why some of the examples were undesirable, and added that they were appalled by some of what is done in their schools. For example, one said that they had someone from their test publisher come in to explain what parts of the state's standards would 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: the standards emphasized by the test are called "power standards.") But many of the principals steadfastly defended every single example of test preparation, even those that were unarguably bad. For example, one of the methods I gave them is plugging in answers. For example, suppose that a given multiple-choice test item requires students to solve an algebraic equation. One can determine the correct answer without knowing how to solve the equation by plugging each of the answer choices in turn to see whether the equality holds. For the most part, teaching students to plug in only helps students answer multiple-choice questions because multiple-choice questions are the only ones that provide answers to plug in. I pointed out that if I simply changed the test to present the same content in a constructed-response format – in which students must generate their own answers and are given no choices to plug in – the gains obtained with this test preparation would vanish. And I pointed out that in the real world, employers are rarely so kind as to present their employees problems in the multiple-choice format. No matter, they replied. One principal defended spending time on plugging in as a good way of teaching "information retrieval skills." I gave them an another example in which a district provided the actual test item in advance, changing only three trivial details, a form

of test preparation that I would classify as simple cheating. That too was fine with them. The tone of some of the principals was hostile.

I can only guess as to the reasons for these responses, but the best guess is the incentives 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, the safely tenured expert, waltz in and explain 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, predominately surveys and case studies, have investigated teachers' responses to high-stakes testing. These studies, which are reasonably consistent, show a mix of desirable and undesirable responses, and they help explain the inflation of scores found in the studies noted above.[vii]

On the positive side, research suggests that high-stakes testing has in some cases motivated teachers to work harder and more effectively. It does lead 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. My colleagues and I have suggested that it is

helpful to distinguish different types of test preparation in terms of their potential to generate either meaningful gains in achievement, score inflation, or both.[viii] We use "test preparation" to refer to all techniques used to prepare students for tests, whether good or bad, and we deliberately avoid terms like "teaching the test" and "teaching to the test," which come freighted with inconsistent and often poorly reasoned connotations. We distinguish seven forms of test preparation:

- Teaching more;
- Working harder;
- Working more effectively;
- Reallocation;
- Coaching;
- Alignment; and
- Cheating.

The first three are what most proponents of high-stakes testing – including NCLB—want and expect. "Teaching more" and "working harder" can both be carried to excess, to a point at which the marginal effects on learning are negative or at which 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, that is, higher achievement as well as higher scores.

Cheating is the other extreme: it can only produce bogus gains in scores. There is limited systematic data about cheating, but there are enough news accounts to make it clear that it is hardly rare.[ix] 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; many educators are completely untrained in testing and do not have a clear notion of the limits of appropriate practice. But for present purposes intent does not matter; cheating inflates scores regardless. My speculation is that cheating is more common in low-scoring schools, again because of the far greater pressure to raise scores, but there are no systematic data to test this hypothesis.

The interesting and controversial types of test preparation are the remaining three: reallocation, alignment, and coaching. All three can produce either real gains, inflation, or both. The general principal 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 test then 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 that keeping educators on the right side will be very hard until we do a better job of creating incentives for them. The disturbing responses to test preparation activities noted earlier are an example of this.

Reallocation refers simply to shifting resources – instructional time, students' study time, parental nagging, and so on – to better fit the demands of a testing program. Many studies have found that educators report reallocating their instruction in response to high-stakes tests. Reallocation can be carried out within subject areas, by emphasizing the particular portions that are emphasized by the test. It can also occur 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 are tested and hence count in the accountability system.[x] For present purposes, I will leave aside reallocation between subject areas. It is important in its own right – and should concern this group because of evidence that it more severe in low-achieving schools – but it has only an indirect connection to the issue of score inflation. Reallocation *within* subject areas, however, 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, one would want the school's teachers to 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 means fewer for Topic B. And if Topic B is important, in a very specific sense, 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 measured parts of the domain (the small sample in the test) 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. For example, an evaluation of Kentucky's KIRIS test-based accountability system (in many respects, the precursor of NCLB) in the early 1990s found that huge gains by state's fourth graders on the state's reading test were accompanied by no gains whatever on NAEP, even though the state's reading test was designed to have a framework similar to NAEP's.[xi] Teachers had found ways to focus so narrowly on the particulars of KIRIS that none of the ostensible improvement appeared on the NAEP – which implies that they would not have shown up in the real world outside of school either.

The more predictable a test is, the easier it becomes to reallocate in a way that inflates scores. For any number of reasons – for example, the pressure of time and costs, a desire to keep test forms similar enough to allow sensible linking of scores from year to year, the creativity needed to avoid similarities – most testing programs 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 (what types of equations should be included in a ninth-grade algebra test?) 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. You can readily find material that would help you reallocate – and inflate scores – online.

Alignment is a cornerstone of current education policy, noted over and over 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.

But alignment is often cast as an unmitigated good, and not infrequently, one will hear alignment presented as a means of preventing score inflation. Not long ago, a principal well known in the Boston area for achieving high scores in a poor, mostly minority school angrily told a crowd of our students that critics who warn of teaching to the test are completely off base. We don't have to worry about teaching to the test, she maintained, because the test her students take (the Massachusetts MCAS) tests important knowledge and skills that the students need to have.

This is utter nonsense. By this I do not intend to disparage the MCAS; her argument would have been specious regardless of which state's test her students took. 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 one constraint: the material gaining in emphasis must be consistent with standards. But whether reallocation – alignment or other forms – inflates scores depends on more than the quality of the material given additional emphasis. It also depends – critically – on the material 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. In fact, there is ample room to take it away from other material aligned with standards (hence test preparation focusing on "power standards.") And research confirms this. The most investigation of inflation to date examined Kentucky's assessment program of the 1990s, KIRIS, which was one of the archetypes of a standards-based system. Studies of this program found severe score inflation in every comparison examined.[xii]

Another way to put this is that 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.

The final form of test preparation is *coaching*, a term that my colleagues and I use to refer to focusing instruction on fine details of the test, such as the format of test items,

the particular rubrics used to score work, 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 it generates gains that evaporate when students are presented with tasks that have no choices to plug in or eliminate. One secondary-school mathematics teacher in a study of mine gave me a wonderful example of coaching based on details of content: her state's test presented only regular polygons, she said, so why would she bother including irregular polygons in her teaching? What she meant was: 'since my goal is to raise scores, why would I…?' If her implicit question had been 'since my goal is to teach plane geometry, why would I…," the answer would have been different but equally obvious.

The lesson from all of this 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 the 1,018 Zogby respondents – rather than improving learning. The incentives teachers face do not favor the good forms of reallocation, alignment, and coaching over the bad. So many educators take the path of least resistance, and by doing so, they inflate scores. The system cheats kids of what they deserve.

A common but mistaken response is that inappropriate reallocation and coaching arise because we use 'bad' tests. If we just built better tests, the argument goes, these problems would be solved. This was an argument made for moving from multiple-choice to performance assessments nearly 20 years ago, and for moving from those to today's standards-referenced tests. Neither change solved the problems of inappropriate test preparation and score inflation, and we are not going to solve them now with better tests. With enough creativity, time, resources, and evaluation, tests could be improved to *lessen* these problems – for example, by deliberately avoiding unneeded recurrences over time, and by building in novel content and novel forms of presentation for purposes of auditing score gains, as discussed below. But as noted earlier, there are numerous factors that limit how much we ameliorate the problem – e.g., 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, when students are given scores, that students within a cohort are administered the same or comparable sets of items. Moreover, there are many important outcomes of education

that are difficult or impossible to measure with standardized testing – a notion that is often branded as "anti-testing" but was in fact carefully explained over half a century ago by one of the most influential proponents of standardized testing.[xiii] And finally, there is the problem of incentives. Under the provisions of NCLB, what would motivate a chief state school officer 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.

How much gain is feasible?

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

Proponents of standards-based reporting of test scores will bristle at the word "arbitrary," but the fact is that in almost no cases are standards or performance targets set on the basis of empirical evidence of attainable improvements. We do have evidence that would be useful for setting targets, but policymakers have generally ignored it. We might start with the data we have on long-term trends in achievement. For example, the achievement decline of the 1960s and 1970s created great consternation and was a major impetus for the waves of education reform that continue with NCLB. Should we assume that schools can quickly implement reforms that produce gains as rapid as the declines of that era? Or, as Bob Linn has suggested, we might examine the rate of gain (on uninflated measures) shown by the most rapidly improving schools, perhaps the top 10 or 20 percent, and use their gains as an upper bound for our expectations. We could also use international comparisons to help us decide what is reasonable. For example, even ideal policies would presumably only bring us to the level of the highest-performing countries over a long period, if at all. (Given international differences in factors that are outside of the control of educators and educational policymakers, it is not clear that

reaching those levels even over the long term is entirely realistic.) Finally, we could use research, development, and evaluation to help set targets, as we do in policy domains as varied as public health and auto safety. That is, we could design new reforms, implement them on a limited but planned basis, and subject them to rigorous tests to determine their effects before putting them into operation nationwide or even statewide. We generally have done none of these things in setting performance targets.

The result has been performance targets that are both arbitrary and in some cases counterproductive because they are unrealistic. Bob Linn did an equipercentile linking of scores on NAEP and TIMSS to show that if states set eighth-grade mathematics standards comparable in difficulty to the NAEP Proficient standard, as innumerable critics of state standards suggest they should, we would be setting targets that roughly a third of the students in Japan and Korea – two of the highest-scoring countries in the world – would fail to meet.[xiv] Is it realistic to expect that virtually all of our students will exceed such a threshold in a mere 12 years? Keep in mind that the current NCLB regulations governing the assessment of students with disabilities mandate that all but 1 percent of students must meet the grade-level standard, which language in one of the Notices of Proposed Rulemaking explained was to make sure that mildly retarded students are held to grade level standards. So if we use standards as rigorous as those of the National Assessment, we are expecting our mildly retarded students to reach a level of proficiency mathematics that a third of the students in the highest-scoring countries in the world cannot reach. And in merely twelve years. That is not only foolish; it is counterproductive – increasing the incentives to cut corners and inflate scores – and cruel to some lower-performing students.

Some years ago, I converted performance targets in Kentucky's KIRIS system, which was in many respects a prototype for NCLB, into the metric of standard deviations per year in order to compare them to long-term trends. I found that the system required far larger gains than we had seen in historical data.[xv] This was not a deliberate policy choice; no one even knew the numbers until I calculated them a few years after the program started. And to my knowledge, no one has yet carried out this exercise for the improvements states require under NCLB. We simply shoot from the hip and remain sanguine, for no real reason, that educators and students can reach these targets by

legitimate means and that we are doing more good than harm. And worse: in most instances, we have in place no credible mechanism for measuring the good and the harm.

How much can the variability of achievement be shrunk?

One of the most positive aspects of NCLB is its belated focus on equity. Many of the key aspects of NCLB – disaggregated reporting; the conjunctive AYP system (which classifies a school as failing if any one of the mandated reporting groups fails to make AYP); the uniformity of the ultimate performance targets; and the method of calculating AYP, which requires greater gains by lower-scoring schools – are motivated by a laudable desire to decrease inequities in educational outcomes. These provisions follow in the footsteps of widespread state initiatives that had the same goals, such as the 'straight-line' accountability systems that required all schools to progress continually from their initial performance to the uniform statewide goal. And they reflect one of the principal policy mantras of the past 15 years: "all students can learn to high levels."

A decade or more ago, I began asking people how they expected all students to reach high standards. Would the distribution of achievement remain as wide as it currently is but skyrocket upwards, so that even kids in the long left-hand tail of relatively low scores would reach the targets? Would the distribution get narrower – perhaps with both the low and high extremes moving in toward the mean, or with the high scorers remaining diverse but the low-scoring tail being pulled up toward the average? No one I asked had a ready answer, but with some thought, they usually said that the both would happen: the entire distribution would move up, but low-scoring kids would move up faster, so the total variation among kids would decrease. (No one wanted to compress the distribution of high-scoring kids, making them more like the average.)

This raises an obvious question that to my knowledge has received almost no attention in the debate about NCLB or other accountability systems: just how much can we shrink the variability of achievement?

To answer this question, it is essential to separate two distinct issues that are often confounded. One issue is variation *among groups* – for example, differences in performance between poor and rich, or between minority and majority children. The second is the total variability of performance of *individual children*. These two are not the same: reducing differences among groups does not have a commensurate effect on the

total variability among individual children. Here I am concerned only with the latter: the total spread of student achievement.

Most people seem to assume that we can shrink this variability a great deal. On its face, this would appear to be a sensible expectation. After all, we have enormous and well-documented inequities in school quality and in the opportunities afforded to students. And despite intermittent progress for several decades, we still have very large gaps in performance between racial/ethnic groups and between the poor and the well-off. It seems only reasonable to expect that if we garnered the political will to combat these inequities – which would take a great deal more than an educational accountability system such as NCLB – the variation in student achievement would shrink dramatically.

But the evidence suggests that it would not. It would shrink, of course, but the variation remaining afterwards would still be huge. One indication of this is international comparisons. Since the first release of data from the Third International Mathematics and Science Study (TIMSS, now the "Trends in International Mathematics and Science Study"), we have had information on international differences in the variability of student performance, and these do not comport with common expectations. There are some exceptions, but most countries show fairly similar amounts of variation. Moreover, there is no consistent relationship between the amount of variability and either social heterogeneity or educational inequities. Japan and Korea, for example, are roughly similar to the U.S. in the variability of students' scores.[xvi]

There may be a variety of reasons for this pattern, but one is a statistical fact that many people find counterintuitive: the mean differences in scores between groups in the U.S., while very large, contribute only modestly to the total variability of performance among individual students. Most of the variation in the entire population arises from the huge variation in scores *within* groups, not from the differences *between* groups. For example, if one entirely eradicated the mean differences between racial/ethnic groups in the U.S., so that scores in each group were distributed just as they now are among non-Hispanic whites, the total variation in student performance would shrink modestly. I calculated this with two nationally representative tests (NAEP and NELS) for both reading and mathematics in grade 8. The reductions in the standard deviations – a

conventional measure of the spread of scores – ranged from about half of one percent to nine percent.

In the ideal world, we would do even more than that; for example, we would diminish the inequities between poor and well-off white students, poor and well-off Latino students, and so on. And by doing so, we would shrink the total variability of student performance somewhat more than these numbers suggest.

But the implication is clear: even if we finally create a more equitable educational system and more equitable community supports for learning, we are going to be stuck with enormous variations in student performance, perhaps considerably smaller than the variation we have now, but still very large indeed.

Therefore, what we need is a system that will put pressure on underperforming schools and schools serving historically low-achieving students – to increase the equity of educational outcomes between groups – while still sensibly and realistically acknowledging the large variability that will persist *within* groups. We have at this time no good models for this, in part because the policy community has not acknowledged the need for them. This gap in our knowledge should be especially worrisome to the civil rights community, which cannot afford to have its core demand for greater equity of opportunity held hostage to unrealistic expectations about the reduction of within-group uniformity. If between-group equity is not clearly distinguished from variability within groups, a failure to meet unrealistic expectations about the latter might lead cynics to become pessimistic about addressing the former.

What are the advantages and disadvantages of focusing on "percent Proficient?"

As one state official said to me recently when discussing how to report performance on his state's test, "Proficiency is the coin of the realm." And NCLB, of course, carries this to an extreme. The accountability apparatus of NCLB hinges on a single statistic: the percentage of students above the Proficient standard.

While reporting performance in this way does have one substantial merit – it helps to focus attention on expectations – it has numerous severe disadvantages. There is not space here to explain all of them.[xvii] However, I will briefly two that are particularly relevant to this meeting.

The first of these disadvantages is obvious: focusing only on percent Proficient leaves all other changes in the distribution of performance unmeasured. And perhaps worse, it leaves all of these other changes out of the accountability system. For example, take a state that has imposed a high standard for Proficiency. If one measures only percent above Proficient, all progress with students below that cut score, no matter how large, goes unnoted and unrewarded. Conversely, a school that makes very small gains among students just below the Proficient standard, just enough to get them over it, will be mistakenly credited with having effected major improvements. Many educators frankly admit that they use this fact as a way of gaming the system, focusing disproportionate attention on students near the standard and giving short shrift to students well below or well above it. There is even a common slang term for the students who are the focus of this approach: "bubble kids." Given the low performance characteristic of many of the schools that disproportionately serve minority and poor youth, this problem ought to be of great concern to the civil rights community.

The second disadvantage is not obvious: reporting in terms of standards distorts comparisons among groups. To be more precise, it distorts comparisons in trends among groups that differ in their initial level of achievement. This fact is a consequence of the distribution of scores, the fact that there are a great many students bunched near the average and progressively fewer as one moves toward high and low extremes of performance. For example, if African American and white students in a given state were making identical progress in a given state, measures such as "percent above Proficient" would create the misleading appearance of differences in their rate of gain.[xviii] This too should be of concern to the civil rights community.

WHAT SHOULD WE DO NOW?

The course we are now on is not working well, and over time, as the unrealistic targets we have set draw near, it is likely to work even more badly. But research has not yet provided us good alternative designs. So what is to be done?

First, we should complement in-school programs with out-of-school interventions. There is currently some debate about the proportion of the variance in test scores that is attributable to out-of-school factors, but there is no doubt that this proportion is large. It is therefore simply unrealistic to expect improved educational services to fully offset the

disadvantages faced by historically lower-scoring groups; even a very effective educational accountability system can be expected to fix only part of the problem. Interventions that go beyond educational accountability could include both supplementary services in school settings and entirely separate services – for example, high-intensity preschool services focused on cognitive development and language acquisition. Many of the needed interventions outside of the classroom are much more difficult and expensive than simply holding educators accountable for scores, but if we are serious about equity, they are necessary.

Second, we need to set more realistic targets for improvement. We need more research on performance targets, an issue to which I will return momentarily. But we already know enough to recognize that the current system is simply not sensible. We need to rely on such data as we have – international and other normative data, historical data, and program evaluations – to set targets that educators are more likely to reach by legitimate means.

Third, we need to use better metrics for reporting and rewarding performance on tests. We need measures that reflect improvement across the entire range of performance and that do not create perverse incentives to ignore students in certain ranges. If we are going to persist in using percent Proficient as part of our reporting and accountability system, we need to supplement this with other measures, and these other measures must count.

Fourth, we need to do all we can to lessen the narrowing of instruction that current test-based accountability systems produce – both the excessive focus on tested subjects at the expense of others, and the excessive focus on the content of the test at the expense of other important content within the same subject areas. Here too, we need more research, but we cannot afford to ignore the problem in the meantime. For example, states and local districts should not disseminate test-preparation materials that focus on test-taking tricks or inappropriate forms of coaching. They can encourage the vendors from which they purchase tests to lessen the unintended recurrences that facilitate coaching. (Test-prep materials can help them identify these recurrences.) Professional development activities can focus on the differences between good and bad test preparation. Principals and others can be on the lookout for undesirable reallocation.

All of these in theory might help, although given the incentives NCLB creates to raise scores at any cost – particularly for schools serving historically low-scoring students – it is naïve to expect their effects to be large until we develop a more reasonable accountability system. All actors in the system, from teachers to state chief school officers, currently have strong incentives to ignore this advice.

Fifth, we must stop taking score gains on high-stakes tests at face value. To be clear, research thus far does not suggest that all gains on all high-stakes tests are spurious. But it does unambiguously show that score inflation is not rare and that it can be very large, dwarfing true gains achievement. And currently, the data commonly reported do not allow us to distinguish routinely between the bogus and the real.

To address the fourth and fifth of these issues, we must begin seriously and routinely evaluating the performance of our accountability systems. This evaluation must include auditing of the gains on high stakes tests. This auditing and evaluation will have two benefits. First, it will give us better information about student performance. We should not tolerate a situation in which real improvements in equity are slowed by illusory gains. Second, it may substantially improve the incentives faced by teachers, thus reducing the gaming of the system that currently inflates scores and cheats children. As of now, teachers who inflate scores stand virtually no chance of getting caught, and admonitions to 'teach to the standards' rather than 'to the test' are empty rhetoric. But if educators know that auditing will sometimes expose score inflation, they will have more reason to avoid the shortcuts that cause it.

At the moment, NAEP can provide an audit measure in many states, but it is not available in most grades, and over-reliance on that one test may lead some people to start gaming NAEP as well. Routine auditing will likely require additional measures, either separate from operational tests or embedded within them. The key is that a good audit test is designed to measure the same core knowledge and skills as the accountability test, but it differs in many of the particulars – the details of content, format, scoring rubrics, and so on, that individual educators and test preparation firms capitalize on in finding ways to inflate test scores. Therefore, if students really learn more of the material, their performance gains will generalize to the audit test, but if their gains depend on the specifics of the accountability test, they will not be echoed on the audit measure. While

this principle is clear and simple, the practical details of constructing audit tests remain unexplored, for the simple reason that the people who buy tests have had no incentive to ask for audit measures. That must change, and the reauthorization of NCLB provides a powerful opportunity to do so.

Finally – and I would argue, most important of all – we need to start, belatedly, on a serious program of research, development, and evaluation to facilitate the design of better educational accountability programs, programs that will do more to improve the achievement of historically low-scoring groups while generating fewer negative side-effects. The list of important questions to which we lack answers remains daunting. How can we best pressure schools to reduce inequities while accommodating the inevitably wide variations in performance within groups? How can we better design assessment systems to reduce the problem of bogus gains in scores? How can we create a better mix of incentives for educators, one that will encourage greater effort with less narrowing of instruction? What types of formative assessments and other test preparation activities produce the largest gains in learning and the least score inflation? The list goes on.

Some people are now arguing that we have a solution at hand that circumvents the need for more R&D: value-added assessments. Value-added systems measure the gains students make while in a given grade, rather than tracking the improvements of successive cohorts of students in a single grade. Value-added systems indeed have many important advantages over the current system. For example, they are less sensitive to bias from differences in the characteristics of the students attending different schools, and they measure what many people consider a more appropriate variable for accountability: what a teacher or school teaches a group of students while they are in the school's care. Value-added assessments, however, confront a truly daunting set of difficulties. For example, if testing is only annual, value-added systems work only in subjects in which the curriculum is largely cumulative; they are highly error-prone, so most of the apparent differences among teachers or schools are measurement error rather than real differences in output; they are seriously problematic where there is substantial differentiation of curricula, for example, in most middle-school mathematics programs; the rankings they provide are not always consistent from one test to another; their results can highly

sensitive to arcane technical aspects of test construction, such as details of test construction and scaling methods; and their results are sometimes sensitive to choices in models used, many of which are extremely complex and not understood by most users of the data.[xix] None of these difficulties, as daunting as they are, argue against exploring value-added approaches as a part of educational accountability systems. But they do argue persuasively against accepting this approach as a new silver bullet that would once again free us from the hard work of rigorous research and evaluation. Children will again be cheated if we again make this mistake.

One of the most serious negative ramifications of NCLB is that it makes the R&D needed for long-term improvements in policy far more difficult to conduct. This may seem an odd claim, given that NCLB has encouraged the creation of vast databases of test scores. These data, however, are generally not useful for serious evaluation of alternative policies because of the problem of score inflation and, in many instances, the limitation of reporting to percents above standards. And the pressure created by NCLB makes experimentation with alternatives too risky. When everyone is in a race – often a desperate race – to raise scores on the few measures that count under NCLB, and to raise them continually, it is simply unrealistic to expect states, districts, and schools to agree to participate in R&D and experimentation. Experimentation runs the risk of smaller gains over the short term in the service of greater benefits over the long term, and this is a trade that NCLB makes very costly. This constraint is especially severe for schools serving the students who are the primary concern of this group, because those schools face particularly great pressure to raise scores by a large amount, and very quickly.

This, then, is my reason for arguing that we should not let concerns about changes to the details of NCLB – even the most important details – blind us to the need for a longer-term plan for creating better educational accountability systems. Building those better systems requires more systematic, empirical data, and that in turn requires a serious agenda of R&D. Whether this R&D is carried out over the coming years will depend substantially on whether the reauthorization of NCLB makes it more feasible. Because of the frantic race for score increases created by NCLB in its current form, serious R&D will be impeded unless the provisions of NCLB are substantially changed. One option might be to provide waivers from NCLB accountability provisions to jurisdictions willing

to do the needed, difficult work of helping to design the programs our children need and deserve. In addition, the incorporation of audit provisions into NCLB might also encourage R&D by reducing the inflated gains of jurisdictions to which the experimenting ones might be compared. And given the costs of the needed R&D – which are too large for many jurisdictions to take on – NCLB could include a mechanism for providing federal support for these efforts.

---

[i] For a discussion of narrowing and goal displacement caused by test-based accountability, see Richard Rothstein and Rebecca Jacobsen (2006, forthcoming), The goals of education, *The Phi Delta Kappan, 88*(4), December.

[ii] E.g., L. Zuckerman (2000). In Airline Math, an Early Arrival Doesn't Mean You Won't Be Late. *The New York Times*, December 26 [internet copy, not paginated]; B. McAllister, A. (1998). 'Special' Delivery in W. Virginia: Postal Employees Cheat to Beat Rating System. *The Washington Post*, January 10, p. A1; A. Hickman et al. (1997), Did Sun cheat? *PC Magazine*, January 6 [internet copy, not paginated]; P. H. Lewis (1998), How fast is your system? That depends on the test. *The New York Times*, September 10, p. E.1; J. Markoff (2002). Chip maker takes issue with a test for speed. *The New York Times*, August 27, p. C3; J. H. Cushman (1998). Makers of diesel truck engines are under pollution inquiry. *The New York Times*, February 11 [internet copy, not paginated]; P. Farhi (1996). Television's 'sweeps' stakes: Season of the sensational called a context out of control. *The Washington Post*, November 17, p. A01.

[iii] M. Santora (2005). Cardiologists say rankings sway choices on surgery. *The New York Times*, January 11 [internet copy, not paginated].

[iv] Hambleton, R. K., Jaeger, R. M., Koretz, D., Linn, R. L., Millman, J., and Phillips, S. E. (1995). *Review of the Measurement Quality of the Kentucky Instructional Results Information System, 1991-1994.* Frankfort: Office of Education Accountability, Kentucky General Assembly, June.

[v] Jacob, B. (2002), *Accountability, incentives and behavior:The impact of high-stakes testing in the Chicago public schools* (Working paper W8968). Cambridge, MA: National Bureau of Economic Research (May); Schemo, D. J., & Fessenden, F. (2003, December 3). Gains in Houston schools: How real are they? *New York Times* . Retrieved December 3, 2003, from http://www.nytimes.com.; Klein, S. P., Hamilton, L. S., McCaffrey, D. F., & Stecher, B. M. (2000). *What do test scores in Texas tell us?* (Issue Paper IP-202). Santa Monica, CA: RAND. Retrieved January 12, 2004, from http://www.rand.org/publications/IP/IP202/; Koretz, D., Linn, R. L., Dunbar, S. B., and Shepard, L. A. (1991). The Effects of High-Stakes Testing: Preliminary Evidence About Generalization Across Tests, in R.L. Linn (chair), *The Effects of High Stakes Testing*, symposium presented at the annual meetings of the American Educational Research Association and the National Council on Measurement in Education, Chicago, April.

[vi] R. L. Linn and S. B. Dunbar (1990), The Nation's report card goes home: Good news and bad about trends in achievement, *Phi Delta Kappan, 72* (2), October, 127-133; B. Fuller, K. Gesicki, E. Kang, and J.

Wright (2006), *Is the No Child Left Behind Act Working? The Reliability of How States Track Achievement*, University of California, Berkeley, Policy Analysis for California Education.; Lee, J. (2006). *Tracking achievement gaps and assessing the impact of NCLB on the gaps: An in-depth look into national and state reading and math outcome trends*, Cambridge, MA: The Civil Rights Project at Harvard University.

[vii] For an excellent overview of this research, see Brian Stecher (2002), Consequences of large-scale, high-stakes testing on school and classroom practice, in L. Hamilton, et al., *Test-based Accountability: A Guide for Practitioners and Policymakers*, Santa Monica: RAND (http://www.rand.org/publications/MR/MR1554/MR1554.ch4.pdf).

[viii] Koretz, D., and Hamilton, L. S. (2006). Testing for accountability in K-12. In R. L. Brennan (Ed.), Educational measurement (4th ed.). Westport, CT: American Council on Education/Praeger; Koretz, D., McCaffrey, D., and Hamilton, L. (2001). *Toward a Framework for Validating Gains Under High-Stakes Conditions*. CSE Technical Report 551. Los Angeles: Center for the Study of Evaluation, University of California.

[ix] For an interesting sample, go to the "cheating in the news" archives of Caveon Test Security at http://www.caveon.com/resources_news.htm.

[x] See also Rothstein and Jacobsen, *op. cit.*

[xi] Hambleton, et al., *op. cit.*

[xii] Koretz, D., and Barron, S. I. (1998). *The Validity of Gains on the Kentucky Instructional Results Information System (KIRIS).* MR-1014-EDU, Santa Monica: RAND.

[xiii] Lindquist, E. F., (1951). Preliminary considerations in objective test construction. In E. F. Linquist (Ed.) *Educational measurement.* Washington: American Council on Education.

[xiv] Linn, R.L. (2000). Assessments and accountability. *Educational Researcher, 29* (2), 4-16.

[xv] Koretz and Barron, *op. cit.*

[xvi] This information appears as a table of standard deviations in an Appendix to each of the TIMSS reports. See, for example, A. E. Beaton et al. (1996), *Mathematics Achievement in the Middle School Years: IEA's Third International Mathematics and Science Study (TIMSS),* Appendix E. Chestnut Hill, MA.: TIMSS International Study Center, Boston College.

[xvii] For an excellent discussion of some of the drawbacks of reporting performance in terms of standards, see Linn, R. L. (2003), Performance standards: Utility for different uses of assessments, *Education Policy Analysis Archives, 11*(3).

[xviii] Koretz, D. (2003). Attempting to discern the effects of the NCLB accountability provisions on learning. In K. Ercikan (Chair), *Effects of Accountability on Learning.* Presidential invited session, annual meeting of the American Educational Research Association, Chicago, April 22. For further discussion of this point and of some other drawbacks of reporting in terms of performance standards, see pp. 557-560 in Koretz, D., and Hamilton, L. S. (2006), Testing for accountability in K-12,iIn R. L. Brennan (Ed.), *Educational measurement* (4th ed.), Westport, CT: American Council on Education/Praeger, 531-578.

[xix] For a discussion of some of these issues, see McCaffrey, D. F., Lockwood, J. R., Koretz, D. M., and Hamilton, L. S. (2003). *Evaluating value-added models for teacher accountability*. Santa Monica: RAND, MG-158-EDU, http://www.rand.org/publications/MG/MG158/.