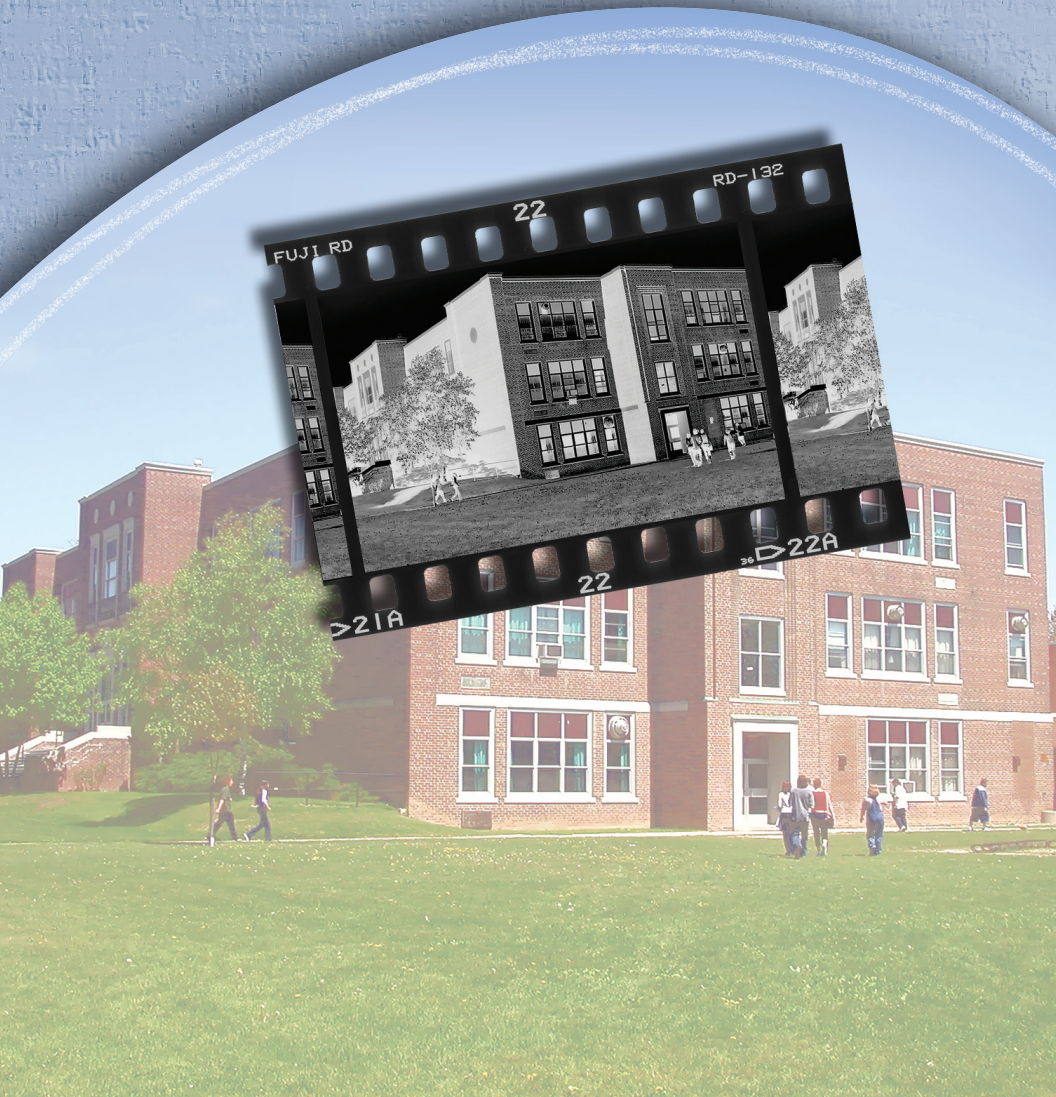
A Union of Professionals

"Failing" or "Succeeding" Schools: **How Can We Tell?**



A Union of Professionals

For more information, contact:
AFT Teachers / Educational Issues Department
American Federation of Teachers, AFL-CIO
555 New Jersey Ave. N.W.
Washington, DC 20001
202/879-4400

To order copies of materials:
Send a check payable to the American Federation of Teachers and mail to: AFT Order Department, 555 New Jersey Ave. N.W., Washington, DC 20001. Shipping and handling costs are included. Please include the item number and publication name.

"Failing" or "Succeeding" Schools: **How Can We Tell?**

By Paul E. Barton*

* The AFT has published this report to promote further discussion of the No Child Left Behind (NCLB) Act and related issues. The views expressed here are those of the author and do not represent official policy of the AFT or any of its affiliates.

Paul E. Barton is an education writer and consultant. He is a former director of the ETS Policy Information Center, and also served as an associate director of the National Assessment of Educational Progress from 1984 to 1989. His recent publications include *Unfinished Business: More Measured Approaches in Standards-Based Reform; One-Third of a Nation: Rising Dropout Rates and Declining Opportunities;* and *High School Reform and Work: Facing Labor Market Realities.*

Several people reviewed early drafts of this manuscript, providing helpful comments that led to a considerable number of revisions. I wish to thank: David Cohen, University of Michigan; Emerson Elliott, NCATE; Chester E. Finn, Jr., Fordham Foundation; Jack Jennings, Center on Education Policy; Michael Nettles, Educational Testing Service; W. James Popham, UCLA Graduate School of Education; Diane Ravitch, New York University; Bella Rosenberg, Consultant; and Dylan Wiliam, Educational Testing Service.

—Paul E. Barton
September 2006

Contents

Introduction

Standardized testing in the public schools has been around a long time. But the use of standardized tests has changed from time to time, and their quantity has exploded in volume as state laws first, and then federal laws, began to require testing for school accountability. Now the federal government requires testing in reading and math every year in grades 3 through 8 and once in high school; soon testing in science will be required.

Over the years in the educational community, standards have evolved to assure reliability and validity of standardized tests. These standards address testing for a variety of purposes: to estimate the knowledge and abilities of individual students at a point in time; to compare students and schools in "norm-referenced" systems; to sort students into tracking arrangements; to promote students to the next grade; to award student diplomas; and to select students ("gatekeeping") for college, graduate schools, professional schools and the military.

What has come to predominate K-12 testing, over the last couple of decades, is testing as a component of accountability systems for measuring school effectiveness. But it takes much more to develop the correct criteria for a total accountability and evaluation system than just a quality test that estimates accumulated knowledge and ability. Accountability systems, backed by strong sanctions that extend to closing down failing schools, are created to determine whether entire schools and school districts are effective. The entire accountability system must be of high quality, not just the tests within it. A test may be the most visible aspect of the system, but its focus is the overall effectiveness of a school.

This brief report summarizes how we have drifted into the accountability systems now in use, either under individual state laws or as mandated by the federal No Child Left Behind (NCLB) Act. The report goes on to describe how slip-sliding into our current accountability requirements has resulted in a system so flawed that it fails in its basic job of identifying which schools are ineffective and which are effective. Not only does this fall short of the intentions of the law—whether state or federal—but it leads to misidentifications with huge consequences for schools, teachers and students.

The nation has clearly embraced holding schools and teachers accountable; the question is how to do it. The objectives of NCLB and individual state accountability systems have broad support.

What, then, constitutes a responsible use of testing as a principal component of an accountability system? How can we get it right? This report highlights the emerging recognition that evaluating school performance with standardized tests requires measuring what students learned in the school during the year of instruction—a quite different matter from measuring the sum total of what students know and can do at a point in time. But to measure what students learn in school poses challenges, some of which are identified and discussed. These challenges are not to be left just to the measurement experts. Public officials and educators must be involved, and new measurement constructs must be understood by them—and make sense to them.

Beyond important choices to make about measuring gain during the year or knowledge at a point in time, other things also must be done right in a test-based accountability system. There must be proper alignment of tests, content standards and the curriculum delivered in the classroom. There must be assurance that the test itself does not become the curriculum to the extent that instruction is narrowed or constrained by what is easy and cheapest to measure.

Drifting Into Test-Based Accountability

This brief review of the events resulting in standards-based reform shows how one thing led to another. What had already developed by the time NCLB was passed became the basis for the federal sanctions-based accountability system; it was not designed from scratch. Building on the past resulted in problems of proper alignment and test validity, and in the even more serious matter of improperly measuring school effectiveness for the purpose of imposing sanctions.

The standards-based reform movement that began in the late 1980s started with the idea of specifying the content of instruction. An important contribution was made when the National Council of Teachers of Mathematics (NCTM) decided to spell out the "content standards" of mathematics instruction—what students should know and be able to do in mathematics.

As the standards for mathematics became known, a movement began to create national content standards in other subjects. The U.S. Department of Education, under Assistant Secretary Diane Ravitch, provided leadership and some modest funding.[1]

The development of standards, such as those for science, took much time and involved several organizations. The standards for history encountered large controversy. National content standards evolved and became starting points for each state to create its own. As it did, the American Federation of Teachers (AFT) and the Fordham Foundation evaluated each state and issued report cards on the rigor and quality of the standards. Albert Shanker, then president of the AFT, was a leader in advocating high standards and tests "with consequences."

End-of-year testing was used by states to determine if goals were being met, with either new tests the states created, tests created for them by testing companies or "off-the-shelf" norm-referenced tests purchased from commercial test publishers. States decided what scores on the tests represented an acceptable level of achievement, and this score (or "cut point") was often called the "proficient" level, although other words were sometimes used, and more than one level was sometimes identified. The National Assessment of Educational Progress (NAEP) provided a model. It set three "achievement levels" for grades 4, 8 and 12, reporting the results as "advanced," "proficient" and "basic."

The setting of national goals in 1989 and the establishment of the National Education Goals Panel kept reform in the forefront of the nation's attention. A National Council on Education Standards and Testing pressed for national standards and a national test—although they were never realized—and the New Standards Project created by Lauren Resnick and Marc Tucker was influential.

Marshall Smith and Jennifer O'Day's article, "Systematic School Reform and Educational Opportunity," set out a framework in which all the elements fit. The state of Kentucky entered the 1990s by throwing out its whole public education system and starting over. The Clinton administration was responsible for the passage of legislation called Goals 2000, which gave the standards movement more momentum.

The 1994 amendments to the federal Elementary and Secondary Education Act (ESEA) codified the emerging elements and concepts related to standards. Discontent with the progress of Title I of ESEA in reducing inequality led to a desire for change, and the standards-based reform movement became the model to build on. The 1994 amendments required each state to set content standards in reading and mathematics, align tests with these standards, and set "performance standards" aligned with content standards to answer the question of how much content a student had to master and what test score would represent student proficiency. The concept of adequate yearly progress toward proficiency was introduced, but without a specific formula or timeline.

The states were slow in their efforts to develop procedures for determining adequate yearly progress, and when ESEA was reauthorized with amend-

ments known as the No Child Left Behind Act in 2001, only a handful of states had approved plans. NCLB built on the development and structure that had already evolved, incorporating ESEA's 1994 amendments. And a standards-based reform movement that, in the beginning, focused heavily on defining content of what was to be taught morphed into a predominantly test-based accountability system—a system with a range of sanctions that progressed to closing down schools.[2] Some states had already turned in this direction before NCLB, which requires states by 2014 to reach test performance levels the states have labeled "proficient"[3] for all subgroups of students. States also are to show their planned trajectory of progress in reaching proficiency, and are held to the yearly targets they set. In the 2005-06 school year, testing requirements jumped to every year in grades 3 through 8, and once in high school. Testing requirements applied to all schools, although the sanctions applied only to Title I schools.

States varied considerably in how high they set the "proficient" level of achievement, but whatever levels were in place before NCLB took effect had to be met under the new law. States, at the time they set these levels, had different views about their intentions. Some saw this new proficient level as a distant goal and set it high. Some wanted a goal that seemed within reach sooner. Some used it as a requirement for passing students to the next grade, and had to make judgments about what proportion of students, as a result, might need to be retained. After the federal law became operative, it would be hard to change the definition of proficient; raising the bar would have considerable consequences, and lowering it would be widely criticized.

Standards for Tests. How would it be possible to know when a state had made enough progress to avoid the sanctions of NCLB? As was the practice before NCLB, student scores on the tests currently in use at or near the end of the school year became the measure—no matter what the tests were, or how much they met the alignment requirements set in the 1994 amendment.

Thus from the beginning, NCLB left the starting line with a problem of validity. The word "validity" should not be taken as some esoteric psychometric concept that only a few testing experts can understand. "Validity" simply means that a test should, in fact, do what it is intended to do. For a test to be valid, it must meet specific conditions. The results of a test—e.g., a score increase—must reflect mastery of the content standards to which the test is supposed to be aligned.

For test results to have meaning on whether standards are being reached, all elements of the system must be in alignment.[4] This is well-described in a guide prepared for the U.S. Department of Education by the Council of Chief State School Officers:

> Systems of performance standards and assessments must be created or selected and matched with the content. In an aligned system, all content standards must be accounted for in some manner...content standards, performance standards, and assessments must be aligned so that what is taught is what is tested and what is tested is taught.[5]

Many states do not meet all the alignment requirements specified in the law, and even when they do, they often do it only minimally.[6] In one recent study, the Fordham Foundation graded the degree of alignment between the tests and the content standards in 22 states and found that there was, on average, only "fair" alignment. The AFT, which has been a leading proponent of standards and tests to measure how well they are met, looked at all the states and concluded that 44 percent of them have tests not aligned to the standards. Achieve, Inc., also conducted in-depth studies of individual states.

Alignment requirements were put into federal law in 1994. Even though they are an integral part of the test-based accountability system, they are not part of what the NCLB amendments added to ESEA and thus have not been covered in evaluations of NCLB implementation, the major ones of which were conducted by the Center for Education Policy and the Education Commission of the States.

Testifying before the House Committee on Education and the Workforce prior to passage of the NCLB, ETS president Kurt Landgraf spoke directly to the need for alignment if test scores were to be used in an accountability system. He called for:

- State curricula linked to state standards;
- Instructional materials linked to curricula; and
- Assessments linked to standards.

NCLB required test scores to be used as soon as it was passed—regardless of whether the tests were aligned, appropriate (i.e., not off-the-shelf, norm-referenced tests) or based on properly set performance standards—to identify "schools in need of improvement," which would trigger NCLB's sequence of sanctions.

A test must measure what it is designed to measure. Once that condition is met, the test must then be used correctly to measure school effectiveness, which is a quite different matter.

Standards for the Use of Tests. Henry Braun and Robert Mislevy, in the March 2004 issue of *Kappan*, point out that constraints on the meaning of tests and test results are being ignored. This results from "intuitive testing:" a prevailing view that "a test is a test is a test" and "a score is a score is a score." Test theory and technology may be up to the task of creating tests that are valid for specified uses, but a large question remains as to whether the officials who demand tests and specify them in RFPs, legislate them and apply them in accountability systems with serious sanctions, are able to use the tests properly. Can standards for accountability systems be as high as needed for schools?

And beyond proper alignment, there is an even larger problem. Although the goal of state accountability systems and NCLB is to gauge the effectiveness of schools, the wrong measures generally have been chosen to make that determination. A performance expectation is set on a scale—the level the state has defined as "proficient." Student scores identify what percent reach that expectation. And progress—"adequate yearly progress," NCLB calls it—is measured by comparing scores of one group of students at the end of a school year with those of another group of students who completed the same grade in the prior year. Thus, the scores at the end of the eighth grade are compared with the scores of last year's eighth-graders, and those of eighth-graders in years before that. This means that accountability for schools is not about measured improvement for any individual student, or even a whole class of students. Test results are disaggregated by subgroups and each subgroup must meet annual targets for progressing toward proficiency—to be reached by 2014—as measured by end-of-year scores of this year's students compared with end-of-year scores of previous years' students.

But this year's students may have different population characteristics from past years' students. Population shifts may mean that some, or many, prior students were either less prepared or more prepared when they entered the eighth grade than this year's students—perhaps because of attending other schools, or having different out-of-school or before-school experiences[7] (see shifting populations, p. 17).

NCLB has made significant advances in ensuring that test results are disaggregated by race/ethnicity and income, and that the scores of minority students are not hidden in average scores. However, the law still measures and judges whether a school is succeeding or failing by measuring the status of student knowledge at a point in time compared with that of some prior year—not by measuring the learning gains of individual students. This is a huge problem.

The following illustration applies equally to all grades: A test at the end of, say, the eighth grade captures not what a student has learned at school during that year, but all the student knows about a subject—from all the experiences and conditions of life that are conducive to cognitive development and knowledge acquisition. Learning may come from parents (through reading to children frequently when they are young, for example); what happens in a student's home (doing homework and turning off the television, for example); summer enrichment experiences; and what students learned in child care, kindergarten and during the first seven years of school—wherever they went to school.[8]

It is important to understand how much students know by the end of the eighth grade, and by race, ethnicity and income. This tells us how well families, society and schools are doing in promoting achievement and reducing inequality. But this does not provide a measure of gain—a measure of how effective the school is in raising the achievement of students during the eighth grade. If schools are to be held responsible, the focus of testing must be on what happens in school—apart from outside influences.

The practice of medicine provides an analogy. A medical treatment program has the purpose of improving a person's health. But patients who enter that treatment program have a life behind them that varies in important respects related to their health: their lifestyle, their diet, whether they smoke, whether they are overweight, any diseases or disorders they already have and family history. The treatment is not judged by the patients' total health at its end, but by how much their health improved during the treatment period. Nor are the physician's and the hospital's competence judged by comparing the health of their patients after treatment with that of different patients during prior years, since the composition of the entire patient population may have changed in important respects.

Present accountability standards do not reliably sort out ineffective schools from effective ones.[9] Many studies have established that there is a low correlation between schools identified through the measures now in use and a measure of the gain in achievement during a particular school year. These results are summarized in an ETS report;[10] see the Appendix for excerpts from the studies. A large proportion of schools identified as ineffective and subject to sanctions might be found effective by a gain measure. Conversely, a large proportion of schools found effective by current standards could be found ineffective by measuring gain. Many schools with students from families with higher income and parental education, when gauged by value-added rather than an absolute standard, may not be doing nearly as well as their upper middle class clientele like to think, while schools with students from families of lower income and parental education may be doing a lot better.

In a percentage of schools in neighborhoods with stable populations, some students will start and finish in the same school, so that students in the fourth grade, for example, will have been in the same school in prior grades. In that case, a rising level of knowledge, based on end-of-year scores, may coincide with an increased gain of knowledge during the school year as compared to the prior year. Such schools may show progress on both measures.

Doing It Right. The educational measurement, psychometric, program evaluation and educational research communities agree that to judge the effectiveness of a program or treatment in education, there must be a measure of what the student knows when the program or treatment begins. This is likely one of the first things taught in Evaluation 101. No proposed study design for determining effectiveness lacking this measurement would get much beyond the door of the U.S. Department of Education's Institute of Educational Sciences.

One particularly good analysis and empirical comparison of the two approaches—status at a point in time and achievement gain—is by Steven W. Raudenbush. He concluded that currently used measures are "deeply flawed:" "Such measures are not plausibly valid indicators of the average causal effects of attending various schools." In other words, measures used now do not capture the effectiveness of schools. He goes on to say that value-added measures (i.e., measures of gain in achievement) "hold schools accountable for the learning that a student exhibits while under the care of the school."[11] This is the measure needed.

Another unfortunate result of using this single cut point on a scale approach—"proficiency"—is that the only concern with student achievement becomes whether students reach this single point on the scale, however high or low it may have been set. Do we not care about, for example, whether minority students in the top half are progressing? The achievement gap exists at all levels. Current practice now tends to parallel the 1970s' minimum competency movement. The two may differ in how high on the scale the minimum is set—and what label is put on it vary; "Proficient" is an attractive label. A measure of achievement gain during the year would apply to students throughout the achievement distribution. It would tell us whether students are gaining as much as they should in schools in well-off suburbs as well as in inner cities.

The accountability system not only fails to address the progress of upper tier students, it also does not check on students' progress in the bottom tier: those who are far below the proficiency cut point but may—or may not—be moving toward it.

Measuring Student Gain

So, all we have to do to correct the present mismeasure is to measure gain in achievement during the year? Raudenbush and others say that developing and using such a measure has its own set of problems, depending on how it is done. The basic proposition of this paper is that if school effectiveness is to be measured through a test, an acceptable way to measure gain must be found—*one that captures the results of the learning that happens in school.* Those who propose and pass laws using test-based accountability take on the responsibility for being sure to have the correct evaluation model and seeing that the measure is sound.

The discussion during the last year or two has centered on gain measures that involve tracking the same students over time, and measuring the yearly gain on a scale.[12] A few such models are in place and can be examined, such as those in Tennessee (under the leadership of William Sanders), and Dallas, Texas (under the leadership of Robert Mendro). A number of research papers have suggested different approaches, and quite a few people are now working on the issue. Some of this research is being conducted on very sizeable databases, where student scores have been looked at in terms of both status and gain. While the concept of measuring gain is clear, there is disagreement over the best way to do it.

My own publications over the last 10 years have argued that, if there is to be test-based accountability to evaluate the effectiveness of schools, measuring gain is necessary.[13]

We should find an approach to measuring gain that meets these criteria:

- Accuracy in capturing what was learned during the nine-month school year, and in a way that does not include differences in summer experiences;

- Alignment of the testing instrument with content standards and instruction; and

- Transparency that allows gain measures used to be understood by students, teachers, parents, administrators and legislators.

The developing view is that students must be tracked through a "student identifier," so that gain can be measured on a "vertical scale." This might be done in a couple of ways.

One way to measure gain vertically is to give a test that covers several years of a subject matter, so that a student's progress can be tracked over those years. This is sometimes referred to as a "stretch" test. Problems already exist in aligning a test for an individual grade with that grade's content standards—that is, designing a test to reflect what is supposed to be taught in a particular grade—so designing a single test that covers several grades is very challenging.

Another way to measure gain vertically is to "equate" the tests given at the end of one year to the tests given at the end of the subsequent year, so that a scale can be constructed on which the year-to-year gain in achievement can be identified. This could be just from one spring to the next spring.[14]

Some statistical models with these multiyear achievement scales may look good to statisticians and psychometricians, but not make sense to teachers in revealing what students learned from what the teachers taught. The kinds of general analytic scales that are necessary, according to W. James Popham, "supply teachers with no diagnostically useful information about which skills or bodies of knowledge a student has or hasn't mastered. …We need to find better ways of measuring students' growth for our adequate yearly progress analyses."[15]

There is another important problem with the value-added models that use only end-of-year scores: They do not take into account large differences in summer experiences. One study established that "the differential progress made during the four summers between second and sixth grades accounts for upwards of 80 percent of the achievement differences between economically advantaged and ghetto schools."[16]

The only way to meet the three criteria for measuring gain—one that can be understood, uses known technologies, and can be clearly aligned with

the content standards and curriculum for the year of instruction—is not part of the current discussion. This approach uses two forms of the same test, giving one at the beginning of the school year and one at the end. The use of before- and after-tests to measure improvement has been perfected for over 50 years.[17]

This approach requires more testing—both at the beginning and at the end of the year—but has the advantage of providing teachers with information on each student at the beginning of the year. The present system of end-of-year testing hardly helps the teacher in the instruction of students during the school year. Although testing at both the beginning and the end of the year will not be as true a diagnostic or formative assessment as a test designed specifically for this purpose, it will tell a teacher what an individual student does and does not know at the start of the school year.

To minimize the frequency of testing, "before and after" testing may not need to be used every year. Schools do not change quickly; indeed, this is a common complaint. Is it really necessary to measure quality of school performance in every year, in every subject, in every grade to promote school improvement? Is it really reasonable to have an accountability system that is on continuous autopilot? Can schools be chosen in such a way that they do not know when their turn for testing is approaching? Can some grades be tested in some years, on a rotating basis? Can schools and subjects be sampled, rather than applying 100 percent testing (a possibility that has been discussed before)? Such alternatives should be explored. Most likely, testing and scoring will improve—with more useful results and more confidence in the system—if there is less frequent testing for accountability purposes.

Repeatedly, some have said that every student should be tested every year to improve instruction. But accountability testing is designed as a "summative" assessment of what students know; it is not designed to be a "diagnostic" or "formative" assessment—i.e., one that tells a teacher what a student is doing wrong or isn't getting. And it is given at the end of the school year, too late to inform instruction during the year. In fact, the growing volume of accountability testing is diminishing the opportunity for testing by teachers to diagnose student needs, as accountability testing takes more time and as the focus of testing is narrowed to measuring school effectiveness. At the ETS Invitational Conference in October 2004—all on teacher and diagnostic assessments—Kurt Landgraf, president of Educational Testing Service,

argued, "We've got to stop using assessment as a hammer and begin to use it appropriately, as a diagnostic and learning tool." Research shows clearly that diagnostic assessment raises achievement; there is a strong case to be made for using this kind of testing with every student, in every subject, in every grade.

How Much Gain Should There Be? This discussion of measuring value-added, or growth and gain, has so far focused on why the present system wrongly identifies schools for applying sanctions, and what approach would be better. Developing of a measure of gain itself, making it operative, and doing so on a disaggregated basis for subgroups would be a huge step.

But an important next step has not been discussed: Standards for accountability must be set regarding how much growth should occur in a particular grade and in a particular subject. Students do grow in school; they know more math at the end of the year than at the beginning. The question is: How much should they grow?

The present testing and standards-setting system of getting all students to a designated proficiency level has been defended by saying that it applies equally to all subgroups—minorities and the majority, poor and non-poor. When the right standard has been set, the standard should apply to all students.[18] The right standard is one based on achievement gain during the year. When there is a standard for how much should be learned during a school year, it should be applied across the board.

To do this, there must be experience with measuring gain during the year. How much gain is typical in a particular subject in a particular grade? How much is "normal"? How much gain occurs in the classes of recognized top teachers, such as those certified by the National Board for Professional Teaching Standards? How wide is the distribution of average gain scores among schools? What approach should be used to establish standards for gain? Once these questions are addressed, then targets and requirements can be set based on what reasonably can be demanded.

Even when goals for gain during the school year are set as high as is feasible, performance gaps will still exist. Lower-achieving students, even if their growth during the year is as great as other students, will require larger investments. Knowing end-of-year scores, as well as gain scores, always keeps

before us the distance we have to go, whether in the schools or through policies that bring equality in experiences and conditions outside of school.[19]

When the standards for each year's gain are set, they should prevail in the system of accountability and sanctions. They should not be considered simply a yearly stepping stone to the current and wrong approach based on total knowledge at a point in time—an approach that does not measure school effectiveness during an academic year or some span of academic years. The expected yearly gain should apply to every year, including 2014, in keeping with the proposition that schools are to be held accountable for what students learn during an academic year.

Early in 2006, at the invitation of the U.S. Department of Education, 20 states were preparing to submit pilot "growth model" projects. These would use some sort of a gain measure. If carefully constructed, they may provide a knowledge base for converting to a gain measure for accountability. By May of 2006, the Secretary of Education had approved only two states' growth models—Tennessee and North Carolina.

However, states with these projects still must finish at the same place in 2014 as they would without the pilot project. States have always had leeway in determining a yearly trajectory to the cut-point score (proficiency) by 2014.[20] These projects do not represent a switch away from an inappropriate measure of school effectiveness and to a standard for achievement gain during a school year, if the same end point is required by 2014.

The distinction between the current method of comparing end-of-year scores against a cut point on a scale, and looking at whether a student's gain meets a standard set for gain, is not a distinction between soft and hard, or between lax and demanding, or between low and high expectations. A large proportion of schools may find a gain standard harder to reach than the current standard—or it may be the other way around.

Proof by Example

A vigorous defense of the current approach (using end-of-year scores and a cut point) often includes examples of how all groups can reach the proficient cut point, or how progress toward it is resulting in a narrowing—or disappearance—of the gap between the percent of majority and minority students reaching it. Such examples are not as simple or straightforward as they might seem, for the phenomenon reported can be the result of several situations or a combination of them.

Performance at the Top. While the correlation is low between adequate yearly progress and achievement gain during the year, some percentage of schools will be effective on both measures. Sometimes a school and its teachers will have hugely improved, and they may seem to have performed miracles. Baseball has its Babe Ruth and basketball its Michael Jordan; education has its Deborah Meirs and Jaime Escalantes—but these superstars are not evidence that all professionals can achieve the same results.

A Low Hurdle. The majority-minority gap—when measured by the percent of students reaching a set cut point, as state and federal accountability systems typically measure—comes about, in part, as a result of where the cut point is set on an achievement scale. Given a gap in the average score, and gaps up and down the distribution of scores, the higher the cut point, the larger the gap in the percentage reaching it. At some very low cut point, all students will reach it and there will be no gap. At a very high cut point, all or almost all students will be below it. As minority scores rise, more students will equal or exceed the cut point. Majority student scores also may be rising. The focus should be on whether the gap has changed in the *average* scores for minority and majority students.[21] With both majority and minority scores rising, the

question is whether the gap in the average scores is shrinking, staying the same or increasing.

Picking Low-Hanging Fruit. A rise in the percent of students reaching the proficient level leads to another concern. Under great pressure to raise this percentage, school personnel may decide to focus on students who score just below the cut point, since getting them over it will produce an increase in the percent reaching proficiency. Such an approach might mean no improvement for students already above the proficiency cut point or far below it, while raising only a relative handful of students who are near the cut point. One effort to measure the effects on achievement in public schools due to competition coming from private schools provides a concrete example.[22] An analysis of scores had to take into account the possibility that "schools affected by competition would target students who score just below the proficiency cut-off." The author illustrates this:

> Roughly 3 percent of students in any given year fail by only one point. If a principal were, for example, to entice one-third of such students to gain a single point, the performance composite would increase by a full percentage point, but the average student-level gain would be tiny and could even be offset by losses made by students safely above or below the proficiency cut-off.

Shifting Populations. The current measure compares students at the end of a year with students of previous years. Some communities have stable populations and some communities change. For example, a poor inner-city neighborhood may undergo a long process of gentrification—something happening in cities all over the country. Housing prices increase, and families with better income, education and/or occupations drive out families with lower income, education and/or occupations. Students in the new families are likely to be more advanced when they start school than the students who moved out, so schools in such communities will likely show better scores—even if there is no change in school quality. Conversely, neighborhoods may change the other way, with a large and perhaps gradual influx of less-prepared students replacing students in upwardly mobile families who move to the suburbs—and they may be of any race or ethnicity.

Community and Family Capital. The communities and families students live in vary, within both minority and majority communities, so it is not enough just to look at the scores of minority students and the achievement

gaps, to prove that the current measure of effectiveness is viable. There are differences in incomes and assets, and differences in the social capital among communities with low average incomes—a difference reflected in cognitive development during early life, achievement during school life and the richness of students' summer experiences. An intensive community-level study by Richard Nathan, director of the Rockefeller Institute of Government Studies, establishes this.[23] Communities vary by the proportion of families having one and two parents. It is no simple matter to "control" for such differences when making comparisons of student achievement, as in the present system.

Test Familiarity. In situations where the test has been "psyched out" and has become the curriculum, test scores will rise, possibly dramatically. This does not necessarily mean that achievement of the content standards has increased. The tendency of test scores to increase each year following the first year a new test is introduced, and then to drop back when another test replaces the first one, has been well-documented in research.

All such examples of "improvement" should be scrutinized carefully. Evaluation is a known science; claims about the effectiveness of any particular treatment intervention need confirmation. Of course, when such examples are cited, few in the nation at large would be equipped with the necessary data to make a judgment. Those reporting them in the public media and elsewhere, however, should check out the stories well before reporting them. And there needs to be a more general awareness of the many reasons why performance may in fact be exemplary—or may only seem to be.

Concluding Comments

Standards-based reform, with its emphasis on defining rigorous content and getting it into the curriculum, has morphed into assessment as the treatment—or at least the leading treatment—in a system of test-based accountability. All eyes are on the test. And if testing now is used chiefly as a hammer, the hammer is hitting the wrong nail. It is time to get back to content and curriculum, and to follow known methods of evaluating effectiveness in accountability systems, of measuring what students learn at school rather than measuring what students have learned in some other school, in earlier grades, before school began, after school and during the summers. If tests are to be the mainstay of accountability, the gain measured should be what happens during a school year, as determined through a test aligned with the content of instruction. The test should be as transparent as possible, and should be understood by students, teachers, the family and the public, as required in the three criteria discussed earlier.

This author sees no other way to meet these three criteria without a test aligned to content standards and instruction, given at the beginning and end of a school year, using two forms of the same test. Perhaps other ways can be found.

Measuring gain during the year will tell us what is typical, what the ranges are among schools, what might be considered normal, and what might be considered as excelling. From this information, we have to decide what is reasonable to expect, and then establish standards for gain across the achievement spectrum. The results must be disaggregated, and the standards set for gain must apply to all subgroups—minorities as well as majorities, poor as well as rich. Once we know how much yearly gain is needed to eliminate

gaps in achievement, the magnitude of the problem can be identified and necessary steps can be taken on all fronts—whether in the schools or out—to narrow and eliminate the gaps.

When experience in a particular school, district or state is offered as proof of the efficacy of a particular approach, the facts need to be carefully examined, as described in the section, "Proof by Example."

Careful thought should be given to whether an evaluation system based on tests is sustainable through every grade in several subjects in ever year, whatever the measurement system. Can a constant and universal test-based evaluation of effectiveness meet ordinary standards of what constitutes proof of effectiveness? Do institutions change that quickly, and do they need measuring every year? Can the accountability testing burden be relieved by implementing it on a sample basis, or a rotating basis, or a surprise basis, or once every three or so years?

Can testing time be freed up for diagnostic assessment that helps teachers in their instruction? The research evidence is clear that teacher use of diagnostic tests during the year can substantially raise student achievement. There is nothing wrong with testing every student every year, if it is the right kind of test and if it clearly will help teachers teach.

No testing agency or testing professional advocates using a standardized test as the deciding factor in high-stakes educational decisions—even if the right test is used correctly. Can this testing indicator be used to trigger an in-depth examination that could lead to sanctions on a school or a district, and at the same time identify the problems that need fixing?

In any event, measuring effectiveness through the use of gain scores is a necessary step forward, if test-based accountability is to measure school effectiveness. Testing for accountability, whether by state law or under NCLB, needs to switch from comparing scores of this year's students with prior years' students and meeting a score cut point on such a test, to measuring the achievement gain of individual students from the beginning of the school year to the end of it. And standards need to be set for how much achievement should be gained.

The bottom line is:

- We need to remember that standards-based reform is about becoming clear on what students should know, and improving curriculum and

instruction. Education reform is not testing; testing is for determining whether reforms are working.

- We need to take seriously what constitutes an appropriate test and the appropriate use of a test. If the law says there must be alignment of tests, content standards and instruction—as it does—test results are not valid until the alignment occurs.

- We need to hold schools accountable for what goes on in schools. The present method of comparing groups of students' end-of-year knowledge with different groups of students' end-of-year knowledge in prior years does not do that. We must measure gain during the school year. To continue with the current approach means failure to identify the failing and succeeding schools, and frequently sanctioning the wrong schools.

- When we do measure gain during the school year, we must measure in a way that is educationally sound; helps teachers teach; and is transparent to students, teachers, parents and policymakers. I argue that testing in the fall and spring, with tests aligned to the year's instruction, meets these criteria. And when it comes to quality in accountability, less is likely more. I suggest accountability testing be performed less often in any one school, freeing up time and resources for diagnostic tests that help teachers tailor instruction to individual students. Just measuring gain is not enough to create an accountability system. We also must determine how much gain is acceptable—i.e., standards for gain.

- We expect to have high standards for schools; we also should expect to have high standards for measuring whether schools are doing their job. At present, we do not.

The nation's state and federal executives and legislators are demanding better educational results. It would be ironic if the judgments of the nation's highly educated professionals in educational measurement, research and program evaluation were ignored in the name of raising educational standards. It is time to encourage conversation openers rather than conversation stoppers.

Appendix

Excerpts from Research Papers

Darrel Drury and Harold Doran
"The Value of Value-Added Analysis" *National School Boards Association Policy Research Brief 3*, no. 1 (Jan. 2003).

Today, in most states and districts, test score data are reported as simple snapshots of student performance, commonly referred to as current-status indicators. Such snapshots represent the average score for students enrolled in a district, school, grade level, or classroom assessed using percentile ranks, the proportion of students meeting a state—or district—designated performance standard, or other means.

Although useful in describing performance for a given student population in a particular year, these indicators may actually provide less information about school quality than the traditional input measures they have largely replaced. Indeed in the absence of other measures, current status indicators are invalid and potentiality misleading for several reasons. …Concerns such as these have led researchers, policy makers, and practitioners to focus increasingly on "value-added" analysis, an approach to analyzing and reporting test-score that address many of these pitfalls.

Whereas current-status measures report the performance of a group of students at a single point in time, value-added analysis focuses on the achievement gains of individual students over time (for example, from spring to spring).

Eric Hanushek
"Should the Federal Government Be Involved in School Accountability?" *Journal of Policy Analysis and Management* 24, no. 1 (2005): 171.

After describing his research showing achievement gains related to states with "consequential accountability," Hanushek explains:

By concentrating on aggregate student performance instead of just value-added of schools, the accountability systems provide rather blunt incentives to schools. ...the tracking of school improvement through the standards of "adequate yearly progress" has ignored information about individual student gains and has relied upon unreliable changes in aggregate scores. The emphasis on whether students "pass" or "fail" a state test does not provide sufficient incentives for student learning across the entire spectrum of student performance.

Martha S. McCall, G. Gage Kingsbury and Allan Olson
Individual Growth and School Success, National Evaluation Association (April 2004): 1-2.

If School A and School B had identical state test score averages, would you think that they were having similar success with their students?

Before you answer, consider that School A started the year with low performing students, and caused every one of them to grow twice as much as the students in School B. What do you think of the two schools now?

Current federal regulations use only the information in the first paragraph (status) to judge school success ...

The study used information from the NEW Growth Research Data Base one of the largest repositories of longitudinal student achievement data in the world. The study includes 840 schools from 22 states. Each school administered NWEA assessments to its students in spring of 2002 and spring of 2003. This allows the identification of student status, the score at a single point in time, and growth, an index of the increase in scores earned over a span of time. More than 270,000 students were involved in the study. . . Some of the primary findings include:

- Schools with similar status levels differ substantially in the amount of growth they cause in students.
- More than 20 percent of the schools with high status levels fall into the bottom quarter of schools in terms of the amount of growth they cause in their students.

- Several schools with low results at a single point in time cause as much growth in their students as the best high-status schools.

The results from this study demonstrate clearly that schools differ in the amount of growth they achieve. Inclusion of information concerning growth is essential for drawing a complete picture of school success.

Steven W. Raudenbush
Schooling, Statistics and Poverty: Can We Measure School Improvement?, ETS Policy Information Center (Feb. 2005): 6-7.

Under NCLB, school quality is indicted by the percentage of students that tests reveal as proficient in various subject areas at a given time. School improvement is the rate at which this percentage increases.

The problem is that if tests flawlessly reveal proficiency, equating percentage proficient with school quality cannot withstand serious scientific scrutiny. Evidence accumulated over nearly 40 years of educational research indicates that the average level of student outcomes in a given school at a given time is more strongly affected by family background, prior educational experiences out of schools, and effects of prior schools than it is affected by the school a student currently attends. To make this assertion is not to say that schools are unimportant or that educators should not be held responsible for their students' learning. Rather, this assertion reflects the reality that, at the time a student enters a given school, that child's cognitive skill reflects the cumulative effects of prior experiences. As that student experiences instruction, the quality of those experiences will begin to differentiate that child's knowledge from the knowledge of similar children who entered other schools with different instructional quality... While I believe the parents have a right to know how well their children are doing at any given time, static measures such as school mean proficiency levels cannot isolate the contribution of school quality, no matter how good the test.

If snapshots of average proficiency cannot reveal school quality, then changes in those snapshots cannot reveal school improvement. For example, this difference in levels of reading proficiency between last year's third graders and this year's third graders may reflect change in the student population served as much as any changes in instructional effectiveness.

Paul E. Barton and Richard Coley
Growth in Schools: Achievement Gains From the Fourth to the Eighth Grade, ETS Policy Information Center (1998).

Another view of the contrast between score levels and score gains is to look at state results over the same period from 1992 to 1996. Among the states participating in NAEP in both those years, Maine had the highest average score in eighth grade mathematics in 1996. Arkansas had the lowest average scores for both years. However, the gain in scores from the fourth grade to the eighth grade was 52 points in Maine and 52 points in Arkansas. Both states moved their students up by the same amount, from where they were when they began the fourth grade.

Joseph Stevens, Susan Estsrada and Jay Parker
"Measurement Issues in the Design of State Accountability Systems" (paper presented at the annual meeting of the American Educational Research Association, New Orleans, La. (April 2000): 14.

After describing the current practice of using successive cohorts of students to measure effectiveness, the authors caution:

> There is agreement in the methodological literature, however, that cross-sectional designs that study different groups of students can shed little light on learning improvement, or other aspects of change.

H. Goldstein
"Better Ways to Compare Schools," *Journal of Educational Statistics* 16, no. 2: Summer 1991: 91-92.

...it is recognized that intake achievement is the single most important factor affecting subsequent achievement, and that the only fair way to compare schools is on the basis of how much progress pupils made during their time in school.

Herbert J. Walberg
"Principles for Accountability Designs," *School Accountability: An Assessment by the Koret Task Force on K-12 Education:* 16, no. 2 ed. Williamson M. Evers and Herbert J. Walberg, Hoover Institute (2002): 161.

Policymakers increasingly recognize that value-added scores better indicate the school's or teacher's contribution to achievement than do test sores at a single point in time.

Keith Zvoch and Joseph Stevens
"A Multilevel Longitudinal Analysis of Middle School Math and Language Achievement," *Education Policy Analysis Archives* 11 (July 8, 2003).

The study discussed below was of a Southwestern school district with over 100 schools, serving a diverse student body of close to 90,000 students.

The present study showed that evaluations of school performance differ depending on whether school mean achievement or school mean growth in achievement are examined. …Evaluation of these estimates showed that the school mean level of performance was not strongly predictive of the school mean rate of growth. Correlation of school growth estimates were only 0.14 for mathematics and 0.41 for language. …Schools with low mean scores were in many cases the schools with the largest growth rate.

Endnotes

[1] An account of developments up to 1995 can be found in Diane Ravitch, *National Standards in American Education: A Citizen's Guide* (Washington, D.C.: Brookings Institution, 1995).

[2] Education reform was moving in this direction by 2000. See Paul E. Barton, *Facing the Hard Facts in Education Reform* (Princeton, N.J.: ETS Policy Information Center, 2001) and Paul E. Barton, *Staying on Course in Education Reform* (Princeton, N.J.: ETS Policy Information Center, 2002).

[3] States set a point on an achievement scale that designates where students can be considered proficient in a subject matter.

[4] Basically, "alignment" refers to coverage of the same content in the tests as in the standards and in roughly equal shares. Test-takers are asked to demonstrate the same level of complexity, cognitive demands and skill requirements called for by the standards; they must demonstrate a level of rigor or difficulty in performance expected by the standards.

[5] Linda N. Hansche, et al., *Handbook for the Development of Performance Standards: Meeting the Requirements of Title I*. (Washington, D.C.: U.S. Department of Education and the Council of Chief State School Officers, 1998): 21-22.

[6] For a summary of all the studies, including those of Fordham University, the AFT, and Achieve, Inc., see Paul E. Barton, *Unfinished Business: More Measured Approaches in Standards-Based Reform* (Princeton, N.J.: ETS Policy Information Center, 2005).

[7] An illustration: Arlington County, Va., faced this situation years ago because of immigration. School population rose sharply and individual schools, almost overnight, enrolled students from many different language backgrounds. Now the situation is reversing, with immigrant families buying their own homes farther out in the suburbs and changing the student population characteristics of Loudon County, Va., schools.

[8] For the conditions and experiences, school and nonschool, that research has found to be related to achievement, see Paul E. Barton, *Parsing the Achievement Gap: Baselines for Tracking Progress* (Princeton, N.J.: ETS Policy Information Center, 2003).

[9] NCLB applies to school districts as well as to individual schools; basically, the same considerations apply.

[10] Barton, *Unfinished Business*.

[11] Steven W. Raudenbush, *Schooling, Statistics, and Poverty: Can We Measure School Improvement?* (Princeton, N.J.: ETS Policy Information Center, 2004): 36.

[12] In this paper, I am discussing gain, or value-added, for evaluating school effectiveness. In the educational community, there is considerable discussion and investigation regarding the use of value-added measures to evaluate individual teachers. Such use involves different considerations; Braun discusses existing models and methodological challenges. See Henry Braun, *Using Student Progress to Evaluate Teachers: A Primer on Value-Added Models*, (Princeton, N.J.: ETS Policy Information Center, 2005).

[13] The last publication, *Unfinished Business*, summarizes many of the available research studies. Also, see particularly the work of Martha McCall, et al., involving 270,000 students; see Appendix, page 23.

[14] A full description of all the measurement considerations is beyond the scope of this brief paper. For example, even when measuring student gains within a year, should the composition of students in a classroom, and how that composition changes, be accounted for? For one early analysis, see Dylan Wiliam, "Value-Added Attacks? Technical Issues in Publishing National Curriculum Assessment," *British Educational Research Journal* 18, no. 4 (1992): 329-341.

[15] W. James Popham, "All About Accountability: Can Growth Ever Be Beside the Point?" *Educational Leadership* 63, no. 3 (Nov. 2005): 83-84.

[16] Donald P. Hayes and Judith Grether, "The School Year and Vacations: Where Do Students Learn?" Cornell Institute of Social Relations 17 (1983): 64, as quoted by Richard L. Allington and Ann McGill Frazen, "The Impact of Summer Setback on the Reading Achievement Gap, *Phi Delta Kappan* (Sept. 2003): 69.

[17] Dylan Wiliam, saying he understands the reason for choosing before- and after-testing to have testing and instruction aligned, in the U.S. context: "I am not totally convinced that cross-years' equating must reduce the instructional sensitivity of the tests. . . . In the UK, a vertical scale was created before subject content for the grades was determined. The effect of this was to focus the attention of the content people on progressions in learning. . . . If you are serious about value-added, one must design the curriculum to support this and not just the assessment." Personal correspondence, Feb. 6, 2006.

[18] This perhaps oversimplifies. Dylan Wiliam informs me that the statistical model that best fits the data is one in which student variability in gain increases with age. There are other differences to understand and take into account. Personal correspondence, Feb. 6, 2006.

[19] The specifics to be dealt with are the gaps in life experiences and conditions that mirror the gap in achievement. Fourteen such before- and during-school factors are distilled from research in Barton, *Parsing the Achievement Gap.*

[20] As Chester Finn has pointed out, some states have chosen to project less progress in the early years, pushing it off until later years and saying it is like taking out a "balloon mortgage."

[21] Better than the gap in the average scores would be the gap in the average for each quartile.

[22] George M. Holmes, Jeff Desimone, and Nicholas G. Rupp, "Friendly Competition," *Education Next* (Winter 2006): 70.

[23] Richard P. Nathan and David J. Wright, *The Flip Side of the Underclass: Unexpected Images of Social Capital in Majority African American Neighborhoods* (Albany, N.Y.: Nelson A. Rockerfeller Institute of Government, State University of New York, 1996).

INSIDE BACK COVER

A Union of Professionals

American Federation of Teachers, AFL-CIO
555 New Jersey Ave. N.W.
Washington, DC 20001
202/879-4400
www.aft.org
Item number 39-0468
SEPTEMBER 2006