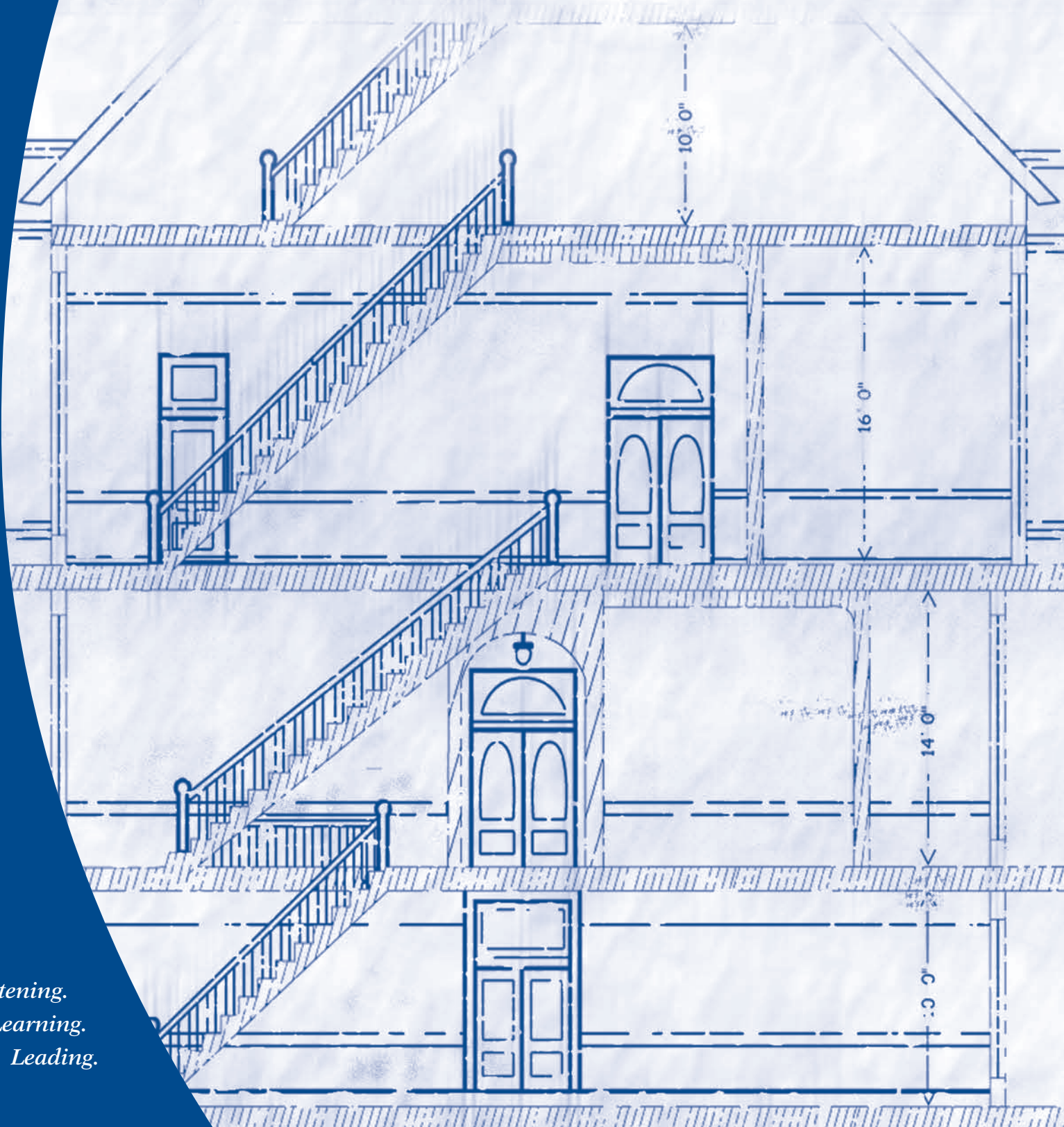
Policy Information Report

Unfinished Business:

More Measured Approaches in Standards-Based Reform



Listening.
Learning.
Leading.

This report was written by:
Paul E. Barton
Policy Information Center
Educational Testing Service

Additional copies of this report can be ordered for $15 (prepaid) from:

Policy Information Center
Mail Stop 19-R
Educational Testing Service
Rosedale Road
Princeton, NJ 08541-0001
(609) 734-5949
pic@ets.org

Copies can be downloaded from:
www.ets.org/research/pic

December 2004

Policy Evaluation and
Research Center
Policy Information Center
Educational Testing Service

Preface

Standards-based reform and test-based accountability have come to be the principal approaches to education reform in the United States, evolving and gathering momentum over the last two decades. As these approaches become ever more important to raising achievement, and as accountability systems become the basis for substantial sanctions and rewards to schools, teachers, and students, it becomes critical that we use the measures that will get it right.

The purpose of this report is to help in the evolution of these systems by examining the measures used, including, but not limited to, tests. The author asks: Are these the best measures? Are they used right? Are there other measures that should be employed? It is the model of reform itself that is examined, and the report does not address specific laws and policies, whether they be at the district, state, or Federal level. It is hoped, however, that the report will be useful to all who frame such laws and policies.

For those who are most interested in knowing what these recommendations are without delving into the supporting research, the Executive Summary reviews the report's key recommendations with the expectation that the reader will turn to the body of the report if more detail is needed. For those who wish a little more detail, the Report in Brief offers a distillation of the research and recommendations contained in the full report. For those who want to obtain the most complete knowledge, the body of the report discusses the supporting research at length and provides numerous references for further exploration. The author has offered, then, a full measure of accommodation to readers' interests and needs.

Michael T. Nettles
Vice President
Policy Evaluation and
Research Center

Acknowledgments

The author appreciates the thoughtful feedback, comments, and suggestions made by the following individuals: Henry Braun, Richard Coley, Michael Nettles and Dylan Wiliam of ETS; Margaret Goertz, Center for Policy Research in Education; Drew Gitomer, NORC; Jack Jennings, Center on Education Policy; and Robert Rothman, Annenberg Institute for School Reform. Lynn Jenkins was the editor; Loretta Casalaina and Susan Mills provided desk-top publishing, and Joe Kolodey designed the cover. Errors of fact or interpretation are those of the author.

At the different governmental levels where standards-based reform and test-based accountability are used as approaches to education reform, there is unfinished business. More and better measures are needed to make these approaches more effective and credible, and we need to be more measured in the criteria used for judging results.

- If we think of accountability as a structure, then the foundation of that structure consists of four walls: the content standards, performance standards defining expected levels of attainment, the curriculum (and the teaching of that curriculum), and the test. In many places where tests are used for accountability purposes, however, the alignment of these four walls is very often deficient in one aspect or another. When this is the case, the foundation is too weak to support the desired progress in achievement, the assignment of failure, the granting of rewards, or the application of sanctions that ensue from the accountability process. To remedy the situation:
  - States and districts pursuing test-based accountability must take advantage of the knowledge currently available to fix the structure—that is, to determine whether proper alignment exists and to improve where needed.
  - The tendency to cut to the chase and use test scores whether or not the required alignment is accomplished needs to be overcome.
- Typically, a single point on a scale—such as "adequate" or "proficient"—is used to interpret test results in accountability systems. But this approach is too limited to represent student achievement and progress adequately. It ties accountability only to the movement of a relatively few students around a point on a scale. Furthermore, the processes used to locate that single cutpoint do not produce a transparent alignment to the content standards. In addition to the percent "proficient," other information drawn from the test needs to be used to reflect what is happening all along the achievement distribution in a classroom, a school, a district, or a state. The results should make it possible to answer questions such as:
  - Have the average scores of students overall, and of those in various subgroups, improved over time?
  - What gains, if any, have been made by low-performing students (e.g., those in the bottom fourth of the performance distribution)? By high-performing students?
  - Is there evidence that the achievement gaps between students in various subgroups (e.g., by race/ethnicity or income) have narrowed?
- The *level* of achievement that students attain is the result of many factors, including not only what happens in school, but also what has happened in early childhood, home life, and after school. To hold schools and teachers accountable, we need to measure the results of what they do while students are in school. To do so means measuring the *growth* or *gain* in learning during the school year, and determining how that changes over time.
  - This "value-added" approach has been applied in some places, as well as in large research studies.
  - There are technical problems to overcome, however.
  - Schools showing success or failure as measured by the *level* of student achievement are very often not the same schools as those whose success or failure is measured by *growth* or *gain* in achievement during the school year. Measures of *both* level and gain are needed. Standards can be set for acceptable growth, as they are now for level.
- While the phrase "teaching to the test" is often used to refer to the problem of the curriculum being narrowed to what is tested, the phrase has taken on so many different interpretations and meanings that it is no longer useful.
  - Lack of agreement exists on what constitutes good and bad practice in preparing students for tests. Different standards are applied in different places, and there has been a lack of clarity in conveying to principals and teachers the desired and undesired practices.

- Legitimate concern exists about whether testing in only a few subjects—with high stakes attached to the results—impacts the curriculum. The concern is whether some subjects, or topics within a subject, get slighted. The distribution of instructional time needs to be measured regularly to gauge both intended and unintended changes.

- Accountability assessment is front and center in the education reform movement. While there has been a lot of rhetoric about how such assessment can help teachers identify students' individual needs, the tests are typically given at the *end* of the year, too late to inform instruction *during* the year.
  - What is needed is an expansion of diagnostic assessment use for helping teachers understand and address student needs.
  - Research has demonstrated that such use of assessments can significantly raise student achievement.

- The reform movement needs to broaden its attention from mostly focusing on *quality* to encompassing concerns about *quantity*. The actual rates of high school completion (ending with the award of a regular diploma) are lower than official rates disclose, and they declined for most states in the 1990s, according to the estimates of independent analysts. Measures must be improved at the level of the school, the district, the state, and the nation.

The bottom line is this: Dealing with unfinished business is essential in order to:

- Be more effective in reaching the important goals that have been set
- Maintain credibility in the use of assessments as a lynchpin in the reform movement
- Avoid unintended consequences, and
- Measure whether intended consequences have been achieved.

Education reform has proceeded apace over the last decade and a half, first taking the shape of what came to be called standards-based reform, and then merging with test-based accountability. The general model has similarities from place to place, but with considerable differences among the states. A recent evaluation of 30 states by the Fordham Foundation, based on six criteria, judged that the systems were only "fair" on average.

The assumption of this report is that the model is—and should be—still evolving based on experience, evaluation, and research. The purpose here is to add to the knowledge and information available for this evolution. It is not to critique or to advocate any particular local, state, or federal formulation.

Many factors are involved in improving schools and teaching, and sometimes there are competing strategies for doing so. This review is by no means all inclusive, although it does suggest some enlargement of the scope of what has become the standards and testing approach. The goal is to improve the measures of success in judging students, schools, and teachers, in determining whether intended consequences are being attained and unintended consequences are being avoided, and in providing more information that will help teachers improve instruction.

Alignment: A Necessary Condition

The tests used for measuring progress in a standards-based reform system must be closely aligned with the content standards that specify what students should know and be able to do. This is the cornerstone of such systems. If alignment is out of whack, there can be no confidence that changes in the scores are a valid measure for accountability. The same is true as regards alignment of the curriculum to the content standards, and of the tests to the actual instruction.

Of course, the need for alignment between actual instruction and the content standards goes beyond interpreting the test scores. The purpose of the content standards is to shape instruction and curriculum. If that is not happening, standards-based reform is not working, and students can hardly be expected to do well on tests where instruction is not aligned with the content standards.

Alignment of Tests to Content Standards. In a project coordinated by the Council of Chief State School Officers and led by Norman Webb, researchers have developed a systematic approach to facilitate alignment and check to see if it exists. Four criteria for alignment are involved.

- Categorical Concurrence—the extent to which both standards and the test incorporate the same content.
- Depth of Knowledge Consistency—the extent to which what is elicited from the students on the assessment is as demanding cognitively as what students are expected to know and do as stated in the standards.
- Range of Knowledge Correspondence—the extent to which a comparable span of knowledge expected of students by a standard is the same as, or corresponds to, what students need to correctly answer the test questions.
- Balance of Representation—the degree to which one objective is given more emphasis than another.

A somewhat similar set of criteria developed by Achieve, Inc. and Lauren Resnick has been used by that organization in a number of states to help them with alignment. In a five state study using these criteria, the authors concluded that although the states tended to limit their tests to material that is in the standards, and that—with some corrections—the test items were generally aligned to the objectives,

> . . . the good news ends here. With few exceptions, the collection of items that make up the tests we examined do not do a good job of assessing the full range of standards and objectives that the states have laid out for their students. What is included and excluded is systematic: the most challenging objectives are the ones that are under-sampled or omitted entirely. Thus, many of the tests in use by a state cannot be judged to be aligned to the states' standards—even though most of the items map to some standard or objective.

In a comprehensive study of the implementation of standards in 2001 by the American Federation of Teachers (AFT)—a leading proponent of standards and

tests with consequences—the conclusion was that 44 percent of states have tests that are not aligned to the standards. Yet alignment between the test and the content standards is critical for the test to be valid in its use in test-based accountability as part of a standards-based reform system.

When the Fordham Foundation rated six aspects of standards systems in the aforementioned study, one of which was alignment of the test to the state content standards, they found that the average rating for the 22 states was "fair," a 3 on a 5 point scale. Three states scored the high of 3.8, and two states were very low, 1.5 in Hawaii and 1.8 in New Mexico; in six of the states studied, not enough information was found to make a judgment. The lack of such information does not bode well for the possibility of alignment.

Alignment of Instruction to Content Standards. Such alignment is critical, and if the curriculum actually used in the classroom is faithful to the content standards, the test can help tell if the content standards are being mastered. States are at different stages of adjusting the curriculum and instructional materials to the content standards. The Survey of the Enacted Curriculum Project, carried out in 11 states, provides approaches to analyzing the degree of fit between what is actually being taught and the content standards. Among the questions asked was: How does math and science content taught in classes compare to the goals outlined in state and national standards? The answers:

- In middle grade math and science, most recommended standards are covered, but the level of expectation and depth of coverage vary widely among schools and classes.
- Data reveal differences in extent of teaching science content across the standards and the extent of articulation between the grades.

Another conclusion of the AFT 50-state study was that fewer than one-third of the tests in use are supported by adequate curricula.

Alignment of Actual Instruction with the Test. The 11-state study referenced above also investigated whether state assessments reflect what is being taught in classes. The study found that state assessment items cover a more narrow range of expectations for students than reported instruction, with tests focusing more on memorizing facts and performing procedures than on solving novel problems and applying skills and concepts. Teaching is broader than what is on the test, and should be. It is the test that needs to be changed.

In two states, it was possible to map the curriculum actually taught with the state test, enabling researchers to draw the following conclusions. For mathematics, "less than half of the intersections of content topics . . . reported by teachers were in common with the assessment items found on the state mathematics test." The same was found for science. The authors of the 11-state study say the results "can provide a database for monitoring the degree to which classroom curriculum is moving toward the standards."

The Passing Score: Performance Standards and Tracking Progress

The rubber hits the road in test-based accountability when the "passing score" is established. This is typically a point on a scale where a test score reaches or exceeds some level labeled "proficient;" this becomes the performance standard in the system of accountability. Such performance standards are supposed to be "aligned" with content standards, with performance standards somehow derived from the content standards, but an eminent scholar in the field contends that no approach has been developed for doing this directly. Several methods are used to set these standards, and have been around a long time. These are reasonably good when the purpose is to set a score for a particular occupation where experts from that occupation make the judgments, and where there is some consistency in the judgment they make. But those conditions do not exist in setting cutpoints, for example, for eighth grade mathematics:

- The cutpoints for setting performance levels on the tests of the National Assessment of Educational Progress (NAEP) have been labeled as "fatally flawed" by the National Academy of Education and the National Academy of Science. Therefore, by direction of Congress, each NAEP report contains a warning label indicating that the performance levels are "developmental."

- Different methods available for setting performance standards can produce quite varying results. Kentucky used three different methods with the result that either 56.6 percent, 29.4 percent, or 15.3 percent were at or above the proficient level. Kentucky took these results into account in deciding on a level to be required by the state.

No matter which one of the available methods is used, there is nothing in these procedures that provides a *direct* link to the content standards in terms of knowing what the score means with respect to how much of the required knowledge has been mastered. Beyond that, using a single cutpoint on a scale is inadequate as a sole representation of performance on a test.

- Using a single cutpoint for accountability means that decisions are based only on the movement up and down of a relatively few students below and above the cutpoint. Progress of the rest of the students is ignored.
- Improvement or deterioration in the achievement of students above the cutpoint, as well as those below the cutpoint, is not revealed by the accountability systems typically in use. The distribution of performance in the United States is very wide—wider than in any other developed country—and the distribution of scores within a school may be very wide, as well. We should therefore be concerned about whether the bottom quarter, or the top quarter of students, for example, are losing or gaining ground.
- Use of a single such cutpoint can be very misleading about the performance of an education system. (Of course, a single test, however used, should not be the sole basis of making high-stakes decisions.)

Take the case of NAEP assessment results in mathematics for Mississippi from 1992 to 1996. No improvement was seen over that period in the percent of students reaching the level of "proficient"' as defined by NAEP. However,

- The average score improved;
- The average score for the bottom quartile improved;
- The average score for the top quartile improved; and
- The gap between the top and bottom quartile was reduced.

Broader measures are therefore needed to capture the level of achievement of students and to compare changes in the levels of achievement over time. There are numerous options. One option would be to develop a composite of measures, which might include but not be limited to the percent reaching the proficient level. For an example, see Table 2 on page 23 showing several indicators of changes in eighth grade science achievement: the average score, the percent "proficient," the average for the bottom quartile, the average for the top quartile, the gap between the top and bottom quartiles, the gap between White and minority students, and the gap between the poor and non-poor students. In the case of eighth grade science, while no state declined on the basis of the percent proficient, several had declines in the bottom quartile, 11 had an increase in the gap between the top and the bottom quartiles, and 7 had an increase in the gap between poor and non-poor students.

Accountability For Growth Due To Schooling

Indicators of the level of student achievement over time, discussed above, are more relevant for measuring progress for the nation or state or community as a whole than for gauging school and teacher effectiveness. No matter which indicators are used, comparing this year's eighth graders with past years' eighth graders is a limited way to evaluate school effectiveness if there is any change in the demographic makeup of the eighth grade class, or if one class was either better or more poorly prepared than the other when it entered the eighth grade.

A lot of factors enter into how much mathematics a student knows at the end of the eighth grade, and a lot of factors enter into what this year's eighth graders know compared to eighth graders five years ago—factors that have nothing to do with how well teachers taught over the prior school year. Schools judged adequate when measured by changes in the *level* of student achievement over time, as is typical at present, are often not the same schools that are judged adequate by changes in the amount of *growth or gain* that take place within the school year, sometimes called the *value added*.

- Among the states participating in NAEP from 1992 to 1996, Maine had the highest scores in fourth and eighth grade mathematics, and Arkansas had the lowest scores. Yet students in Maine and Arkansas both gained 52 scale points from the fourth to the eighth grade. So, does Arkansas have as effective a school system as Maine?
- Tennessee for well over a decade has had a system based on gain scores as well as on levels of achievement. For contrast, see the results in Bradley County for 2003 on grades given by the state for performance in six subjects. The grades for the level of achievement were one A, three Bs and two Cs. For gain, the county had two Cs, two Ds and one F. In other words, the county had high achieving students, but they were lagging in how much they were improving. The opposite was true in other counties. The Tennessee system, which was changed recently by the state, represents only one of several ways to measure gain.
- A recent research study conducted in a large urban school district in southwestern United States examined middle school students' test scores. When the researchers compared the difference between the mean achievement of students in the same grade over time, and the growth in achievement of the same students, they concluded that:

  > evaluations of school performance differs depending on whether school mean achievement or school mean growth are examined . . . Evaluation of these estimates showed that the school mean performance was not strongly predictive of the school mean rate of growth . . . characterization of school performance is substantially different depending on whether mean achievement or mean growth is examined.

- The Consortium on Chicago School Research, led by Anthony Bryk, has been using a "Learning Gain Index" to produce a school productivity profile for about a decade. According to Bryk, the approach stems from a belief that "a school should be held responsible for the learning that occurs among students actually taught in the schools."
- A study of 230,000 students by the Northwest Evaluation Association found that many schools with high scores had low growth in achievement during the year.

Given these problems, many researchers have advocated the use of gain or "value-added" measures to evaluate the effectiveness of schools and teachers.

- Goldstein, describing school effectiveness studies in Britain, indicates that "it is now 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."
- Walberg points to an increasing recognition "that value-added scores better indicate a school's or teacher's contribution to achievement than do test scores at a single point in time . . . non-value added scores, however, can complement value added scores, and together, they give policy makers more information and are less misleading than either one alone.
- Lowery and Kubzdela write: "Currently, the most accurate, accepted and utilized method for measuring teacher quality is *value added* for *gain analysis* . . . value-added analysis follows the progress of individual students by tracking changes in their test scores from year to year."

The measure of *growth* and *gain* is necessary for holding schools and teachers accountable and determining their effectiveness. However, there are choices to be made as regards the best way to do this and methodological problems to be dealt with. The measure of *level of achievement* and its change over time tells us how well the state, community, or nation as a whole is doing with regard to progress in education achievement. Both measures are needed. And as Stephen Raudenbush cautions us in a new report from the ETS Policy Information Center, a test alone, whichever approach is used, needs to be supplemented by other information if high-stakes decisions are to be made.

Teaching and the Test

Two aspects are examined. The first is how instruction is changed to prepare students in tested subjects. The second is what happens to instruction in non-tested subjects.

Teaching in Tested Subjects. With 11,000 entries in Google, the term "teaching to the test" has been used to denote so many different situations that it has become virtually useless in conveying any common meaning. There are shades of gray in how much instruction is specifically tailored to what appears or is expected to appear on a test, and differing judgments about what is desirable, what is educationally sound, and what is legitimate versus what is cheating. It is not a simple matter for the teacher to know what to do, or for a principal to know what to encourage teachers to do. When a test score has real consequences attached to it, distortions in or departures from "regular" teaching can be expected.

- Efforts to rank practices from "good" to "bad" disclose disagreements. Readers have to make their own judgments, outright cheating aside. Although the use of practice tests was ranked as bad by some in a national survey, over half of teachers used such tests from their state a great deal (24 percent) or somewhat (28 percent) to help students prepare for the state test. While instructing students in test-taking skills was ranked as suspect by others surveyed, most teachers used this test-preparation approach a great deal (45 percent) or somewhat (46 percent).
- Clarity by education officials up and down the hierarchy is critical, so that teachers operate in a situation of known standards and expectations.
- Where there is poor alignment between the test and curriculum actually in use, a not infrequent case, and where there is limited alignment between the test and the state content standards, then even when the curriculum is aligned to the standards, teachers are on the horns of a dilemma. How do they deliver on test scores without finding ways to prepare students for material not covered? What is and is not appropriate under such circumstances—circumstances not created by the teacher?
- An understanding is needed of when a score result can be relied upon to represent a degree of achievement of the standards and when it cannot. When does a blood pressure reading represent real blood pressure and not a distortion due to the way it is taken? More measured approaches can be used to check whether the test results from these large scale test operations really represent the full domain established by content standards.

Teaching in Subjects Not Tested. Concern is frequently expressed about whether subjects not covered by test-based accountability are being neglected. Whether the objective is to continue the emphasis on the other subjects or to reduce instructional time in them, measures are needed that permit judging what is happening. And educational authorities need to be clear about what they do and do not expect.

- In Anne Arundel County, Maryland, after the school administration reduced some middle school offerings, the Coalition for Balanced Excellence in Education succeeded in getting the decision overturned at the state level.
- In North Carolina, the past president of the state's Council for Social Studies said that, with testing in just a few subjects, ". . . social studies is left behind because there is not testing."
- In 2003, *The Wall Street Journal* ran an article entitled "Schools Say 'Adieu' to Foreign Languages."
- A Boston College survey found that in states with high-stakes testing, one-fourth of the teachers reported cutting back in untested subjects.
- A four-state study by the Council on Basic Education found increases in instructional time devoted to subjects tested, and decreases in many subjects not tested.

A more measured approach involves the kind of tracking of the distribution of instructional time that would permit education policy makers to assure themselves that there are not unintended consequences for subjects not covered by test-based accountability—and intended redistributions in tested subjects.

Assessment to Inform Instruction

In the current standards-based reform model, testing is used for accountability, with tests given at the end of the school year to evaluate teachers and schools. From its earliest beginnings, however, testing has had the promise of being used *during* the school year to inform instruction and help teachers identify and address students' individual instructional needs. Such testing—"formative" and "diagnostic"—needs to be given a central role in education reforms.

A synthesis of research on the achievement effects of such assessment to inform instruction reveals a substantial impact, with "effect" sizes ranging from 0.4 to 0.7.

- An effect size of 0.4 would mean that the average pupil involved in an innovation would record the same achievement as a pupil in the top 35 percent of those not involved.
- An effect size gain of 0.7 in the recent international comparative studies in mathematics would have raised the score of a nation in the middle of the pack of 41 countries (e.g., the United States) to one of the top five.

The standards-based reform movement has not specifically encompassed the use of testing for these formative and diagnostic purposes. With increases occurring in the use of tests for accountability, such use could even be diminishing. There are many examples, however, of this testing.

- The Council of Chief State School Officers found that one distinguishing characteristic of the five high-performing schools they studied is that the staff at each school use standardized assessment data "to identify areas where students can improve and where their own teaching strategies can be adjusted to meet students' needs."
- A California study identified 16 schools with high performance of minority students and 16 with low performance, all schools with similar socioeconomic characteristics. One key characteristic of the successful schools, as compared with the others, was the frequency of testing to guide individual student instruction.

Measuring School Completion

Standards-based reform and test-based accountability are focused on raising achievement levels of students. In that context, good statistics are needed on how many students leave school without getting a regular high school diploma, and education reform ought to be about quantity as well as quality. The questions are always asked: Will the higher standards result in more students leaving school? Will higher standards keep more students in?

We need to know more about the terms of trade between higher standards and school completion, so that informed decisions can be made. Of course, even when the terms are known, judgments will differ on how to strike the balance. While much is written, little is known for certain about whether there has so far been any widespread impact on school completion with a diploma. Schools and students are, as might be expected, under considerable pressure.

- Independent estimates of non-completion rates at the national level indicate that completion rates have fallen over the last decade. But what this is related to has not yet been established. By itself, it is a very serious matter.
- There are instances of students being dismissed from school because of their poor prospects for meeting standards—or, more likely, transferred to a GED preparation class where their achievement scores are not counted in the accountability system.
- New York City recently settled a lawsuit in which one of its schools was charged with discharging poor performing students. These students are being readmitted. Two other suits are still pending.

Official government measures, whether state or federal, have come in for considerable criticism from researchers over the past several years. These rates typically have not shown a decline, for a number of reasons. This report explains and compares high school completion estimates, state by state, using five different methods, including one by the National Center for Education Statistics and one by this author. It also shows the completion rates submitted by the states to the Department of Education in September 2003, as required by the No Child Left Behind Act. A few examples will illustrate the comparisons.

- For Georgia, the NCES calculates a high school completion rate of 71 percent, compared with 54, 54, 57, and 58 percent under the four independent methods; under NCLB, Georgia reported 62 percent.
- For New York, the NCES reported 82 percent, compared with 70, 60, 67, and 65 percent; under NCLB, New York reported 75 percent.

- Much less variation occurred in some states. For example, in Idaho, the NCES rate was 77 percent, compared to 78, 75, 71, and 73 percent; under NCLB, Idaho reported 77 percent.

As for state-by-state trends, the NCES estimates have been the only ones available, and they are not available for all states. Few states were in this system a decade ago, so trend comparisons over a decade are not possible. This author made estimates for the period from 1990 to 2000. While five states raised their high school completion rates over that decade, the general pattern was one of declining rates:

- 16 states declined up to 3.9 percentage points;
- 18 states declined from 4 to 7.9 percentage points;
- 9 states declined from 8 to 11.9 percentage points; and
- 1 state declined 13 percentage points.

Challenges and possibilities are offered in the report. The U.S. Department of Education is now working to improve the estimates, as should each state. Education reform should be both about the *quality* and the *quantity* of educational achievement. More measured approaches need to be applied to the matter of quantity.

In order for the standards-based reform model to remain credible and continue to evolve, it will be necessary to attend to unfinished business. Content standards must be fully translated into curriculum and instruction, all components of the system must be aligned, and accountability assessment approaches must be used that better measure what is learned based on what is taught in the classroom and that measure gain in achievement as well as level. Clear understandings need to be conveyed to schools and teachers about what is correct and educationally sound in getting students ready for tests. Standardized testing for informing instruction *during* the school year needs to become an integral part of the model, as does measuring the distribution of instructional time among subjects and tracking changes—particularly among tested and untested subjects. The quantity as well as the quality of education needs to be attended to, as well, with better measures of high school completion and a better understanding of why completion rates are so low—and falling.

The bottom line is that we badly need better and more comprehensive measures if we are to have an accurate picture of how the system is functioning, where more effort is needed, or where policy adjustments are required. What is at stake are effectiveness, credibility, and the avoidance of unintended consequences.

Introduction

Over the last 15 years or so, the standards-based reform model, increasingly accompanied by test-based accountability, has emerged to become the principal approach to improving public education in the United States. According to this model, education systems must determine what students should know and be able to do (content standards), decide how much competence they should demonstrate (performance standards), align curriculum and instruction with the content standards, and conduct testing to measure if students are learning the desired content.

The rigor and style with which standards-based reform is being implemented in different states and localities have varied considerably. In some places, the model is being carefully applied. In other places, it is being done poorly, and unintended consequences have resulted. The No Child Left Behind (NCLB) Act, passed in 2001, spurred the model's dissemination and will probably result in its being applied more uniformly across the states, particularly in the use of testing for accountability purposes.

This author, in previous publications for the ETS Policy Information Center, has traced the emergence of the standards-based reform model, as well as the role of assessment in U.S. education.[1] It was the underlying assumption of these reports that the model, which originated in the 1980s with the standards established by the National Council of Teachers of Mathematics, was a good one that could be applied at the school, district, or statewide level. These reports also had a considerable amount to say about the many variants of the emerging model, and the constructive use of standardized testing within it.

This new report similarly accepts the current reform model and its four main components: content standards, performance standards, curriculum and instruction, and testing. How these are structured, and at what level of governance, leaves a lot of area for debate and disagreement, however.

If standards-based reform is the house that is being built, then this analysis is intended to aid in the design and construction of the building. The aim is to lay a strong foundation so that the structure is not continually swaying in the winds of disillusion, distrust, and disgruntlement. Thus, this report addresses the standards-based reform model generally, rather than its formulation by one district or one state or the nation as a whole. This is not, for example, a critique of the now long standing Kentucky model, or of the NCLB Act. The title, "More Measured Approaches," refers to the need for measured steps based on a thorough understanding of potential missteps and consequences. It also refers to ways of monitoring whether the expected changes are, in fact, occurring, as well as whether side effects that are *not* intended are emerging.

Specifically, the report addresses six key areas.

First is the *alignment of the test with the state content standards, the actual curriculum in the classrooms with the content standards, and the test with what is actually taught*. Of course, complete alignment of the first two would assure the third. Without alignment, the meaning of test scores, and changes in scores over time, are called into question, as are accountability decisions made with such testing. Test scores are being used, with serious consequences, where alignment does not fully exist. This, too, will have serious consequences.

Second are the *processes used to define the level of student performance that is sought*, and to determine the adequacy of performance on the accountability tests, and to indicate whether progress over time is or is not being made. There are, it is argued, serious weaknesses in current practice.

Third are *considerations involved in assigning responsibility for student achievement, and lack of it, through accountability testing*. How is the effectiveness of teachers and schools being measured? How can measures be developed that hold teachers accountable for what they teach in a given year, as opposed to measures that reflect all that students learned from prior teaching or life experiences?

[1] Paul E. Barton, *Too Much Testing of the Wrong Kind; Too Little of the Right Kind in K-12 Education,* Policy Information Perspective, Policy Information Center, Educational Testing Service, March 1999; Paul E Barton, *Facing the Hard Facts in Education Reform,* Policy Information Perspective, Policy Information Center, Educational Testing Service, July 2001; and Paul E. Barton, *Staying on Course in Education Reform,* Policy Information Perspective, Policy Information Center, Educational Testing Service, April 2002.

Fourth is *"teaching to the test,"* and the need to assure that the education sought is actually occurring, as interpreted through test scores, based on validity studies. Also, there is the question of the desired and undesired effect of accountability testing on the scope of what is taught, with particular attention to differences between tested and untested subjects.

Fifth is an element that is not specifically addressed by the model: *the use of testing for more than accountability*. This concerns the use of tests in the classroom during the school year to aid instruction. The terms frequently used are "formative" and "diagnostic" assessment, as contrasted with "summative" assessment used for evaluation of schools and teachers in an accountability system.

Sixth is *the systematic assessment of school completion*. Because current measures are inadequate, too much remains unknown about the relationship between higher standards and graduation rates. Some believe that higher standards will increase the number of students dropping out (or being "pushed" out), while others dispute this. Some believe that higher quality is desirable even if fewer students may achieve at the higher level.

The following sections of this report address these issues in turn, with the goal of assisting those who are constructing laws and policies related to the standards-based approach. The underlying assumption is that the model is—and should be—continuing to evolve, based on experience and expanding knowledge.

The tests used for measuring progress in a standards-based reform system must be closely aligned with the state or district content standards that specify what students should know and be able to do. This is the cornerstone of such systems. If the alignment is out of whack, there can be no confidence that changes in test scores are a valid measure of accountability.

Even if the test is closely tied to the content standards, however, performance on the test is meaningless if the curriculum and instruction do not give students the opportunity to achieve those standards. Students cannot be expected to do well on a test that measures content they have not been taught. It is therefore also essential for curriculum and instruction to be aligned with the content standards and the test.

While the integrity of the accountability system depends on proper alignment, and there are tools and approaches for achieving it, there are undoubtedly challenges to be met. Given that content standards tend to be quite broad and varied, for example, how can a single test encompass it all? We want teaching to be broad, not narrowed because of problems in translating between standards and tests.[2]

This section addresses methods developed to create the necessary alignment and measure the degree of fit. It also discusses the still limited (but growing) body of information on the extent to which alignment has been achieved.

Of course, the need for alignment goes far beyond giving meaning to the test scores. The purpose of the content standards is to shape instruction and the curriculum. If that is not happening, standards-based reform is not working.

Alignment of Tests to Content Standards. As noted earlier, alignment between the test and the content standards is critical for the test to be valid in its use as part of a standards-based accountability system. The purpose of the test is to measure the degree of achievement of the standards. The point is made forcefully in the American Educational Research Association's (AERA) *Research Points*:

> Today's calls for alignment are built upon a foundation of more than 70 years of research on the development, evaluation, and use of tests. *Standards for Educational and Psychological Testing*, the recognized authority on educational testing, stresses that a "valid" test must show that it actually measures the constructs—knowledge, skills, abilities, processes, and characteristics—it was intended to measure. When a test is used to measure the achievement of curriculum standards, it is essential to evaluate and document both the relevance of a test to the standards and the extent to which it represents those standards.[3]

Similarly, researcher Robert Rothman has emphasized that the validity of the test results depends on proper alignment:

> If a test measures only some of the expectations the standards hold for all students, can a score on a test truly represent a measure of performance against the standards? . . . If a test measures only some of the knowledge and skills expected for all students, what does a passing score indicate? Does it mean that students who attain the score have demonstrated proficiency on the test or on the standards?[4]

The validity problem is especially serious when standardized norm-referenced tests are relied upon for accountability purposes, with stakes attached to the results. The U.S. Department of Education's *Handbook for the Development of Performance Standards* specifically warns against this practice, since "it is unlikely that any 'off-the-shelf' test will fully align with the breadth and depth of a state's or local system's content standards."[5] Furthermore, the results of norm-referenced tests compare students with other students and schools with other schools, rather than indicate what

[2] One way to ensure broad coverage of subject matter on a test is to use "matrix sampling." In the National Assessment of Educational Progress (NAEP), for example, different samples of students take different sets of items; the compiled results accomplish what a single test of several hours would reveal.

[3] *Research Points*, Spring 2003, Volume 1, Issue 1, p. 4.

[4] Robert Rothman, "Benchmarking and Alignment of State Standards," *Redesigning Accountability Systems for Education,* edited by Susan H. Furman and Richard F. Elmore, Teachers College, 2004, p. 112.

[5] Linda N. Hansche, with contributions by Ronald K. Hambleton, Craig N. Mills, Richard Jaeger, and Doris Redfield, *Handbook for the Development of Performance Standards: Meeting the Requirements of Title I*, prepared for the U.S. Department of Education and the Council of Chief State School Officers, 1998, pp. 21-22.

students know relative to what they are supposed to know.

Accepting that there must be alignment between the test and the content standards, the question then becomes, what is the current status? Are problems of misalignment widespread, or are they rare? According to the AERA review of alignment studies cited earlier:

> While specific findings may vary from study to study, all of the research points to one central conclusion: Alignment needs to be improved. For some extreme cases, studies have found that alignment between state standards and tests is so weak that the standards from one state more closely match the tests used in another state.

Other research efforts have also concluded that alignment problems are fairly common. For example, a comprehensive study of the implementation of standards in 2001 by the American Federation of Teachers (AFT)—a leading proponent of standards and tests with consequences—concluded that 44 percent of the states have tests that are *not* aligned to the standards.[6]

In another recent study, the Fordham Foundation graded the degree of alignment between the tests and the content standards in 22 states on a scale of 1 through 5 (with 5 meaning outstanding performance, and 1 meaning poor performance). Researchers found that the average rating was 3, or "fair." The lowest states were Hawaii (1.5) and New Mexico (1.8). The highest were Georgia, Pennsylvania, and Virginia, all with 3.8. All of the top-performing states were using tests that were both state-developed and criterion referenced.[7]

Some researchers have developed criteria that are not only useful in evaluating alignment between tests and content standards, but also in improving that alignment. One important project of this nature has been the work of Achieve, Inc. staff and Lauren Resnick of the University of Pittsburgh. They began by examining the typical process that test developers use to show, usually in a matrix presentation, how the items or tasks on the test match the statement in the content standards. The authors argued that while "this [approach] seems pretty straightforward, . . . it masks myriad difficulties."

At the request of several states wanting to improve their processes, the authors set out to develop a methodology. They stated their purposes as follows:

> We wanted to judge not only the quantity of individual items, but also the overall qualities of the tests—range, balance, and degree of challenge—represented by the set of test items as a whole. Furthermore, we sought a method that recognized that alignment is not an attribute of either standards or assessments, per se, but rather of the relationship between them. And because it describes the match between standards and assessments, alignment can legitimately be improved by altering either one of them or both.[8]

The authors established four dimensions for reviewing alignment:

- Content Centrality "provides a deeper analysis of the match between the content of each test question and the content of the related standard by examining the degree or quality of the match."
- Performance Centrality "focuses on the degree of the match between the type of performance (cognitive demand) presented by each test item and the type of performance described by the related standard."
- Challenge is a criterion that is "applied to a set of items to determine whether doing well on these items requires students to master challenging subject matter."
- Balance and Range addresses whether the tests cover "the full range of standards with an appropriate balance and emphasis across the standards."

---

6 American Federation of Teachers, *Making Standards Matter 2001*.

7 Richard W. Cross, Theodore Rebarber, Justin Torres, and Chester E. Finn, Jr. (editors), *Grading the Systems, the Guide to State Standards, Tests, and Accountability Policies,* Thomas B. Fordham Foundation, January 2004. States were graded on six components, including the alignment between test and content standards. Often, the basis for rating a state did not exist. To understand these ratings, readers are advised to look at the specific criteria used in the evaluations.

8 Robert Rothman, Jean B. Slattery, Jennifer L. Vranek and Lauren B. Resnick, *Benchmarking and Alignment of Standards and Testing*, CSE Technical Report 566, Center for the Study of Evaluation, Graduate School of Education and Information Studies, University of California, Los Angeles, May 2000.

When Robert Rothman and his colleagues applied these criteria to tests and standards in five states, they concluded the following:

> They have, for the most part, limited their tests to material that is in the standards—a primary requirement for a fair accountability test . . . Further—at least after our correction of the test blueprint—included test items are generally quite well aligned to the standards or objective to which they are mapped.
>
> But the good news ends here. With few exceptions, the collections of items that make up the tests we examined do not do a good job of assessing the full range of standards and objectives that states have laid out for their students. What is included and excluded is systematic: the most challenging standards and objectives are the ones that are under-sampled or omitted entirely . . . Thus, many of the tests in use by a state cannot be judged to be aligned to the state's standards—even though most of the items map to some standard or objective.

Achieve, Inc. subsequently worked with individual states to improve alignment between tests and content standards.

The Council of Chief State School Officers has coordinated another effort to assist states in measuring the degree of alignment between tests and standards, and in developing tests that reflect what the content standards require. The results were reported in a study by Norman L. Webb.[9]

The first step was to develop the criteria for determining alignment between the test and the content standards, and to train reviewers in the process of determining alignment. Four criterion were established:

a. Categorical Concurrence. This criterion "provides a very general indication of whether both documents incorporate the same content. The criterion of categorical concurrence between standards and assessment is met if the same or consistent categories of content appear in both documents. This criterion was judged by determining whether the assessment included items measuring content from each standard."

b. Depth-of-Knowledge Consistency. "Depth-of-knowledge consistency between standards and assessments indicates alignment if what is elicited from students on the assessment is as demanding cognitively as what students are expected to know and do as stated in the standards." Four depth-of-knowledge levels were defined for each of the four content areas (reading, writing, math, and social studies), and elaborated on in considerable detail:

   - Level 1. Recall or simple reproduction of information.
   - Level 2. Use of skills and concepts.
   - Level 3. Strategic thinking.
   - Level 4. Extended thinking.

c. Range-of-Knowledge Correspondence. "The range-of-knowledge criterion is used to judge whether a comparable span of knowledge expected of students by a standard is the same as, or corresponds to, the span of knowledge that students need in order to correctly answer the assessment items/activities."

d. Balance of Representation. "The balance-of-representation criterion is used to indicate the degree to which one objective is given more emphasis on the assessment than another. An index is used to judge the distribution of assessment items."

A Source of Challenge Criterion "is used only to identify items on which the major cognitive demand is inadvertently placed and is other than the targeted construct (skill, concept or application)."

An example of a summary for a state in one grade and subject area is provided in Table 1, for language arts in grade 11 for state F.

In different grades, states, and subjects, the degree of alignment varied considerably. The states volunteering to participate in the study were undergoing reviews and changes in their programs, and the study gave them information they could use. The degree of alignment in these states has likely changed since the study.

[9] Norman L. Webb, *Alignment Study in Language Arts, Mathematics, Science, and Social Studies of State Standards and Assessments for Four States,* Council of Chief State School Officers, December 2002.

Table 1:

Is the Alignment Acceptable?
Language Arts in Grade 11, State F

Standards	Categorical Concurrence	Depth-of-Knowledge Consistency	Range-of-Knowledge Correspondence	Balance of Representation
1. Reading Process	Yes	Weak	Yes	Yes
2. Responding to Text	Yes	No	Yes	Weak
3. Information and Research	No	Insufficient Number	Insufficient Number	Insufficient Number
4. Grammar/Usage and Mechanics	Yes	Yes	Yes	Yes
5. Literature	No	No	Yes	Yes

Source: Norman L. Webb, *Alignment Study in Language Arts, Mathematics, Science, and Social Studies of State Standards and Assessments for Four States,* Council of Chief State School Officers, December 2002.

In summary, much work has been done to enable matching tests to content standards, and to determine whether the desired match has been achieved. Some advocates of high-stakes testing are inclined to view this work as an effort by experts to erect impediments to action and results. But if we think about the matter in common sense terms, it is clear that alignment is vitally important. Content standards represent what education policy makers want students to know and be able to do. The test is an instrument to see whether or not that goal is achieved. If the test is not "done right," then the question of whether the goal is reached remains unanswered—and the risk of undue negative consequences for teachers and schools is greatly increased.

Alignment of Instruction to Content Standards, and Tests to Instruction. If the accountability test is aligned with the content standards, and if the curriculum actually in use in the classroom is also faithful to the content standards, then the test can help tell if the standards are being mastered. In other words, when what is taught and what is tested are both connected to the content standards, the three pieces fit together. Where they don't, a number of things may happen.

- The test may be measuring achievement of the state content standards and students do poorly because that is not what they are taught;
- Curriculum and instruction may be addressing the state content standards, but since the test is not aligned to them, it does not measure achievement of the standards and it is hard to tell what the scores mean;
- The test may be measuring what is taught but not the achievement of the state content standards, because what is taught is not aligned to the content standards;
- What is taught does not align with the content standards, and neither does the test, so no one knows what the test is measuring.

These scenarios are not equally likely to happen. Furthermore, it takes a brave teacher to teach the content standards when the test doesn't match them, as in the second approach. More often, it appears that teachers shift their instruction to what is going to be tested, knowing that their (or their school's) effectiveness may be judged according to students' performance. This is discussed later in the report.

Based on the studies and evaluations this author has seen, states vary widely in their progress toward achieving the desired match between content standards and curriculum. In one study of reading, all but three states claimed there was alignment.[10] Yet the AFT 50-state study cited earlier found that fewer than one-third of the tests in use are supported by adequate curriculum.

Adjusting the actual curriculum and course content to the prescribed standards involves considerable effort, and often considerable investment, as well.

- Course descriptions and content must be communicated to curriculum development staff and to teachers in useable forms.
- Teachers must have opportunity through professional development efforts to get up to speed with curricula and instructional approaches that are more demanding or different from what they have been accustomed to.
- Schools and students must have textbooks, workbooks, and possibly software that are congruent with the content standards.

Clearly, then, numerous alignment requirements must be met if there is to be a chain of evidence that establishes how well students have mastered what is expected of them in the content standards. Some of these requirements have been examined in depth in certain places, and methods have been developed and applied in the field to facilitate this work.

One project that has the promise of providing tools to examine various aspects of alignment is the Survey of Enacted Curriculum Project, carried out by the Council of Chief State School Officers, the Wisconsin Center for Education Research, and an Eleven-State Collaborative.[11] (The term "enacted curriculum" is used to denote what is actually being taught in the classroom.) The described purpose of this detailed survey in eleven states is to provide:

> A practical research tool for collecting consistent data on mathematics and science teaching practices and on what is taught in classrooms. The enacted curriculum data give states, districts, or schools an objective method of analyzing current classroom practices in relation to content standards and the goals of systems reform.

The study dealt with several questions.

1. *How does math and science content taught in classes compare to the goals outlined in state and national standards?*
   - In middle grades math and science, most recommended standards are covered, but the level of expectation and depth of coverage vary widely among schools and classes.
   - Data reveal differences in the extent of teaching science content across the standards and the extent of articulation between grades.
   - Schools differ in the extent of emphasis on algebra, geometry, and data statistics at elementary and middle grades.
2. *What effect do state and national standards for science and math learning have on the curriculum taught in classrooms?*
   - State frameworks/standards and national standards are reported by most teachers as strong positive influences on their curriculum.
   - Survey data allow comparisons of degree of influence on curriculum of state and national standards, textbooks, state and district tests, and teacher preparation and knowledge.
3. *Do state assessments reflect what is being taught in classes?*
   - Analysis of teacher reports and state assessment items show that tests cover a more narrow range of expectations for students than instruction does, with tests focusing more on memorization, facts, and procedures and less on solving novel problems and applying skills and concepts. It is good that what is taught is broader; it is the tests that need to be changed.

[10] K.K. Wixon, et al., *The Alignment of State Standards and Assessments in Elementary Reading,* CIERA Report #3-024, University of Michigan School of Education, 2002.

[11] Council of Chief State School Officers, *Summary Report from the Enacted Curriculum Project*, May 2001.

- The data on alignment between teacher reports on instruction/content and state assessments allow teachers and assessment staff to examine the areas of weakness and strength of tests and classroom practices.

After analyzing the taught curriculum in light of the state test, two states reached the following conclusions:

- For mathematics, "the alignment statistic of 0.37 means that less than half of the . . . content topics . . . reported by teachers were in common with the assessment items found on the state mathematics test and the NAEP test."
- In science, "the alignment statistic of 0.33 means that less than half the . . . content topics . . . reported by teachers were in common with the assessment items found on the state science test."

Finally, the Enacted Curriculum Surveys "can provide a database for monitoring the degree to which classroom curriculum is moving toward the standards. Standards are written with specific benchmarks or indicators of student performance, and the survey data are reported both by broad categories matched to standards and by specific item profiles and teacher expectations that match the benchmarks."

If curriculum and instruction are properly aligned with what is being expected of students, and if what is expected of students matches what is being tested, then the test results will be valid for the purposes for which they are being used. Unfortunately, this is frequently not the case. In many places, test scores and changes in test scores from year to year continue to be used for accountability regardless of whether or not alignment exists, raising the possibility of serious consequences for students, teachers, and schools. Shouldn't accountability systems have higher standards? Shouldn't the test scores be applied, and consequences meted out, only after the required alignments have been obtained? Fortunately, effective approaches have been developed to determine whether alignment has been achieved—and if not, how to accomplish it.

While content standards establish what students should know and be able to do, performance standards indicate how much of the content the students have mastered. The rubber hits the road in test-based accountability at the setting of the performance standard: the "cutpoint" on a scale where a test score reaches or exceeds some level labeled, for example, "proficient." The performance standard is accompanied by a written description of what students at that point on the scale know and are able to do. Just as the curriculum and test being used must be aligned with the content standards, so too must the performance standards be aligned with the content standards and test.

Alignment of Performance Standards. In a handbook prepared for the U.S. Department of Education and the Council of Chief State School Officers to guide establishment of standards-based reform systems, this alignment is described as critical:[12]

> 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 tested and what is tested is taught. No surprises, no questions, no controversy, and no confusion. *Although the primary use of a system of performance standards may appear to be its connection with the tests or assessment as results are reported, the system remains rooted in content*. (Italics added for emphasis.)

As the Handbook put it: "The system is based on specific content, and components are interpreted relative to content standards."

The question, then, is by what method do we move from content standards to performance standards and test score cutpoints so as to know whether the student has learned the content and established some required degree of mastery?" This author's search for the answer to that question was stimulated by a statement made by Bert Green, an eminent scholar in the field of educational assessment:

> The performance standards have to reflect the content standards. Logically, it would seem preferable for the judges to set standards just on the content domain. They could identify what parts of the domain are basic, what parts go with proficient persons, and what parts would be mastered by advanced students. **It is not at all clear how to do this,** (emphasis supplied), but a way might be found. Judges might also be useful in evaluating this bridge from content to performance. This would seem a more straightforward test than imaging the test behavior of marginally competent test takers.[13]

The known methods for setting cutpoints on tests were developed before Bert Green made this statement, and to the author's knowledge, there is still no method for setting performance standards in this manner. What I looked for in the *Handbook*, and other places, was something like this: "Starting with content standards, follow these steps to determine how much of the content standards are mastered at a particular score on a test that is aligned to the content standards." What I found instead were descriptions of traditional methods of setting cutpoints on standardized tests. These standard-setting processes, in and of themselves, do not tell us whether the breadth of the content standards are being achieved or what proportion are being achieved.

Methods for Setting Standards. One method that is used to set cutpoints and define performance standards entails analyzing the test questions that a borderline student—for example, a student bordering on a level designated as proficient—would be able to answer. Based on this analysis, a description of what it means to be proficient is developed, consisting of a paragraph or so. (This contrasts sharply with content standards, which usually require a fairly thick document for just one subject at one grade level.) The following example is a description of the "proficient" level on the National Assessment of Educational Progress (NAEP) mathematics assessment in grade 12.

[12] Hansche, 1998.

[13] Bert Green, *Setting Performance Standards: Content, Goals and Individual Differences,* W. Angoff Memorial Lecture Series, Policy Information Center, Educational Testing Service, 1996.

NAEP *Proficient* Level For Grade 12 Mathematics
(Scale Score at 336)

> Twelfth graders performing at the *Proficient* level should demonstrate an understanding of algebraic, statistical, and geometric and spatial reasoning. They should be able to perform algebraic operations involving polynomials; justify geometric relationships; and judge and defend the reasonableness of answers as applied to read-world situations. These students should be able to analyze and interpret data in tabular and graphical form; understand and use elements of the function concept in symbolic, graphical, and tabular form; and make conjectures, defend ideas, and give supporting examples.

Another approach, the "bookmarking method," arrays all the test items on the order of their statistical difficulty. Members of a panel of qualified people go up the list until reaching the point where they think a described level of performance is reached. It is hard to see how such an approach would assure capturing the breadth of content that would represent a reasonable mastery of what is in the content standards.

Such methods have been developed for establishing cutpoints on tests where it is very necessary to establish such cutpoints, such as in licensure. These methods have depended heavily on professional knowledge of what a person must know to be able to perform in an occupation. For example, a panel of cosmetologists may be assembled to make judgments about what test items the tests taker would need to get correct in order to be competent in the occupation. A reasonable degree of similarity in the ratings of the panelists is required for the standard-setting process to be valid.

In the early 1990s, I asked William Angoff, the late measurement expert at ETS and creator of the "Angoff method" that is frequently used to set cutpoints, whether it would be possible to use his method to set literacy scores required for a particular occupation. He answered yes, provided that there was a considerable degree of agreement among panel members drawn from that occupation. Of course, in setting school achievement targets, such panels are expected to come from different backgrounds, to be "representative;" they have different perspectives on what students should know and be able to do to be proficient. And there is no tangible performance requirement, as there is for a functioning cosmetologist.[14]

The methods available can do the job of setting cutpoints where it is really necessary to make pass and fail decisions. The methods prescribe rigorous processes, but they still rely on judgment, and there is no consensus as to which method is best. In 1994, A Joint Conference on Standard Setting for Large-Scale Assessments was held with 18 presentations by scholars. The results were summarized this way:

> Even though controversies and disagreements abounded at the conference, there were some areas of general agreement. Authors agreed that setting standards was a difficult, judgmental task and that procedures used were likely to disagree with one another. There was clear agreement that the judges employed in the process must be well trained and knowledgeable, represent diverse perspectives, and that their work should be well documented.[15]

The difficulty of using existing methods to determine the test score required to establish "proficiency," or some other level of achievement with recognized legitimacy, has been well demonstrated by the decade-long effort of the National Assessment Governing Board (NAGB) to do so with scores on NAEP. NAEP first set performance levels in 1990, generating considerable and prolonged debate. The Joint Conference referred to earlier was an effort to settle on a valid and acceptable way to establish these performance standards. But the debate continued. While some measurement experts defended the achievement levels, evaluations by The National Academy of Science (NAS) and The National Academy of Education (NAE), pronounced them "flawed."[16] In 1994, Congress required

[14] Dylan Wiliam, in commenting on this section, pointed out that "most standard-setting methods emphasize reliability over validity, and . . . the notion of validity even in methods such as the Angoff method, is very weak, being primarily about content or face-value considerations, rather than (say) maximizing accuracy of predictions. Ultimately, all standard-setting methods are seeking to draw an arbitrary line along a continuous variable." – Personal Communication

[15] L. Crocker and M. Zieky, *Proceedings of the Joint Conference on Standard Setting for Large-Scale Assessments*, Washington, DC, National Assessment Governing Board and National Center for Education Statistics, p. ES13. Crocker and Zieky are quoted in Robert L. Linn, "Performance Standards: Utility for Different Uses of Assessments," *Education Policy Analysis Archives*, September 2003, p. 8.

[16] A reading of the NAS and NAE reports would be instructive in setting cutpoints, in terms of the problems to be dealt with.

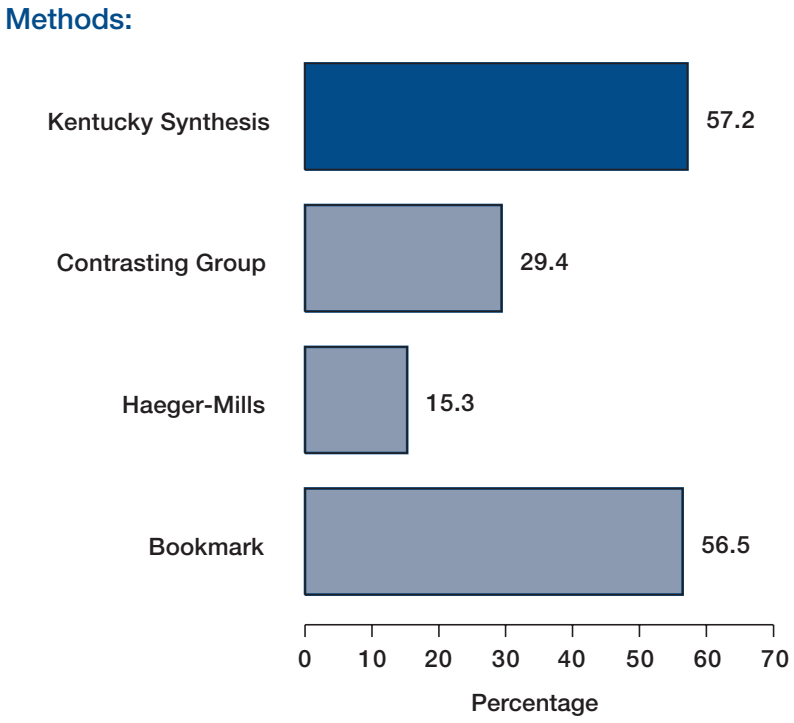
that achievement levels be used on a "developmental" basis until the Commissioner of Education Statistics determines that the achievement levels are "reasonable, valid, and informative to the public."[17] This determination has not yet been made. Each NAEP report summarizes the process and the evaluations made, and notes the Congressional requirement that the levels be labeled "developmental." These standards are still a work in progress.

The research on variations in "passing" scores on state assessments is summarized by Robert L. Linn in his article on performance standards. In one study, four different methods were used. As one example, for grade eight, these four methods set passing scores on a 60-item reading test at 28, 39, 43 and 48 items correct.[18]

These differences have led some to suggest that several standards be used to inform choices that are made about standards. Obviously, to do so would be expensive and time consuming. Kentucky has taken this approach, however. Three different methods were used, the results synthesized, and a standard chosen for six subjects at the elementary, middle school, and high school levels. Figure 1 illustrates the results for reading at the middle school level.[19] While in this case the chosen standard (the Kentucky Synthesis) was the highest of what was produced by the three methods, it varied considerably in relation to the three in other grades and subjects.

Figure 1:

Percentage of Students at or above the Proficient Level, Based on Different Standard-Setting Procedures



Source: Adapted from tables in Green, Timble, and Lewis (2003).

[17] *The Nation's Report Card: Mathematics 2000*, National Center for Education Statistics, NCES 2001-517, August 2001, p. 13.

[18] Linn, 2003, referring to a study conducted by J.P. Poggio, D.R. Glasnapp, and D.S. Eros reported in *An Empirical Investigation of the Angoff, Ebel, and Nedelsky Standard Setting Methods*, a paper presented at the annual meeting of the American Educational Research Association, April 1981.

[19] Reported in Linn, 2003, drawing on the work of D.R. Green, C. Trimble, and D.M. Lewis, "Interpreting the Results of Three Different Standard-Setting Procedures," *Educational Measurement: Issues and Practices*, 22, 1, 2003, pp. 22-32.

A Single Cutpoint to Measure Progress? It is understandable that psychometricians and educational measurement experts are often reluctant to use these cutpoint methodologies except in situations where there is a clear need to pass or fail an individual on a test. Robert Linn says he believes "it would be desirable to shift away from the standards-based reporting for uses where performance standards are not an essential part of the test use."[20] Linn also notes Bert Green's concern that in such systems a single item provides little information, but a single item may make the difference between meeting or not meeting the required score.

Indeed, whatever the validity of a particular cutpoint arrived at by these methods, there are better approaches to using test scores in accountability systems. There are better ways to know whether or not student performance is improving. While there is a certain logic in determining whether students are reaching an established standard, there is no transparency in the way this is done that connects the score to the content standards so as to reveal how much of the content students have mastered.[21] Not knowing this, the required alignment between the performance standards and the content standards has not been achieved, and year-to-year changes in the percent of students reaching these cutpoints do not tell us about progress toward the learning specified in those content standards. The word "proficient" is there beside a point on a scale, but proficient to what degree in terms of mastering the content described in the state standards?

Because the meaning of such cutpoints is problematic, in terms of the alignment objectives of standards-based reform, the use of a cutpoint—however attained—as the sole measure of what a test reveals in an accountability system is a poor choice. It is just one point along an achievement scale, and in terms of change from year to year it reveals only the performance of the small proportion of students clustered around that level of proficiency. It can appear that performance for schools as a whole has improved when in actuality, just a handful of students may have improved—and possibly by just a little bit. Similarly, performance overall may appear to deteriorate when the scores of just a handful of students decline.

One thing to worry about is that when the stakes are high, students just below the "proficient" level may be targeted for improvement. But the real concern is that, since all students take the test, we should have some way of knowing how all students have done when we hold schools accountable for improvement. Do we not care if the lower tier of students, those well below the chosen cutpoint, have improved or deteriorated? Do we not care if the higher achieving students, those above the cutpoint, have improved or deteriorated? This consideration is especially important in a country where there is such a wide distribution of achievement.

A limited view of change in achievement, such as just the percent proficient, can produce a very distorted picture of what is happening in schools. Take the case of NAEP assessment results in mathematics for Mississippi from 1992 to 1996.[22] No improvement was seen over that period in the percent of students reaching the level of "proficient," as defined by NAEP. However,

- The average score for all students improved;
- The average score for students in the bottom quartile improved;
- The average score for students the top quartile improved; and
- The gap between the top and bottom quartile was reduced.

So, was there no improvement in achievement in Mississippi in this time period? Apply the same scenario to an individual school in a high stakes accountability system. Are penalties in order?

In the fall of 1989, not long before cutpoint method was first used in NAEP to report scores, the Education Summit of President George H. Bush and the nation's governors was convened in Charlottesville, Virginia, to set goals for the nation for educational achievement. The Summit set the following objective

[20] Linn, 2003.
[21] We are here discussing accountability systems. Passing scores, of course, are often assigned to individual students on standardized tests.
[22] Paul E. Barton, *Raising Achievement and Reducing Gaps: Reporting Progress Toward Goals for Academic Achievement,* National Education Goals Panel, 2001.

to be achieved by 2000: "The academic performance of all students will increase significantly in every quartile, and the distribution of minority students in each quartile will more closely reflect the student population as a whole." The nation's governors and President Bush were right to recognize the wide range of student achievement in our country, and the need to improve it up and down the line, at the same time that gaps were being narrowed.

Options for Measuring Progress. There are many options for the use of broader measures, and perhaps a combination of measures. Change in the average score would draw on all student scores. And we could follow changes in the average scores of each subgroup of students. Changes in the average could be compared, at least roughly, across schools, districts, and states, using a standardized measure of change such as standard deviations. That is not now possible across states, since each state sets its own cutpoints, and they vary considerably.

Alternatively, a set of measures from testing could be used instead of just averages. This author has written three reports showing progress toward the goals as set by the 1989 Education Summit and enacted into law by Congress, two published by the National Education Goals Panel and one by ETS's Policy Information Center. Table 2, drawn from the last one, addresses science achievement for the eighth grade from 1996 to 2000.

In science at the eighth grade level, we see that while only three states raised their average scores and two had declines, seven states increased the percentage of students reaching the proficient level or above. Only one state showed improvement in the bottom quartile, while 17 states had improvement in the top quartile.[23] No state reduced the white-minority score gap, based on a comparison of average scores, and seven states deteriorated in the poor/non-poor gap, with no states improving. Other indicators could be added, such as the average scores for each subgroup, or an index could be established for each subgroup. From such a set of indicators, an index of improvement could be established.

In summary, the widespread practice of using a single measure of progress—the percent of students reaching a cutpoint on a test—conceals a lot more than it reveals about educational progress. Of course, when tests are used to judge the performance of

Table 2:

Changes Between 1996 and 2000 In NAEP Eighth Grade Science Achievement on Seven Indicators

Changes 1996 – 2000	States Improving	States Unchanged	States Worse
Average Score	3	28	2
Percent "Proficient"	7	26	0
Top Quartile Average	17	14	2
Bottom Quartile Average	1	25	7
Gap: Top and Bottom Quartile	0	22	11
Gap: White and Minority	0	33	0
Gap: Poor and Non-poor	0	26	7

Source: Policy Information Center, Educational Testing Service, "*A Deeper Look at NAEP Science Results*," *ETS Policy Notes*, Vol. 11, No. 1, Fall 2002.

[23] Florida is an example of a state that includes change in scores in the bottom quartile in its accountability system.

individual students, the use of passing scores may be more understandable. To judge whether teachers and schools are effective, however, an evolving accountability system needs to develop a broader base of student achievement measures. We should care about the performance and progress of students *below*, as well as above, the cutpoint. The single measure now often used also would seem to be destined to fail to meet the standard set by standards-based reform systems: that the measure used to gauge performance be clearly aligned to the content standards established by or within the states, at least until a specific methodology comes along.

The kinds of improvements described in this section would provide a more complete and accurate view of student achievement based on test scores; however, neither a single test, nor standardized test scores alone, should be the basis for high-stakes decisions about students, teachers, or schools. It is important to remember, however, that how much students know at the end of the eighth grade, for example, is the result of the experiences and life conditions both before school and after school hours, what was learned in the first seven years of school, as well as the quality of instruction during the eighth grade. In a new report being issued by the ETS Policy Information Center, University of Michigan Professor Stephen Raudenbush analyzes both the use of tests in the now almost universal practice of comparing the level of student achievement over time, and the measurement of growth or gain during the school year. He concludes that both approaches can provide useful information, but that such test results cannot stand alone in handing out rewards and punishments, and that other information must also be brought to bear.[24] We argue in the next section that schools and teachers should be held accountable for what students learned *during* the eighth grade—no more, no less.

[24] Stephen W. Raudenbush, *Schooling, Statistics, and Poverty: Can We Measure School Improvement?,* The Ninth Annual William H. Angoff Memorial Lecture, Policy Information Center, Educational Testing Service, 2004.

While the previous sections have addressed the level of student achievement, this section is about progress, or gain, in achievement. The general question addressed below is whether we want to reward or sanction schools on the basis of the level of performance achieved by their students, or on the basis of their progress.

As the nation entered the 1990s and the standards-based reform approach was gathering steam, the existing K-12 testing system was mostly designed to rank students by percentiles and sort students into tracks or ability groups in classes. Students and schools would know how well they compared with other students and schools; this is the norm-referenced approach. Coming along beside the ranking and sorting was the criterion-referenced approach, where students were measured against some set level of attainment, an approach much more adaptable to use in the new reform movement.[25]

For measuring change over time in teacher and school effectiveness, the norm-referenced approach suffers from not measuring students against some fixed standard of achievement.[26] Students and schools could move up in the rankings, or down in the rankings, depending on how students and schools elsewhere performed on the tests. While the trend is generally away from such tests, many states and districts still use them. It is hard to see how strictly norm-referenced tests can appropriately be used in a high stakes accountability system. On the other hand, they may be useful in enabling a state to compare achievement of its students to those in other states.

Grafting onto Old Testing Systems. The problems with grafting accountability onto the existing system go far beyond the use of norm-referenced tests. Existing standardized tests can be used to gauge performance in eighth grade mathematics, for example, at a point in time. However, such tests may capture not only what the students learned during the eighth grade, but also prior learning—what the student learned in the prior grades, in formal early childhood programs, and at home. This presents a very serious problem for any accountability system that is used to measure whether a teacher or school is doing a more effective teaching job with this year's eighth graders, compared with last year's eighth graders or eighth graders of five years ago.

The obvious challenge is how to compare the effectiveness of teaching this year's eighth graders as compared with those in prior years, when we don't know how different the education and development of these students was for the period before the eighth grade, and when the demographic makeup of the entering class of eighth graders may be changing over time. For example, if eighth graders entering school this year are less prepared and less knowledgeable than their predecessors two years earlier, then better teaching in the eighth grade math class this year could result in no better performance on the accountability test.

An analysis of data available from NAEP illustrates how such comparisons of achievement of 13-year-olds (mostly eighth graders) over time can represent incomplete information. NAEP regularly reports achievement results by comparing students at the same age or grade level—for example, today's eighth graders with those of four years ago or thirty years ago. Since this is done at ages 9, 13 and 17, and data are available for four year intervals, we can also look at growth in achievement from age 9 to age 13—for example, from age 9 in 1992 to age 13 in 1996. We will call this "cohort growth."[27]

From 1971 to 1996, reading scores were up at both age 9 and age 13, in terms of the level of achievement. But in terms of cohort growth in the period from age 9 to 13, over the same span of time, achievement was actually unchanged (see Table 3). Thus, achievement from age 9 to13 in the period from 1991 to 1996 was the same as it was from 1971 to 1975. The gains made

[25] The author has described these developments in two brief reports: *Too Much Testing of the Wrong Kind; Too Little of the Right Kind in K-12 Education, 1999,* and *Staying on Course in Education Reform*, 2002, both published by the ETS Policy Information Center.

[26] As used here, this approach might more accurately be called "cohort referencing." With a properly established norm group, examinees are always tested against students who have already been tested, so that if a student improves his or her performance, the score would go up (at least between re-normings).

[27] See Paul E. Barton and Richard J. Coley, *Growth in School: Achievement Gains from the Fourth to the Eighth Grade*, Policy Information Report, Policy Information Center, Educational Testing Service, 1998.

Table 3:

Trends in Cohort Growth Compared to Average Score Trends for 9- and 13-year-olds

	Cohort Growth, Age 9 to 13	Average Score Trend, Age 9	Average Score Trend, Age 13
Science	Level	Up	Up
Mathematics	Down	Up	Up
Reading	Level	Up	Up
Writing	Level	Level	Level

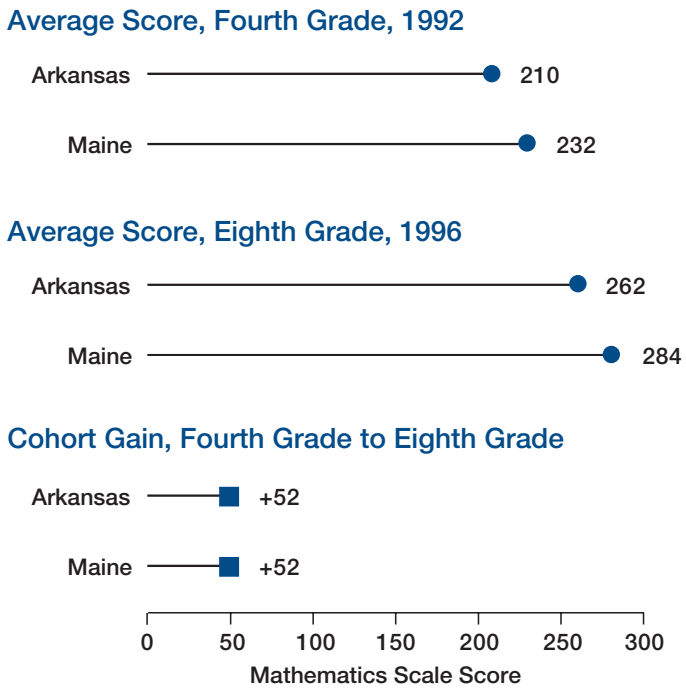
Source: National Assessment of Educational Progress data analyzed by the ETS Policy Information Center.

at age 13 over this period were the result of gains made before those students were age 9. The currently used accountability systems would have recorded these students as improving, and thus would have judged that the schools teaching them in that interval were improving.[28] The story is very similar for science and mathematics.

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 of those years, Maine had the highest average score in fourth grade mathematics in 1992 and also the highest average score in eighth grade mathematics in 1996 (see Figure 2). 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.

Figure 2:

Average NAEP Mathematics Scores and Cohort Growth, Arkansas and Maine



Source: National Assessment of Educational Progress data analyzed by the ETS Policy Information Center.

[28] The *Growth in School* report was an analysis by the Policy Information Center; NAEP is not reported by the Department of Education in terms of cohort growth. An update of this report describes some of the caveats involved in using NAEP data to measure trends in cohort growth scores. See Richard J. Coley, *Growth in School Revisited: Achievement Gains from the Fourth to the Eighth Grade,* Policy Information Report, Policy Information Center, Educational Testing Service, November 2003 (www.ets.org/research/pic).

Designs Appropriate for Accountability. Among the state testing systems, few are specifically designed to measure learning gains from the beginning to the end of a school year. For more than a decade, however, Tennessee has been using a system pioneered by William Sanders that shows average scores for school systems, schools, and individual classes, as well as gain in scores (value added) during the year of instruction. Each county is given a grade for the average score and a separate grade for the score gain in each of six subject areas. The contrasting results can be seen for Bradley County for 2003: the county's grades for average scores were one A, three Bs, and two Cs, while their grades for score gains were two Cs, two Ds, and one F. In some other counties the situation was reversed, with the average grades lower than the gain scores. Some high-performing schools did poorly in terms of the gains being made, and in others, it was the other way around.[29]

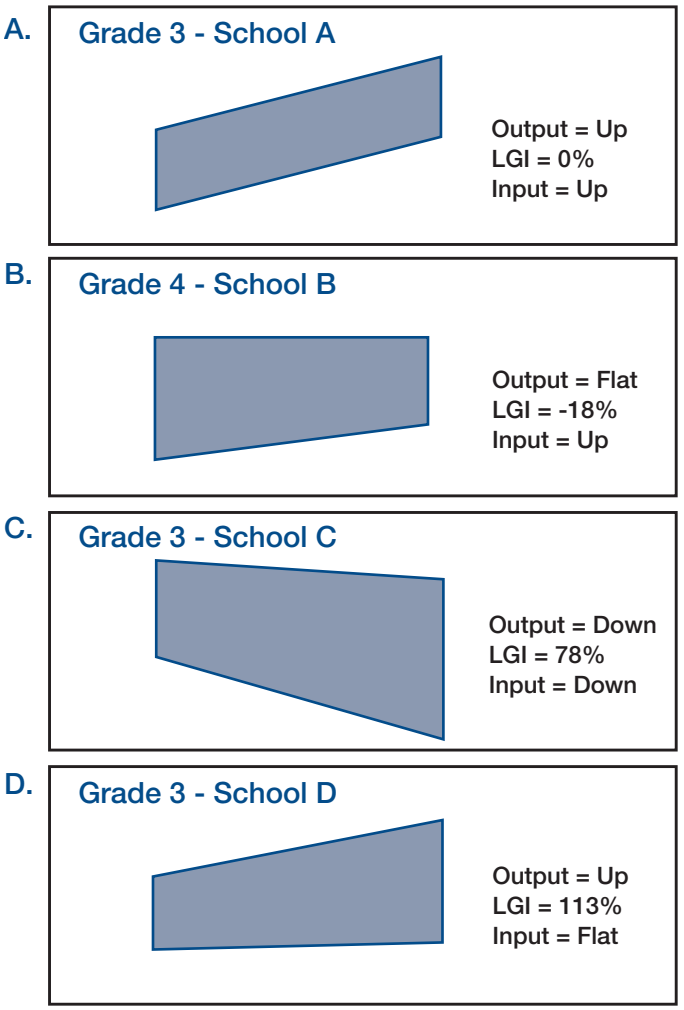
The Chicago Consortium on School Research has been using a "Learning Gain Index" to produce a school productivity profile for about a decade. The Consortium describes the process as follows:

> The productivity profile is built up out of two basic pieces of information for each school grade: the **input status** for the grade and the **learning gain** recorded for the grade. The input status captures the background knowledge and skills that students bring to their next grade of instruction. To estimate this input status, we began by identifying the group of students who received a full academic year of instruction in each grade in each school, and then retrieved their Iowa Test of Basic Skills (ITBS) scores from the previous spring . . . As for the learning gain for each grade, this is simply how much the end-of-year ITBS results have improved over the input status from this same group of students.[30]

Anthony Bryk says that this approach stems from the belief that "a school should be held responsible for the learning that occurs among students actually taught in the school." Figure 3 displays examples of grade productivity profiles using the Learning Gain Index (LGI). A school with its output up may have an LGI of 0 percent, because the input was up by an equal amount (School A). A school with its output down had a positive LGI, because its input had dropped more than the output (School C). Other combinations are also shown.

Figure 3:

Grade Productivity Profiles



Note: LGI = Learning Gain Index, computed for 1992-1996.
Source: Anthony S. Bryk et al., *Academic Productivity of Chicago Public Elementary Schools*, Consortium on Chicago School Research, March, 1998. Reproduced with permission.

[29] State of Tennessee School System Report Cards for 2003.

[30] A full description can be found in Anthony Bryk et al., *Academic Productivity of Chicago Public Elementary Schools*, Consortium on Chicago School Research, March 1998. It was necessary to build a gain score with a test in use in the schools, which was the Iowa Test of Basic Skills. One difficulty to be dealt with was that this is a norm-referenced test; a criterion-referenced test would have been preferred.

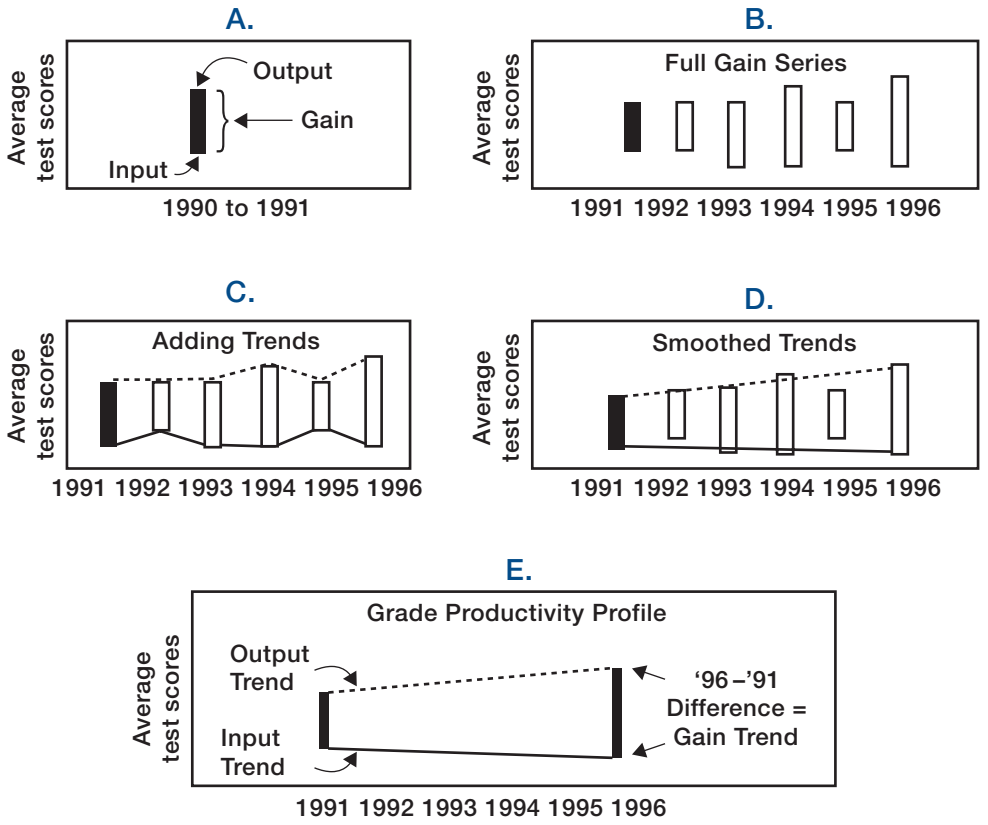
Figure 4, from the work of Anthony S. Bryk and Barbara Schneider, provides an example of how to depict trends in learning gains over time.[31] Block A shows "The Gain"—the difference between scores at the beginning and ending of the sixth grade in Prairie School in 1991. Block B shows the gain for each year from 1991 to 1996. Blocks C and D, which show fitting trend lines for the inputs and outputs over the period, reveal a positive gain from 1991 to 1996.

To determine whether schools or teachers are becoming more or less effective requires a measure of achievement gains *within* the school year. This kind of measurement has its own technical challenges to be worked out. In the present accountability systems, we are typically not measuring changes in the effectiveness of schools at any particular grade level over time.

Joseph Stevens describes the current practice of using successive cohorts of students to measure effectiveness where, for example, the mean fourth grade achievement test scores in a school for the year 2000 cohort of students would be compared to the mean fourth grade scores for that school for 1999. Stevens says "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.[32]

Figure 4:

Constructing a Sixth-Grade Productivity Profile for Prairie School, 1991 to 1996



Source: Anthony S. Bryk and Barbara Schneider, *Trust in Schools: A Core Resource for Improvement*, New York, Russell Sage Foundation, 2002. Reproduced with permission.

[31] Anthony S. Bryk and Barbara Schneider, *Trust in Schools: A Core Resource for Improvement*, Russell Sage Foundation, 2003, p. 102.
[32] Joseph Stevens, Susan Estrada, and Jay Parker, *Measurement Issues in the Design of State Accountability Systems,* a paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA, April 2000, p. 14.

Others, such as Goldstein and Raudenbush, as well as Linn and Haug, have pointed out the difference between the level of achievement and gain or growth in achievement for use in accountability. Goldstein, describing school effectiveness studies in Britain, writes:

> . . . it is now 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.[33]

Herbert J. Walberg, in enumerating principles for accountability designs, advises that "policymakers increasingly recognize that value-added scores better indicate the school's or teacher's contribution to achievement than do test scores at a single point in time . . . [N]on-value-added scores, however, can complement value-added scores, and together, they give policymakers more information and are less misleading than either one alone."[34] I suspect, however, that researchers are more likely to recognize this than are policymakers.

George R. Lowery of Roosevelt University and Katarzyma Kubzdela of the Chicago Board of Education, looking at evaluations of teacher effectiveness, state that "currently, the most accurate, accepted and utilized method for measuring teaching quality is *value-added* or *gain analysis*. This type of analysis has been independently developed and used by a number of researchers and institutions worldwide."[35] They have found that without this approach, schools serving high-performing students are unable to identify situations where students are not improving or may even be losing ground.

Sanctioning the Right Schools. Keith Zvoch and Joseph Stevens have recently summarized this research, and gone beyond it.[36] They point out that there have not been actual comparisons, using real school data, of the results of using measures of the achievement level *and* the achievement gain of a cohort of students over a period of time. They set out to do this in a large urban school district in the southwestern United States. The district has over 100 schools, serving a diverse student body of close to 90,000 students. Middle school students were used because they provided the only cohort for which three consecutive years of data were available; the test scores of students who did not attend all three of these years were excluded.

With regard to the question of the difference between the mean achievement of students in the same grade over time, and the growth in achievement of the same students, the authors concluded the following:

> The present study also showed that evaluations of school performance differ depending on whether school mean achievement or school mean growth in achievement are examined. There was significant variation in the mean achievement of students in mathematics and languages, both from student to student and from school to school. The analyses also showed that there was significant variation in the rate of achievement growth from student to student and from school to school for mathematics and from school to school for language . . . Evaluation of these estimates showed that the school mean level of performance was not strongly predictive of the school mean rate of growth. Correlations of school growth estimates were only 0.14 for mathematics and 0.41 for language . . . *characterization of school performance is substantially different depending on whether mean achievement or mean growth is examined* (emphasis added). In several cases, schools with low mean scores were not always "poor performing" schools. In fact, schools with low mean scores were in many cases the schools with the largest growth rate. Conversely, a high mean achievement score was not always a clear indicator of "good performance."

---

[33] H. Goldstein, "Better Ways to Compare Schools," *Journal of Educational Statistics,* 16(2), pp. 89-92.

[34] Herbert J. Walberg, "Principles for Accountability Designs," *School Accountability: An Assessment by the Koret Task Force on K-12 Education,* edited by Williamson M. Evers and Herbert J. Walberg, Hoover Institute, 2002m p. 161.

[35] George R. Lowery and Katarzyma Kubzdela, "Increasing Urban School Teaching Quality: A Value-Added Approach," *Improving Academic Achievement in Urban Districts,* Education Commission of the States, December 2003.

[36] Keith Zvoch and Joseph J. Stevens, "A Multilevel Longitudinal Analysis of Middle School Math and Language Achievement," *Education Policy Analysis Archives*, Volume 11, November 20, July 8, 2003.

Zvoch and Stevens point out that "accountability systems now in use in many states apply evaluative methods that cannot . . . validly disentangle school effects from factors that are outside the control of educational policy and practice at the school." In short, this study demonstrates that if achievement gains were not measured for the same students over a period of time, then the wrong schools would be sanctioned—perhaps even closed—in this city of over 100 schools. Schools capable of high gains in student achievement would be sanctioned, while those with low gains would be ignored.

A study by the Northwest Evaluation Association (NWEA) analyzed achievement growth based on tests administered to 230,000 students in 723 schools from 22 states in the spring of 2002 and the spring of 2003. The conclusion was that the addition of a growth indicator "adds essential information about school effectiveness." The study found that schools with similar status levels differed substantially in the amount of student achievement growth they caused. Also, they found that more than a fifth of the schools with high-scoring students were in the bottom fourth of the schools in terms of the amount of growth that occurred.[37]

A study by Carlson utilized data for a state that enabled him to compare results from a gain model ("quasi-longitudinal") with a model based on score changes for successive groups of students. He was able to compute correlations between estimates of school gain scores and estimates of school change scores for successive groups of students. He found the correlations between the two to be quite low; from .14 for one grade to .48 for four grades. Carlson comments: "The two longitudinal models have the advantage that they provide direct estimates of gains in achievement and do not depend on the comparisons of successive groups of students." [38]

The various accountability models in use in the states, and variations in them, are described by Robert Linn in a 2004 book entitled *Redesigning Accountability Systems for Education.* He also discusses some advantages and disadvantages of each.[39]

Together these analyses make it clear that, in evaluating the effectiveness of programs and interventions, it is the effect of those programs or interventions that have to be measured, and other factors that contribute to outcomes must be factored out.[40] For an ongoing system of evaluating school and teacher effectiveness, administering two forms of a test—one at the beginning of the year and one at the end—is about as close as one can come to this, as a practical matter. Further, the resulting gain or growth scores must be interpreted. The goal could be to demonstrate continuous progress, or it could be to achieve a certain amount of growth during a year in a particular subject, in a particular grade. Standards for growth can be set, just as standards are currently set for the level of achievement.

In a standard program evaluation approach, one form of the test would be administered at the beginning of the year, and one at the end. However, there are instances where the test results at the end of one year are compared to the test results at the end of the prior year. This is attractive in terms of reducing the test burden, but it leaves the problem of the differential experiences students have over the summer. Substantial research has found that in reading, for example, the different experiences of students over the summer can make big differences in achievement. For example, in one study, the achievement gains made during the school year were similar among students in both high-poverty and low-poverty schools. However, spring to fall comparisons over the summer vacation showed a quite different pattern, with poor students losing ground: "The differential progress made during the four summers between second and sixth grade

---

[37] Martha S. McCall, G. Gage Kingsbury and Allan Olson, *Individual Growth and School Success,* Northwest Evaluation Association, April 2004, page 1 of the Research Brief.

[38] D. Carlson, "All Students, or the Ones We Taught?," a paper presented at the Annual Conference on Large Scale Assessment, Council of Chief State School Officers, Snowbird, VT, June 2002, as cited in Robert L. Linn, "Accountability Models," in Susan H. Fuhrman and Richard F. Elmore (editors), *Redesigning Accountability Systems for Education,* Teachers College Press, 2004, pp. 89-90.

[39] Linn, 2004, pp. 73-95.

[40] For taking all the relevant factors into account, see Celia Rouse, *Accounting for Schools: Econometric Issues in Measuring School Quality*, a paper delivered at the ETS Invitational Conference, October 3-4, New York, New York, 2003.

accounts for upwards of 80 percent of the achievement differences between economically advantaged and ghetto schools."[41] There is a case to be made for assessment that reveals what happens over the summer as well as changes during the school year. And there is also a case to be made for programs that provide reading opportunities in the summer.

Comparing the end of the seventh grade in mathematics with the end of the eighth grade requires a developmental (or vertical) scale of achievement that also allows comparisons of growth over the year. This, too, has its measurement challenges. The testing and scaling have to be developed with this purpose in mind. Typically, present test results in grade 7 tell how students did on the subject matter for that grade, and likewise with grade 8. Such a vertical scale has to bridge from one year to another. Another approach, well known and well used, is using two forms of the same test, one given in the fall and one in the spring.

Different approaches to measuring gains have their advantages and disadvantages and there are some real technical issues to be dealt with.[42] But if standardized tests are to be used for consequences, the tests need to measure the learning that occurs within a school year. While this "value-added" approach is a big departure from prevalent practice in accountability systems, something like this has been carried out on a large scale for more than a decade in Tennessee, and which has recently changed significantly. Theirs is only one of many possible approaches, however.[43]

We need to know much more than we currently do about the level of achievement at any one grade and how that changes over time. We need to know how well we are doing, as a community and a society, in dealing with all the factors affecting school achievement. Accordingly, we need measures not only of changes in score levels over time, but also of growth or gain in scores during the school year.

[41] Donald P. Hayes and Judith Grether, "The School Year and Vacations: When Do Students Learn," *Cornell Journal of Social Relations,* Vol. 17, 1983, p. 64, as quoted by Richard L. Allington and Ann McGill-Frazen, "The Impact of Summer Setback on the Reading Achievement Gap," *Phi Delta Kappan*, September 2003, p. 69.

[42] For some issues involved, see Daniel F. McCaffrey et al., *Evolving Value-Added Models for Teacher Accountability*, Rand Corporation, 2003.

[43] At this writing, Tennessee's approach was being re-examined, and legislation opposing the testing system had been introduced in the state legislature. See Lynn Olson, "Tennessee Reconsiders Value-Added Assessment System," *Education Week,* March 3, 2004. For a recent view of the value-added system in Tennessee, as well as a discussion of a number of issues involved in value-added testing, see Henry Braun, "Value-Added Modeling," forthcoming in the ETS Teacher Quality series.

High-stakes tests for accountability impact teaching and the shape of instruction in two ways. One is what happens to the teaching of the subjects being tested. The other is what happens to the teaching of the subjects *not* being tested. The responses of teachers to the testing can vary across classes within a school, and across schools within a district.

Teaching in the Subjects Tested. A great many of the discussions and writings about contemporary education and the use of testing have something to say about "teaching to the test." In fact, a search on Google turns up about 11,000 entries for this phrase. The question of whether undesirable distortions in instruction are taking place, or whether teaching is being better focused or improved, is of primary importance. However, the framing of the discussion in terms of the pros and cons of "teaching to the test" seems to have become almost useless in understanding differences in viewpoints and in describing what is happening in classrooms.[44]

Shades of Gray. Experts versed in psychometrics and educational measurement would likely say something along the lines of the following about standardized testing. Tests are a collection of questions and tasks that represent a sampling of a "domain" of knowledge, or a subject area such as eighth grade math. Instruction should address that domain, and the tests should provide an estimate of how much the student has mastered the domain of knowledge. In fact, of 16 textbooks on testing, all published since 1990, 11 did not have the phrase "teaching to the test" in the index. Five of them did, and contained discussions ranging from relatively brief to substantial.

But in the classroom, things are not that simple. It is not just a matter of standard testing theory. Principals and teachers have their backs to the blackboard. They have to show their results in terms of advancing test scores, and if they do not, there are serious consequences. Education officials rely on the integrity of the scores as a gauge of whether standards-based reform policies are having the desired effect, and that integrity is assumed unless shown to be otherwise. Whenever a number is a stand-in for judging the success of an enterprise, ways will be found to get that number. The quarterly bottom line on corporate balance sheets, for example, has become the primary measure of executive performance. The frenzy to get good numbers distorts discussions as short-term goals supersede long-term ones. And accounting scandals reveal the extent to which good numbers are obtained by artful manipulation.

The debate over distorted corporate decision making and doubtful accounting practices reveals a broad gray area that extends from the barely legal to outright fraudulent. Similarly, in test-based accountability systems, comparable effort to get good numbers—numbers that represent the preparation of students who are proficient—ranges from poorly designed systems and policies to debatable practices to some outright cheating.

Ranking Practices. For a more simplified look at these gradations, we will start with the assumption that the elements of the system are aligned: the curriculum and the test are aligned with the content standards. Several authors have attempted to classify the kinds of practices teachers may or do engage in, in terms of ethical or unethical behavior, or good or bad educational practices. The most comprehensive attempt is by the Center for Education Policy, which classified them as follows:[45]

Bad Practices

- Getting actual test questions from a current test form and teaching students the answers;
- Giving students actual test questions for drill, review, or homework; or
- Copying, distributing, or keeping past versions of a test that have not been officially released as a practice exam.

[44] For an excellent summary of research performed over the years, see Joan Herman, "The Effects of Testing on Instruction," *Redesigning Accountability Systems for Education,* edited by Susan H. Furman and Richard F. Elmore, Teachers College Press, 2004. Also, a study in New Jersey describes how teachers responded to math and science standards and assessments; the study also addressed test preparation practices. See William A. Firestone et al., *The Ambiguity of Teaching to the Test,* Lawrence Erlbaum Associates, 2004.

[45] *Testtalk*, Center for Education Policy, June 2002.

Middling Practices—Use With Care

- Teaching students how to fill in the bubbles on answer sheets, narrow answers in a multiple-choice question, or write a short answer response or paragraph;
- Assigning homework and practice questions that resemble real test items;
- Teaching from state-endorsed or commercially-developed practice materials designed to go along with a particular test; or
- Giving writing assignments in the same format as the writing portions of a specific test.

Good Practices

- Covering the most important knowledge, skills, and concepts contained in the standards for that particular subject;
- Addressing standards for both basic and higher order skills;
- Using test data to diagnose areas where students are weak, and focusing instruction in those areas; and
- Giving students diverse opportunities to apply and connect what they are learning and demonstrate true mastery of standards.

Blaine Worthen and colleagues, in their 1999 textbook, attempt to rank a set of practices in a decreasing order of legitimacy as follows:[46]

1. Using specific instructional objectives to guide your teaching, without knowing what objectives are covered by the particular standardized test items in your district;
2. Motivating students to do their best on tests and teaching general test-taking skills;
3. Structuring the curriculum so that it corresponds to the objectives included in the standardized tests used in your district;
4. Teaching the specific format and objectives used in the test as a major part of the instructional activities;
5. Teaching the *specific* content of an upcoming test to future examinees, but without using the actual test items;
6. Under the guise of instruction, using one parallel form of a test for students to "practice," prior to administering another parallel form of the test to students;
7. Having students "practice," using the same form of the standardized test, or providing copies of actual test questions to examinees in advance, whether in instructional materials or any other form.

The authors comment: "Few would disagree with the first and second items in this list, but the third and fourth items may generate some disagreement. Probably most measurement professionals will agree that the last three items constitute inappropriate teaching to the test. Surprisingly, however, such activities are not uncommon." They cite a study finding that half the teachers interviewed did not consider it cheating if they had students practice on previous versions of the test currently in use by the district. Almost a quarter thought it was acceptable for a teacher to teach a specific item if the teacher happens to remember that the question was on a previous test. The study was conducted in 1985, before such high stakes were attached to tests.

The reader can draw his or her own line as to where good instruction and test preparation shade into undesirable or unacceptable practices. There would, I believe, be considerable variation in where that line would be drawn. And it would vary depending on the place, and on whether the line drawer was the teacher, principal, test creator and administrator, policy official, or parent. There also would be, I believe, considerable variation within each group.

Take the first approach in the Worthen list. The teacher does not know what educational objectives drive the creation of the test. In a great many places, teachers would be expected to tune into the objectives that frame the questions, and it would be the intention of test makers and givers that instruction be influenced by those objectives. On the other hand, in places where there was confidence that the content

[46] Blaine R. Worthen, Karl R. White, Xitas Fan, and Richard R. Sudweeks, *Measurement and Assessment in Schools* (Second Edition), Longman, 1999, p. 46.

standards, the instructional program, and the test had all been brought into alignment, a teacher might feel it appropriate to take the first approach. That is true in very few places, however.

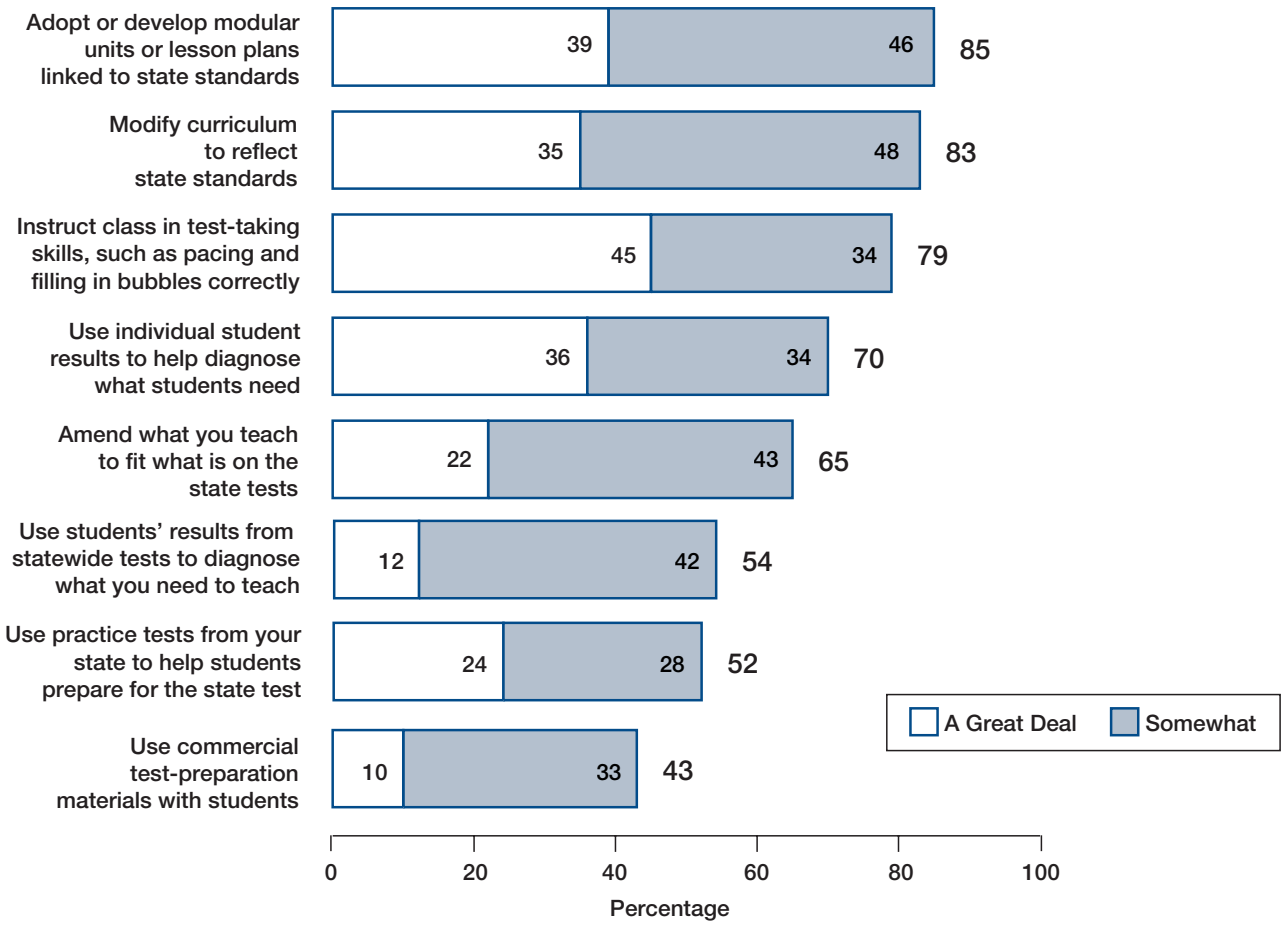
The fifth approach—teaching the content known or suspected of being tested (but without using the test items)—gets very close to teaching to the objectives of the test, as does the third approach, aligning the curriculum to the known test objectives. Some fine gradations of differences are involved here. Of course, if the district has not aligned the curriculum objectives to what is tested, that is the fault of the system, not the teacher.

Even the seventh approach, at the bottom of the list—having students "practice" with old forms of the test—is not so clear cut, according to *Testtalk*.[47] "Most state and national testing programs do officially release prior test versions for practice, and it's all right for teachers to use them in this way." Even in the case of the SAT, the College Board sells a book containing prior tests, to be used for preparing to take the test. Of course, using forms of non-released state tests is a problem.

What Teachers Do. What do we know about the prevalence of different kinds of teacher practices in the face of high-stakes testing? *Education Week* surveyed these practices in 2000. The results are shown in Figure 5.

Figure 5:

Teacher Practices Related to High Stakes Testing



Source: *Education Week*, "National Survey of Public School Teachers," *Quality Counts*, 2001.

[47] *Testtalk, 2002.*

The survey questions lump together some good instructional uses of tests with test preparation practices. The first two questions go to the matter of aligning curriculum with state content standards, and would be widely recognized as a good practice (depending, of course, on what one thought of the state standards). Many address preparation for testing, and these practices are at different points along the scale of what was reported above on "legitimate" practices.

Another study examined approaches to test taking practices and preparation on the part of reading teachers and supervisors in Texas.[48] They were asked to rate their use of five approaches on a scale: 1 stood for Never, 2 for Sometimes, 3 for Often, and 4 for Always. Most said Often or Always (see Figure 6). All of these activities were most frequent in low-scoring schools. Such practices were used throughout the year, peaking in the months before the state test.

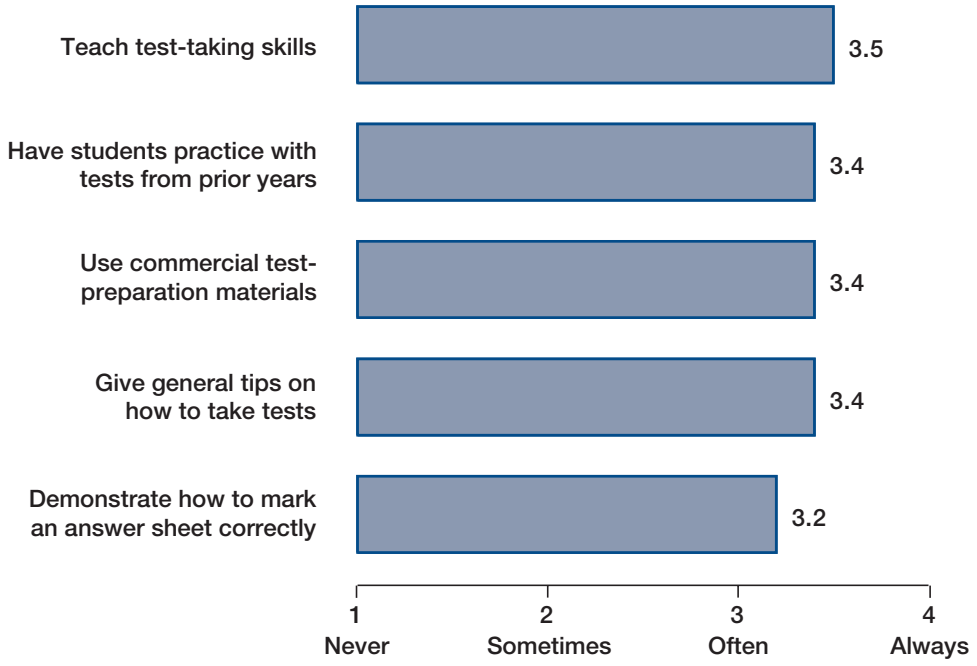
In a 1996 study in Maryland, about three-fourths of the principals in a survey said they encouraged teachers "a great deal in using released items from past tests and other materials for test preparation."[49]

Correct Approach? This report does not seek to delineate the correct or ideal teaching approaches to testing. There is clearly a lot of disagreement, and considerably more gray than black and white. Rather, it is useful to focus on *specific* practices and try to reach a better understanding of what is desirable and what is not.

Clarity on these matters within a state, district, or school is critical to the success of standards-based reform. To the extent that various practices make a difference in student scores, they make a difference in school comparisons, and to the extent the practices change, they make a difference in score trends. Further, there is a need to understand the teaching prac-

Figure 6:

Frequency of Texas Teachers' Approaches to Preparing Students for Tests



Source: Hoffman, Assaf, and Pario, 2001.

[48] James V. Hoffman, Lori Czop Assaf and Scott G. Pario, "High Stakes Testing in Reading: Today in Texas, Tomorrow?" *The Reading Teacher*, Vol. 54, No. 5, February 2001. Cited in *Testtalk*, 2002.

[49] Daniel Koretz et al., *Final Report: Perceived Effects of the Maryland School Performance Assessment Program,* National Center for Research, Evaluation, Standards, and Student Testing. Cited in *Testtalk,* 2002.

tices being used in order to determine whether higher test scores do, in fact, represent more advanced knowledge of the domain the tests are designed to sample—or whether they simply represent better knowledge of the test.

To some extent, the judgment about the degree to which results of tests, given under different teacher approaches, reflect knowledge of the domain may be deductive.[50] Or, empirical research may be needed to check up periodically on what is happening. For comparison, students who have been tested with a standard instrument and prepared for the test in different ways could be given another much longer test that is a much larger sampling of the domain—perhaps using matrix sampling, as NAEP does—so that one student does not have to be subjected to long periods of testing. The results would be compared with those of regular operational testing. This would be the kind of "evidence-based" approach advocated by the U.S. Department of Education. It would take resources. But the credibility of test-based accountability is at stake. It is essential to know how to use testing in ways that accurately depict progress toward achieving the subject matter required in the content standards.

Ways are also available to check results of tests, such as comparing results of a state criterion-referenced test to those of a nationally normed test, or looking at whether the students who met the standards made reasonable progress the next year.

Horns of the Dilemma. In the current situation, teachers and principals are often left in a twilight zone, if not in the dark. They must produce results, but the means of doing so may not have been made clear to them. To this point, as stated at the outset of this section, the assumption was made that the test, instruction, and content standards were at least reasonably aligned. More likely, all three are not. When they are not, the problems for the teacher and the principal are greatly enlarged. If the testing is for accountability only, a matrix approach such as that referred to above could be used operationally, assuring a broad coverage of the content standards.

If a district's prescribed curriculum, for example, is not in line with the content standards but the test is, the teacher faces a problem of covering the prescribed curriculum and also getting students ready for a test that is going to cover materials not regularly taught. So teachers must scramble and latch on to the approaches they can find to help their students obtain good test scores.

If the curriculum is aligned with the standards but the test is not (perhaps where an off-the-shelf test is used), the teacher has to do the best he or she can to find the differences and do what can be done to prepare the students for the test.

These kinds of misalignments are not rare; they are common. They put the achievement of the state content standards at risk. They also put teachers and principals at risk: they must deliver the desired test score distribution by various ad hoc means of stimulating acceptable student performance.

The Philadelphia school district announced in February 2004 that eleventh grade English and math teachers had to spend 20 minutes a day for the next 10 weeks using a Kaplan-provided test preparation program. It was designed to prepare students for the late March administration of the Pennsylvania System of School Assessment, a test based on the state standards. The city's high schools, it was explained, did not have a standard curriculum, and this new test preparation program would ensure that students learned the material that was likely to be covered on the test. Some teachers said it interfered with term papers and the study of literary works. This is a case where the system scrambles in a situation where instruction and the standards did not seem to be aligned.[51]

These misalignments are bound to result in distortions of instruction, and the use of valuable class time teaching to the machinery of test taking, and trying to obtain intelligence on what will be covered on the test. The dilemmas of teachers are created by failures of the system to fit the pieces together. Where there is a fit, the teacher can teach with a confidence that the test questions will be a fair sampling of the material the teacher is covering.

But a hurdle is often encountered in getting a "fit" between standards and instruction. The content standards can be very broad, so much so that teachers cannot cover the entire ground. *Education Week* asked

[50] This would not necessarily be breaking new ground; some such studies have been done.
[51] Susan Snyder, "Philadelphia Schools Adopting a Test-Preparation Program," *Philadelphia Inquirer,* February 8, 2004.

teachers whether they had enough time to cover the state standards; 24 percent "said they had 'far too little' and 46 percent 'somewhat too little' time."[52]

To sum up, more attention should be given to monitoring teacher practices, clarifying what teachers and principals are expected to do and not do, gauging the effects of different test preparation practices on scores and what they mean, completing the alignment of the key elements of the system, and getting content standards shoehorned within the day-to-day lesson plans.

The important point about introducing specified test preparation practices has to do with the impact they may or may not have on validity through compromising the generalizability of the test tests. There is a need to say: "If you permit X, then you cannot conclude Y from the test results."

Teaching in the Subjects Not Tested. A question frequently asked is whether instructional time shifts from non-tested subjects to tested subjects. The number of subjects tested varies among the states. All must now test in reading and mathematics, and later in science, under the No Child Left Behind Act. Statements are heard that there is no reason why other subjects should be slighted just because all are not tested. At the same time, some hold views that progress to proficiency is so critical in these basic subjects that there may be justification for redirecting instructional time.

Evidence of shifts in instructional time in the face of high-stakes testing is incomplete. Although there are scattered stories of redirection of instructional time and impact on the curriculum, and some surveys have been done, this author is not aware of statistical systems of regular reporting on the allocation of instructional time among subjects. If such systems were available, administrators and policy makers could monitor instructional time. Given the pressure being exerted on teachers, principals, and school systems, there is a need for such information to be available. Schools need to know that slighting of other subjects will be spotted. And if policy officials are, in fact, trying to get some redistribution, they need to know whether it is happening.

It is undoubtedly a constant struggle to fit everything into the school day and meet score targets for the subjects tested for accountability. The lead item in a 2004 news story read:

> Los Angeles School Superintendent Ray Romer found himself caught Thursday between two serious problems: the epidemic of childhood obesity and students' abysmal knowledge of science. To meet the need to raise science scores he proposed scrapping some health classes to create more time for science. But with 40 percent of the district's students obese, there was an outcry about reducing health instruction. At the meeting where the proposal was made, the president of the California Association of School Health Educators said, "Dead students do not do well in testing." The Superintendent is looking at alternatives.[53]

There is a similar concern in North Carolina regarding social studies. A past president of the North Carolina Council for Social Studies stated that with the testing in just a few subjects, ". . . social studies is left behind because there is no testing. . . We are putting it on the back burner." She said that a lot of groundwork laid in elementary school is no longer provided. "It's still part of the standard course of study, but a teacher has only so much time in the day."[54]

In Anne Arundel County, Maryland, the school administration reduced middle school offerings, including some courses required by the state. However, the Anne Arundel Coalition for Balanced Excellence in Education fought the reduction and won an appeal to the Maryland State Board of Education. The Board told the County to follow state requirements.[55]

In October 2003, *The Wall Street Journal* published an article titled "Schools Say 'Adieu' to Foreign Languages," based on a survey of cutbacks in foreign language offerings. The article states: "In general, school districts are making these cuts reluctantly. Some are having to shift resources to help students prepare for standardized tests in subjects like math and reading."

[52] *Education Week*, "National Survey of Public School Teachers, 2000," *Quality Counts 2001*.
[53] Jennifer Radcliffe, *Los Angeles Daily News*, January 30, 2004.
[54] Todd Silberman, *Raleigh News and Observer,* October 30, 2003.
[55] Anne Arundel Coalition for Balanced Excellence in Education, http://www.aacbee.org/home.htest, downloaded 12/29/2003.

In Kentucky, alignment of what is taught with state standards was driven not so much by the state curriculum framework, *Transformations*. Studies found that the document's lack of specificity and its length (500 pages) resulted in minimal use. But Kentucky had another document, *Core Content for Assessment,* that specified the content to be assessed. It was meant to be only part of the comprehensive curriculum described in the curriculum framework, but often was more useful in developing instructional plans. A fourth grade teacher was quoted in 1998 as saying, "Trying to get all the subjects is really difficult; so you tend to go with the ones you are really tested with and let other subjects slide." In statewide surveys, teachers reported deemphasizing untested areas in favor of tested ones.[56]

A comprehensive study, carried out in 1996 by Rand, was based on a sample of teachers and showed shifts in the curriculum. For example, 95 percent of teachers said that they increased writing instruction (which was tested by the state) and none said that they decreased it. Mathematics instruction was also increased considerably (45 percent said that emphasis was increased), and just 13 percent said that it was decreased. In social studies (also tested), 30 percent of the teachers said that they increased emphasis, while 33 percent decreased emphasis. In art (not tested), 34 percent of the teachers said that they decreased emphasis, while 10 percent increased it. The comparable percentages for music (not tested) were 21 percent and 5 percent, and for physical education (not tested), 22 percent and 3 percent.[57]

A comprehensive survey of teaching practices was conducted in 2001 by Boston College's National Board of Educational Testing and Public Policy. Lynne Olson summarized the results in *Education Week*. "Teachers are changing what and how they teach in response to state testing programs . . . Those changes are greatest in states where more consequences are attached to test results." One-fourth of the teachers in states with high stakes for students and schools reported cutting back on instruction "a great deal" in untested areas, compared with nine percent of teachers in states with moderate or low stakes.[58]

In March 2004 the Council on Basic Education released the results of an in-depth survey of instructional time conducted in four states: Illinois, Maryland, New Mexico, and New York. The key findings are signaled by the report's title: *Academic Atrophy: The Condition of the Liberal Arts in America's Public Schools*.[59] The study found increases in instructional time in subjects tested: reading, writing, mathematics, and science. About three-fourths of the principals surveyed reported that there were increases in instructional time in the first three of these subjects, as well as in professional development of teachers in these subjects. Close to half reported increases in science. In grades 6-12, there were "considerable increases in instructional time and professional development for social studies, civics, and geography."

Of great concern to the authors, however, were:

- Decreases in instructional time for the arts, especially in high-minority schools;
- Decreases in instructional time and teacher professional development in foreign languages; and
- Decreases in instructional time, in elementary (K-5) schools, for social studies, civics, and geography—with the decreases "especially evident" in high minority schools.

Reallocation of teaching time occurs within subjects as well as across them, one way or another, giving priority emphasis to what is the expected emphasis on the test. A forthcoming paper by Laura Hamilton at Rand provides a helpful review of several studies.[60]

---

[56] Patricia J. Kannapel et al., "What Can Be Said About Reform Progress," *From the Capital to the Classroom: Standards-Based Reform in the States,* Susan H. Fuhrman, Editor, University of Chicago Press, 2001, pp. 246-247.

[57] Daniel Koretz et al., *The Perceived Effects of the Kentucky Instructional Results Information System (KIRIS)*, Rand, 1996.

[58] Lynne Olson, "Survey Shows Testing Alters Instructional Practices," *Education Week,* April 24, 2002.

[59] Claus von Zastrow, with Helen Janc, *Academic Atrophy: The Condition of the Liberal Arts in America's Public Schools;* Council for Basic Education, March 2004.

[60] Laura Hamilton, "Assessment as a Policy Tool," *Review of Research in Education,* expected in late 2004. This article contains a comprehensive review of recent research of the impact of high stakes testing in the classroom and in the school.

Because content standards have typically been developed separately subject by subject at the state level, and entirely so for those developed at the national level, school systems are typically faced with a document indicating what constitutes complete coverage of a subject in a particular grade. But what about the balance of subjects within a grade? Guidance from above may be much more limited on this question. Standards-based reform systems tend to button down all the moving parts for an identified subject in an individual grade, much more so than the total content of what is taught from the beginning to the end of the day.

Clearly, decisions about what to teach are influenced by what is tested. So despite statements about the importance of geography, history, civics/social studies, health, art and music, if these subjects are not tested, the time spent on them in the classroom appears likely to be reduced in favor of subjects that are assessed—particularly those with high stakes attached to the results. There can be disagreement as to the implication of these increase and decreases. It is not the purpose of this report to offer judgments about the appropriate balance in a curriculum. But it is necessary to know what is happening in the distribution of instructional time.

A more measured approach therefore involves the kind of tracking of the distribution of instructional time that would permit education policy makers to assure themselves that there were not unintended shifts away from subjects not tested in accountability systems. Such tracking also would reveal whether intended results were being achieved, with respect to redirection of instructional time, as well as what subjects "win" and "lose" in the curriculum.

As to the governmental level at which instructional time should be systematically tracked, it would seem appropriate to do so at the level where such decisions get made. NAEP provides statewide achievement information that is highly regarded. Perhaps NAEP or some other program could also report on the distribution of instructional time, although this would take NAEP beyond its basic mission. Information about achievement and information about instructional time do go hand in hand.

The components of standards-based reform and test-based accountability have been taken up one by one, including more measured approaches to tracking both intended and unintended effects, such as changes in achievement, the distortion of the curriculum, and school non-completion rates. The purpose of this section is to point out a neglected element in this reform model: the use of assessment as a tool for informing instruction. Assessments—standardized and teacher made—hold promise as an integral part of instruction.

There has been considerable rhetoric about how testing requirements will give information helpful to improving instruction. But accountability assessments are "summative" evaluations, given at *the end* of a school year. These can hardly help teachers with individual students *during* the school year. In any event, there are unlikely to be enough test questions taken by a student in a specific area to yield detailed information to help inform instruction.

To be sure, such tests are not useless to teachers and principals. They yield broad information about how well students in a class performed, and may indicate areas of strength and weakness, depending on the nature of the test and how results are reported to teachers. Also, more states are using "student identifiers" that permit showing how individual students are growing from grade to grade.

But the thrust of accountability testing is to judge final results. Tests in education have been used extensively in the United States for measuring IQ, which was long mistakenly believed to be fixed; for gate-keeping for promotion or in getting into advanced education; for determining "readiness" to start school; for comparing students with other students; and for sorting students within classes (into "redbird" and "bluebird" learning groups, for example), and into different curriculum tracks. Compared to these other uses over the history of standardized testing, the use of testing to provide feedback to teachers to improve instruction has been minimal.

This neglect is indeed ironic. Alfred Binet's pioneering work in testing originated in France and then led to the birth of IQ testing in the United States—but Binet developed his test to help spot learning problems in young children, to help with instruction.

According to researchers Paul Black and Dylan Wiliam, external testing and accountability models that test for results at the end of a year's instruction focus on what goes into and comes out of the classroom, but not on what happens *within* the classroom—what happens in the "black box." Their interest is in "formative assessment"—knowledge from assessment that is actually applied by the teacher—and the central importance of such assessment to effective teaching and student achievement. They refer to assessments designed to reveal student shortcoming and errors as "diagnostic" but their actual use as "formative." Accountability testing, on the other hand, is called "evaluative" and "summative."[61]

Black and Wiliam synthesized about 580 articles or chapters of books from around the world. They deal with the bottom line question of whether use of formative assessment actually increases student achievement, based on around 20 studies selected by the quality of the methodology, and ranging over age groups from 5-year-olds to university graduates, and across several school subjects in several countries. They found typical "effect" sizes between 0.4 and 0.7, considered large in experiments with education innovations. They explain the meaning of these effect sizes:

- An effect size of 0.4 would mean that the average pupil involved in an innovation would record the same achievement as a pupil in the top 35 percent of those not so involved.
- An effect size gain of 0.7 in the recent international comparative studies in mathematics would have raised the score of a nation in the middle of the pack of 41 countries (e.g., the United States) to one of the top five.

Black and Wiliam report that "many of these studies arrive at another important conclusion: that improved formative assessment helps low achievers more than other students and so reduces the range of achievement while raising achievement overall." This is a very important finding as the United States struggles with large achievement gaps.

[61] Paul Black and Dylan Wiliam, "Inside the Black Box: Raising Standards Through Classroom Assessment," Phi Delta Kappan, 1998, p. 1. http://www.pdkneth.org/kappan/kbla9810.htm

Although a lot can be learned about formative assessment as a part of teaching, this cannot be undertaken in this brief review.[62] Those seeking more detailed information may want to explore the *Kappan* article, which provides advice on how to improve it; Black and Wiliam's detailed research synthesis;[63] and other works that address the use of assessment in classroom instruction.[64]

Standardized Formative and Diagnostic Testing in the United States. It is hard to pin down the frequency with which tests are used to give diagnostic information to teachers, as well as the volume of such use relative to accountability testing. A great many standardized tests are available for formative evaluation purposes, however. In a fairly recent review of the available tests and their characteristics, Linn and Gronlund draw a distinction between achievement test batteries and tests designed for diagnostic purposes. The former are "survey" tests that provide a general measure but with "too few items measuring each skill to provide much help in making instructional decisions." Survey tests have only a few items in an area of achievement, insufficient "for describing what individual students had learned, what they had yet to learn, and . . . what types of errors they are making." [65]

Many more items are used in tests designed for such diagnostic purposes. An example is the battery of *Metropolitan Achievement Tests*. The tests provide subscores for interpreting students' strengths and weaknesses with features built in to help teachers. Among others, they include:

- Various comparisons between mathematical scores to determine whether problem-solving performance is due to lack of computational skills, low reading ability, or carelessness.
- A higher-order thinking skills score from critical thinking items used in several subject-matter tests.
- Specific information for instructional planning for those students needing remediation, as well as for those performing above average levels.

Linn and Gronlund advise that such group-administered diagnostic batteries of tests are useful for identifying students "who could benefit from remedial teaching and individual help," but that for more serious learning problems, individually administered diagnostic tests and careful study of students' total development would typically be required. In this computer age, there is also likely a role for interactive online tests to bring assessment and instruction together.

Instructional Uses in Schools. The Council of Chief State School Officers conducted a study of Texas schools that are both high performing and high poverty to try to determine what factors helped them succeed.[66] Chosen from among schools that exempted few students in the Texas accountability system, these five high performing schools were identified on the basis of high test scores and high attendance.

What sets these schools apart from others? One distinguishing characteristic is that the staff at each school use student data "to identify areas where students can improve and where their own teaching strategies can be adjusted to meet students' needs." Administrators and faculty use assessment data to develop an intervention strategy where test results reveal it is needed, and to identify students in need of one-on-one tutoring, small group instruction, and other types of support.

The Ogg School, for example, makes intensive use of assessment in the early grades. In addition to the end-of-year and benchmark testing developed by the local district, staff collect student performance data every week "to quickly identify and address areas where students are having difficulty," and the data are used to guide changes in instruction.

---

[62] Two volumes describe how such ideas were put into practice: P. Black, C. Harrison, C. Lee, B. Marshall, and D. Wiliam, *Working Inside the Black Box: Assessment for Learning in the Classroom.* London, UK: King's College, London Department of Education and Professional Studies, 2002; and P. Black, C. Harrison, C. Lee, B. Marshall, D. Wiliam, *Assessment for Learning: Putting It Into Practice,* Buckingham, UK: Open University Press, 2002.

[63] Paul Black and Dylan Wiliam, "Assessment and Classroom Learning," *Assessment in Education,* March 1998, pp. 7-74.

[64] Richard J. Stiggins and Nancy Faires Conklin, *In Teachers' Hands: Investigating the Practice of Classroom Assessment,* State University of New York Press, 1992 and W. James Popham, *Test Better, Teach Better: The Instructional Role of Assessment,* Association for Supervision and Curriculum Development, Alexandria, VA.

[65] Robert L. Linn and Norman E. Gronlund, *Measurement and Assessment in Teaching,* Eighth Edition, Prentice-Hall, Inc., 2000.

[66] Council of Chief State School Officers, *Expecting Success: A Study of Five High Performing, High Poverty Schools,* April 2002.

In the upper grades, the benchmark TAAS tests are given every nine weeks. But third, fourth, and fifth grade students are assessed on specific objectives every one or two weeks. According to a fourth grade teacher, "We can pinpoint those kids (who are experiencing difficulty). We have a lot of graphs and charts. We'll look to see who's lowest and then we work with those kids harder to try to get them up to par with the other kids." The assistant principal of the schools says he "can tell you which child missed which question on what day."

At Peck School, the third grade teachers were finding that even with much effort and resources brought to bear, they still were not getting the results they wanted. So they started giving short tests of 17 to 20 questions that were correlated to the state curriculum standards. But rather than grade the test, each teacher went back to the students, discussed the problems she saw and then graded after the second time they took the test. The teacher reported that she saw a marked improvement in students' self esteem and academic performance.

A study by the Bay Area School Reform Collaborative looked at four years of California school testing data, identifying 16 schools where Black and Latino students were successful and 16 schools where they were not. The schools had similar ethnic and low-income populations. In one group, the Black and Latino groups were achieving as well as or better than their White and Asian classmates; in the other, there were the typical achievement gaps. The researchers spent a year visiting these schools to see how they differed. One key characteristic of the successful schools, as compared with the other 16, was the frequency of testing to guide individual student instruction.

An example of a success was a second grader at San Francisco's Treasure Island Elementary school. He couldn't read, and the teacher did not know why. The test revealed his problem. He did not know the sounds that corresponded to the letters, and now the teacher has him reading. The researchers found that the most notable thing about this school and the other 15 successful schools was that teachers diagnosed student needs frequently to guide their teaching. The teachers in these schools also analyzed the available data more frequently and used it to improve teaching. Furthermore, closing the achievement gap was often a primary goal of the principals.[67]

While accountability assessments given at the end of the year do little to shape individual instruction, some useful information can be obtained from these assessments if the effort is made. The data could be more useful, for example, if the student results follow the student from year to year. In the fall of 2003, just 21 states had a "student identifier" that records a student's achievement throughout K-12 education. Says Chrys Dougherty, research director for Austin-based National Center for Education Accountability, "You just can't interpret (a student's test score) unless you know how they were when they came into school . . . It's absolutely essential if you're going to do any kind of analysis at the middle and high school level."

The report recommends the following state actions:

- The state makes available to teachers student-level test score information on state exams that can be broken out by specific skill areas within a subject.
- The state uses the statewide database to measure student academic growth. This depends on the design of the test and the ability to track individual student records over time. [68]

These are examples of close ties between testing and instruction, and of frequent testing. Without more information it is hard to be sure whether they are all examples of the use of tests to convey diagnostic information to inform teachers; frequent testing could also be used in a narrow way, as part of efforts to focus in on preparing for a particular accountability test. What is argued here is that the use of assessments in the current education reform movement goes beyond their role in accountability systems, and includes the use of assessments to inform instruction during the school year. This was testing's first promise, but it ends up being a low priority in the total education system. Scientific studies have found that such use of testing pays off in terms of raising student achievement.

---

[67] Nanette Asimor, *San Francisco Chronicle,* December 8, 2003. The full report, entitled *After the Test: How Schools Are Using Data to Close the Achievement Gap*, can be found at the Bay Area School Reform Collaborative's website: http://basrc.org.

[68] David J. Hoff, "States Need Updates for Managing Data, Analysis Concludes," *Education Week,* October 22, 2003.

Standards-based reform and test-based accountability are focused on raising achievement levels of students. The question is always asked: Will the higher standards result in more students leaving school? At some level of standards, students not faring well will choose to leave if they find they cannot reach the standards. Students may be encouraged to drop out or even be pushed out by some schools. Conscious decisions may be made to raise standards even when there is an expected increase in the non-completion rate, but measures still need to be available to inform those choices. Of course, some people maintain that more challenging content makes more students opt to stay in school, because many students are bored by the low level content being provided them. This may well be true for some number of students.

What we need to know is the tradeoff between required achievement levels and school completion. This requires good statistics on how many students leave school without completing it.

One of the national education goals defined in 1989 by former president Bush and the nation's governors was to reach a high school completion rate of 90 percent by 2000. Each year, the National Education Goals Panel duly reported the available completion rates, which showed that no progress was being made. Despite the lack of progress, however, there is broad acceptance of the proposition that we should strive both for higher achievement and a higher rate of school completions, with the award of a diploma, and that we need accurate measures of the completion rate, as well as of achievement.

Heightened Attention. During the past few years, a number of studies—notably by Jay P. Greene, Andrew Sum et al., and Christopher B. Swanson and Duncan Chaplin—have examined existing measures of high school completion and found them wanting.[69] For those deeply interested in these evaluations of existing statistics, the reports by these authors may be consulted. Below are a few of the problems with the available measures identified:

- In reports derived from the Census Bureau's Current Population Survey, regular diplomas and GED certificates have been lumped together. This practice has now been discontinued. While the GED is an important means of getting a credential, separate statistics for regular diplomas are needed to track how schools are performing.
- The census reports also have the problem that some minority populations are undercounted. Further, the results rely on self reports, with one person in the family answering the questions for all family members.
- Since the early 1990s, the National Center for Education Statistics (NCES) has been providing estimates based on the Common Core Data system, with estimates for 12 states in 1995 rising to 39 states for the 2000-2001 school year. The estimates of the completion rate are built from the states' reports of student dropouts for each of the four years. These dropout rates seem to be widely underreported, leading to considerably higher estimated rates of completion than obtained under several other methods of estimating state level rates. These are described below, and compared to the NCES estimates.

The above studies have received some media coverage and have heightened concern about the accuracy of dropout rates. Recently, a considerable number of news stories about inflated school completion statistics have appeared. These stories have been placed in the context of the severe pressure on schools to raise achievement scores, as well as completion rates, and the response to move low-achieving students out of the schools. Some reports charge outright fraud in reporting. Others charge that creative ways have been used to report students as transfers rather than as dropouts. Still others charge outright expulsion of students, or forced transfer, for example, to adult GED programs.

[69] These studies are referenced later in this section.

One example is New York City. In July 2002, a panel of administrators of the city's GED programs for adults were invited to the American Youth Policy Forum in Washington, DC. They described how they were now getting many young people who had just left their high schools after being "advised" to leave and enroll in a GED program. The resources for the GED programs were being strained by the influx.

One particular school has been singled out in a class action lawsuit against the New York City Education Department in Federal District Court in Brooklyn. According to *The New York Times,* the suit charged that, although New York State law provides that students have the right to remain in school until they are 21, officials at the Lane school routinely told difficult students it was time to go elsewhere. The school, the article said:

> . . . now has 3,200 students, down from 4,500 five years ago. Its official statistics show a striking decline in enrollment in the upper grades. As of last October, Lane's annual report said, there were 1,266 students in ninth grade, 1,070 in tenth grade, 341 in eleventh grade and 325 in twelfth grade.[70]

On January 8, 2004, *The New York Times* reported that the city Department of Education settled the lawsuit and agreed to take back students discharged from the Lane school since January 1, 2000. They can re-enroll in Lane or attend another school. In addition, the Department would also start a neighborhood social center, called the Young Adult Success Center. It will offer at least 12 hours a week of academic and other support services for Lane students, current and past. Lawsuits against two other schools, with similar charges made, are still pending.[71]

Alternative Methods for Estimating Completion Rates. The studies referred to earlier have critiqued existing completion estimates and have advanced alternatives, based on available data. Depending on the study, estimates are made for the state, district, or city. In this brief presentation of the measurement problem, the results of these different approaches are shown at the state level in Table 4. The approaches used are described below, including my own estimates for 2000. For comparison purposes, the statistics reported by the states in September 2003 under the requirements of the No Child Left Behind Act (NCLB) are also provided.

[70] Jennifer Medina and Tamar Lewin, "High School Under Scrutiny for Giving Up On Its Students," *The New York Times*, August 1, 2003.
[71] Tamar Lewis, "City Settles Suit and Will Take Back Students," *The New York Times*, January 8, 2004.

Table 4:

Completion Rates by Different Methods

	1 NCES 2001	2 Greene 1998	3 Swanson & Chapin 2000	4 Sum et al. 1989	5 Barton 2000	6 State Rpts NCLB 2001
US	--	71	66.6	68.7	69.6	--
AL	80.0	62	61.3	61.6	65.1	--
AK	75.2	67	59.3	62.5	72.1	84.5
AZ	68.3	59	--	54.9	55.0	70.8
AR	79.1	72	69.2	72.5	72.6	85.1
CA	--	68	68.3	67.0	68.8	86.9
CO	--	68	70.3	62.2	67.4	81.8
CT	86.6	75	76.3	81.0	85.6	87.3
DE	81.6	73	67.0	73.3	64.8	83.1
DC	--	59	53.5	72.1	48.0	63.5
FL	--	59	49.9	60.3	59.2	64.7
GA	71.1	54	53.5	57.2	58.1	62.0
HI	77.7	69	62.3	72.6	82.6	78.9
ID	76.9	78	74.7	70.7	73.1	77.1
IL	75.8	74	73.9	81.2	71.8	85.2
IN	--	74	70.8	73.0	67.7	91.0
IA	89.2	93	77.6	70.4	83.9	89.4
KS	--	76	73.3	70.6	74.3	85.1
KY	79.9	71	63.7	68.5	70.8	80.7
LA	65.0	69	59.5	60.8	63.9	--
ME	86.5	78	72.5	76.6	80.0	86.1
MD	83.2	75	72.7	76.3	79.6	84.7
MA	86.3	75	75.5	77.7	74.4	--
MI	--	75	74.0	69.3	69.0	86.0
MN	82.5	82	79.5	81.1	81.8	87.9
MS	77.3	62	59.2	59.1	59.3	72.0
MO	81.0	75	71.3	71.1	72.4	82.5
MT	82.1	83	76.5	73.1	79.1	84.1
NE	83.9	85	77.7	82.8	83.7	84.0
NV	73.5	58	55.2	61.7	60.4	63.7
NH	--	71	72.8	78.5	68.4	84.5
NJ	88.0	75	81.6	75.1	82.7	88.7
NM	74.4	65	60.1	63.0	67.2	76.6
NY	81.6	70	60.2	67.5	65.3	75.0
NC	--	63	60.3	63.3	61.2	92.4
ND	90.1	88	79.7	81.3	83.5	90.6

Table 4:

Completion Rates by Different Methods—continued

	1 NCES 2001	2 Greene 1998	3 Swanson & Chapin 2000	4 Sum et al. 1989	5 Barton 2000	6 State Rpts NCLB 2001
OH	81.0	77	70.7	73.3	76.6	82.8
OK	79.2	74	67.3	70.7	72.1	68.8
OR	76.4	67	62.6	62.1	65.8	79.5
PA	84.0	82	75.2	77.9	76.7	86.4
RI	79.8	72	72.6	75.1	63.2	71.4
SC	--	62	48.4	58.9	57.7	77.6
SD	84.6	80	78.0	71.0	77.3	97.0
TN	79.5	60	48.6	60.7	61.2	75.7
TX	--	67	62.9	65.3	67.7	82.8
UT	82.6	81	79.4	76.1	73.2	86.1
VT	81.9	84	72.9	85.5	88.2	82.0
VA	83.8	74	77.5	70.9	71.4	84.7
WA	--	70	62.3	70.1	71.3	79.0
WV	83.4	82	70.2	78.4	79.8	--
WI	90.0	85	76.6	76.6	79.1	90.8
WY	76.5	81	74.7	71.0	78.1	77.2

There is considerable similarity among the estimates made by independent researchers, shown in columns 2, 3, 4, and 5. Nevertheless, some variation in the estimates is evident, particularly in certain states. In Iowa, for example, estimates of high school completion ranged from 70 to 93 percent. Some of the variation may be explained by the fact that these estimates are not for the same years, but state rates would not be expected to change greatly in just one or two years.

The greatest difference is between the estimates by private researchers in columns 2, 3, 4, and 5, and the state reports submitted under NCLB (column 6), which are mostly higher. The NCES estimates (column 1) are also mostly higher than these "outside" estimates. The methods used are described below.

NCES Method (Column 1). The numerator in the NCES method is the number of high school completions in a particular year, including regular high school diploma and other credentials (but excluding GEDs). The denominator is that number plus all who dropped out in each of the four years.[72]

Greene Method (Column 2). Greene starts with the eighth grade public school enrollment in the fall of 1993. Then, he uses the number of high school diplomas awarded in the spring of 1998 when the 1993 eighth graders would be graduating. Also, to adjust for the possibility that students moving into or out of an area would distort the graduation rate, Greene adjusted the 1993 eighth grader counts for the student population change in that jurisdiction and for each ethnic/racial subgroup between 1993-94 and 1997-98 school years.[73]

Swanson and Chapin Method (CPI, Column 3). In this approach, an estimate is made of the "probability that a student entering the 9th grade will complete high school on time with a regular diploma." "Promotion

[72] Beth Aronstamm Young, *Public High School Dropouts and Completions from the Common Core of Data: School Year 2000-01*, November 2003, p. 4.

[73] Jay P. Greene, *High School Graduation Rates in the United States*, the Manhattan Institute, April 2002, p.7.

rates" are calculated for the future years by examining the difference in enrollment by grade for the past years. These were applied to estimate the number graduating in the 1999-2000 school year.[74]

Sum, et al, Method (Column 4). This approach uses just two numbers: the annual number of high school diplomas awarded (pubic and private schools) as reported by NCES, and the 17-year-old population as reported by the U.S. Bureau of the Census. This is the same approach used by NCES for national completion rates going back to the 1880s. Sum also makes the calculations based on the 18-year-old population, to test whether a single age group could represent the cohort of students that started the first grade, whose ages would vary somewhat. He found that there was very little difference between the two sets of completion rates.[75]

Barton Method (Column 5). This approach is very similar to the Sum method. However, I also tried to get a measure of trend over a ten-year period, and chose to use census years 1990 and 2000 for this purpose. Intervening years are based on estimates by state for individual age intervals by the U.S. Bureau of the Census. This seemed a better approach for a trend comparison. Since private school data on graduations is collected only every other year, I used 1989 and 1999 data to add to public school enrollments for 1990 and 2000. Private school graduations are about 10 percent of total graduations, and steady at the national level; there were 285,000 in 1989 and 273,000 in 1999. In comparing the populations of 17- and 18-year-olds for each year, I found that while they were very similar for 2000, there were many states in 1990 with a significantly higher number of 18-year-olds than 17-year-olds. In view of this, I decided I could not use a single age group to represent the age of the graduating cohort in order to make comparisons between 1990 and 2000. So I constructed a cohort of both 17- and 18-year-olds, based on the age of graduating seniors, using 23.8 percent of 17-year-olds and 76.2 percent of 18-year-olds.[76]

State Reports Under NCLB (Column 6). The No Child Left Behind Act required the states to report high school completion rates by September 2003 for the 2000-2001 school year. The rate is the percentage of students who began high school and graduated with a "regular diploma." Alternatives definitions could be used if approved by the U.S. Department of Education. I have taken these rates from *Telling the Whole Truth (or Not) About High School Graduation Rates*, an analysis by the Education Trust issued in December 2003.

It should be noted that not all these methods could be used at the individual school level, and that getting rates for a school is much more challenging than for a state as a whole. When students leave a school it is often hard to track them, but they are likely to be in schools somewhere within the state, unless they have dropped out. Under any method, it is extremely challenging to apply the concept of school completion at the level of the individual school for accountability purposes. This is particularly true in poorer neighborhoods where there is considerable mobility so that an individual school is only one among several that may be involved with a particular student.

Given the requirements for reporting under local, state, and federal laws, however, the challenge must be met, or the requirements changed. At the federal level, the U.S. Department of Education issued a statement on December 22, 2003, in response to a critical report released by The Education Trust. Deputy Secretary Gene Hickok said the following:

> Dropout data collection has been an issue for years because no state collects the same way. Recognizing that this issue has been a Gordian knot, Secretary Paige last week announced the awarding of a contract to the National Institute of Statistical Sciences for convening a group of experts to review the methods for reporting high school dropouts and on-time graduates.

[74] Christopher B. Swanson and Duncan Chaplin, *Counting High School Graduates When Graduates Count,* Education Policy Center, the Urban Institute, February 24, 2003, pp. 19-20.

[75] Andrew Sum, et al., *The Hidden Crisis in the High School Dropout Problems of Young Adults in the US: Recent Trends in Overall School Dropout Rates and Gender Differences in Dropout Behavior*, prepared for The Business Roundtable, Center for Labor Market Studies, Northeastern University, February 2003, pp. 34-35.

[76] To arrive at these proportions, ETS research staff extracted the age of seniors at the time they took the NAEP assessment in 2000. Then, these seniors were "aged" to see how old they were at the time of graduation. Since all but about 10 percent were either 17 or 18, I used these two age groups, on a proportional basis, to get an estimate of the size of this cohort of graduating age.

Trends in Completion Rates. All of the calculations made at the state level and shown in columns 2, 3, and 4 in Table 4, except for the statistics published by NCES, are available only for a single year. But how have the state rates changed over this period of standards-based reform? Are the methods for estimating a rate good enough for estimating the change in the rate from year to year? There are no NCES estimates of completions by state that go back as far as 1990. The NCES estimates began with a handful of states, 12 in 1995, and grew slowly to 39 for the 2000-2001 school year.

While the results are not included in the summary table, a new study also uses the cohort approach and references a different set of cohort-based estimates. This 2004 study by Walt Haney and his colleagues at Boston College is the most ambitious yet, making state-by-state estimates going back to 1970.[77] Haney uses enrollment data and diplomas awarded, with no attempt to make adjustments for such things as population growth. He then compares the results of his straightforward computations (all 10,000 of them) with the complex procedure used by John Robert Warren of the University of Minnesota in a paper published in 2003.[78]

The Warren estimates were cohort based, but used census data to adjust for migration patterns and current population survey data to adjust for migration and grade retention. Haney's simple estimation procedure showed similar results. Haney concluded that, with few exceptions (noted in his text), "the simple grade 9 to graduation rate provides a good approximation to the more complex adjusted rate calculated by Warren . . . Though not reported by Warren, the correlation between the simple graduation rates and his adjusted rate is 0.903. And when the two outlying cases of the District of Columbia and Nevada are excluded, the correlation rises to 0.960." Accordingly, Haney concluded that the simple method is, under most circumstances, a good proxy for more complex calculations.

In addition to making estimates for 2000, I also made estimates for 1990, using the same procedures for both years. The results are shown in Figure 7. Increases occurred in the completion rates in only seven states: California, Connecticut, Maryland, Rhode Island, West Virginia, Utah, and Vermont. While the increases for four of the states are modest, the increase for Vermont is a whopping 23 points, with the increase due to both public and private school enrollments. The Haney estimates, on the other hand, show a small decline for Vermont. The reason for such a difference bears looking into.[79]

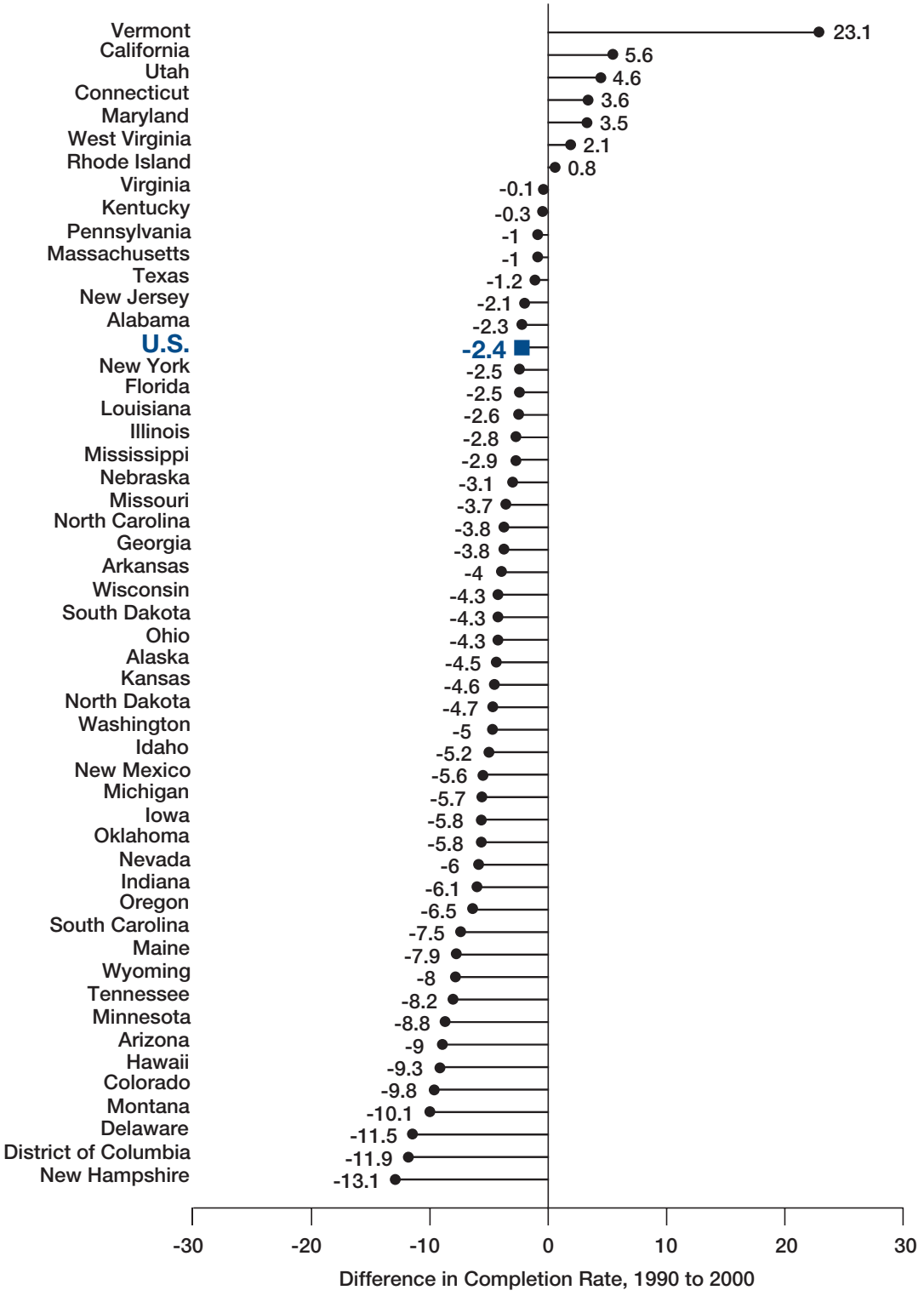
[77] Walter Haney et al., *The Education Pipeline in the United Sates, 1970-2000,* The National Board on Education Testing and Public Policy, Boston College, January 2004.

[78] John R. Warren, *State Level High School Graduation Rates in the 1990s: Concepts, Measures and Trends,* paper prepared for presentation at Annual Meeting of the American Sociological Association, Atlanta, August 2003, cited in Haney et al., 2004.

[79] In the Barton method, for example, the rate is affected by diploma awards to students in residence in other states.

Figure 7:

Change in High School Completion Rates, 1990 to 2000, by State*



*Source: See description of Barton method, page 48.

The general picture, however, is one of *declining* completion rates between 1990 and 2000. Most of the states lost ground:

- 16 states declined up to 3.9 percentage points;
- 18 states declined from 4 to 7.9 percentage points;
- 9 states declined from 8 to 11.9 percentage points; and
- 1 state declined 12 or more percentage points.

These decreases are widespread, and many of them are substantial. A pattern is not obvious, although further analysis might find one. While the magnitudes sometimes differed, the Haney estimates also showed most states in decline, ranging up to 15 percentage points, but with 17 states having declines of 5 points or less.

How have rates changed over the longer term? Completion rates are available for the nation as a whole as far back as the 19th century, but yearly calculations cannot be made for individual states prior to 1986 because adults receiving diplomas were included in the counts. Calculations can be made for subsequent years, however, preferably at two-year intervals when NCES counts private diplomas, and using the census population estimates for years between the decimal census. In contrast to 1990, there would be few years when 17- and 18-year-old populations are not nearly identical.

For almost all states, the approach described and used for national data would probably yield a rate that is comparable over time. Estimates for the cohort could be improved by using both the 17- and 18-year-old population counts and weighting them based on the proportion of students who are 17 and 18 at graduation. In any event, calculations of completion rates on this basis are a good reality check on rates derived in other ways, and on statements about how rates in a state have changed over time. Of course, such an approach cannot be used at the school level. These estimates of trends and levels from 1990 to 2000 are generally very similar to the estimates by Haney referred to earlier.

Some Measurement Considerations. The seemingly simple matter of counting how many students complete high school is actually very complex, particularly at the level of the individual school. When such a measure is applied to the high school, how is responsibility apportioned among the high school and the feeder schools responsible for the first eight grades? Students getting behind early are candidates for dropping out later.

A few observations arise as a result of this somewhat cursory investigation.

1. There is potential for better tracking of high school completion rates by applying to the individual states the NCES approach that has long been used for national data, producing yearly estimates back to the mid-1880s. NCES is now publishing data on diplomas awarded by public schools by race and ethnicity, and that could make estimates of completion more useful. If such data also become available for private school graduates, estimates could be provided about state completion rates by race and ethnicity. While there are limitations to this approach, it relies on only two numbers: the Census Bureau population counts and the number of diplomas awarded. The latter figure is unlikely to be distorted the way other school counts may be. This certainly would not be the only measure; there is a need for triangulation using several approaches.
2. It seems worthwhile to perfect and use the cohort approach used by Jay Greene and others. This provides a rate for public schools only. And it requires only enrollment data and the count of diplomas awarded. Enrollment data, particularly at the individual school level, is far from perfect, however. For years, in the 1990s, the DC school system did not know how many students it had in school; record-keeping was very poor. And there are questions about what constitutes "enrollment." One Superintendent of Schools, then a member of the Blue Ribbon Panel on the SAT Score Decline, commented that some schools have the tacit understanding that if a student shows up in the morning for the count for average daily attendance—the basis for payment from the state to the school—no one will hold the student accountable for being there the rest of the day.

3. The need and demand for accurate statistics on completion will grow considerably as the use of tests for high school graduation grows. According to The Education Trust, about 30 states will be using such tests by 2009.[80] The policy question will be asked as to whether the tests result in a growth in non-completions, and what the tradeoff is between higher achievement and the proportion getting a diploma. Better decisions will likely be made if they are based on sound data.

4. The meaning of what constitutes "graduation" and a high school "diploma" is evolving, and will likely continue as exit tests phase in. With Maryland moving toward requiring five subject matter tests for graduation, State Superintendent Nancy Grasmick proposed five diplomas, based on how many tests are passed.[81] For 2000-2001, NCES broke down the total of high school completions into those receiving diplomas and those receiving some other kind of certificate.[82] In terms of state completion rates, the "other completer" component ranged from a high of 7.2 percent in Tennessee down to 0.1 percent in Missouri, among the 39 states for which NCES was able to provide data. The differentials are already large enough to factor in comparisons based on the total number of completers. In terms of the total percent completing, these two states are about equal. But 81 percent of those in Missouri completed *diplomas*, compared to only 72 percent in Tennessee. A single "high school graduation rate" increasingly may not tell enough of the school completion story.

5. The assignment of a completion rate to an individual high school is hugely complicated by student mobility. This is particularly the case at the lower socioeconomic levels. In fact, we have almost no measure of student mobility rates. A 1994 GAO report found that a quarter of Black and a quarter of Hispanic third graders had already changed schools three or more times, while only 13 percent of White students had done so.[83] Students changing schools a number of times are particularly challenged. One larger problem is that their school records and test scores are not likely to follow them to the new school. A system of tracking individual students could help both in terms of the achievement of these mobile students as well as improve statistics on the power of schools to retain students, as transfers are separated from dropouts. However, we do know that it is hard to track the same individuals over more than a few years in longitudinal surveys. Perhaps an identifying number, such as a Social Security number, could be used for tracking throughout the schooling period.

6. This discussion has been about the adequacy of the measures of school completion. Even with good measures, when comparisons are made between schools and states, different standards may be used to determine what constitutes the necessary level of achievement for graduation.

[80] Yhan Q. Mui, "Maryland to Give Class of '09 Exit Exams," *The Washington Post*, December 4, 2003, p. A1.
[81] Mui, 2003. The proposal was rejected, however.
[82] Young, 2003, Table 5.
[83] Paul E. Barton, *Parsing the Achievement Gap: Baselines for Tracking Progress*, Policy Information Report, Policy Information Center, Educational Testing Service, October 2003, p. 23.

In Conclusion

The standards model, accompanied by test-based accountability, seems still to be gathering momentum, at all levels of governance. The reason for the model's popularity seems quite obvious: it is logical to define what students should know, get instruction and curriculum in line with this, decide what constitutes required performance, determine through quality assessments whether gains are being made, and set goals for progress. The devil is in the details of implementation, of course.

It is clear that there is substantial unfinished business and considerable unevenness in the way the standards-based model is being applied. State-level content standards are often not translated into the curriculum, lesson plans, instructional materials, and professional development, or at least not fully so. So what is taught is often not aligned to what is expected. The test is frequently not aligned to these state content standards, or to what is actually taught. However, the tools and knowledge are available to measure and improve alignment.

There is also unfinished business in the methods being used to measure students' progress and gauge the effectiveness of teachers and schools. Too often, these measures are not based on the learning that actually occurred in the classroom in a year of school, but instead encompass school and life prior to that year. But good models and experience exist to lead the way in this area, too. While it is important to measure students' levels of achievement for school accountability purposes, measuring "gains" made during the school year are also needed. And where the level of achievement is the focus, measures should go beyond the percentage of students reaching a point on a scale labeled "proficient;" more of the test results need to be used to examine, for example, changes in the top and bottom of the score distribution.

Where alignment of the various parts of the standards-based reform model has not been achieved, and evaluation has not focused on what is actually learned in the classroom, the meaning of test results is called into question. Credibility will be a key to sustaining the reform effort.

Where alignment is out of whack, teachers and principals struggle with getting students ready for tests with real consequences. Where the curriculum and textbooks are not clearly delivering the content on which these tests are based, teachers are up against the blackboard. They do things that are labeled as unwarranted "teaching to the test." If they are not left in the dark as to how to deal with such mismatches, they are frequently left in a twilight zone.

The choice of which subjects to test may be diverting instructional time away from subjects that are not tested. Whether there are intended or unintended shifts in instructional time will be known only if their measurement becomes part of the standards-based reform system. Shifts in instructional time have clearly occurred in some paces, as revealed by special studies such as the recent one by the Council on Basic Education.

Standardized testing began with a promise of providing information that would help inform instruction; that was Alfred Binet's declared purpose. But the use of educational testing as an integral part of instruction has been eclipsed by other uses—from measuring IQ, to sorting students, to measuring accountability. Diagnostic and formative testing have proven promise of helping to raise achievement, however, and they need to be incorporated as a key feature in the reform movement.

The emphasis of reform so far has appropriately been about the quality rather than quantity of education. There is one area where quantity matters greatly, though, and that is the amount of schooling completed. Recent research has shown that school completion rates are lower than has been regularly reported, and rates appear to be falling despite goals to the contrary. The measures of school completion, and the rates of completion, need to be improved at all levels: school, district, state, and nation.

No attempt is made here to critique particular laws at particular levels of governance. There is no judgment that the standards-based reform movement is way behind schedule. Rather, the assumption is that the model is—and should be—evolving and continuously improving based on knowledge and experience. The more measured approaches herein suggested are offered to assist in that evolution.

Policy Evaluation and Research Center

Policy Information Center

Visit us on the Web at www.ets.org/research

Listening.
Learning.
Leading.

85995-43375 • U114E6 • Printed in U.S.A.