A POLICY INFORMATION PERSPECTIVE

Too Much Testing

of the Wrong Kind; Too Little of the Right Kind in K-12 Education

by Paul E. Barton

POLICY INFORMATION CENTER
Research Division
Educational Testing Service
Princeton, New Jersey 08541-0001

Additional copies of this report can be ordered for $9.50 (prepaid) from:

Policy Information Center
Mail Stop 04-R
Educational Testing Service
Rosedale Road
Princeton, NJ 08541-0001
(609) 734-5694
Internet – *pic@ets.org*
http://www.ets.org
Copies can also be downloaded from
www.ets.org/research/pic

March 1999

Table of Contents

Preface

Any discussion in the United States of how well American students are educated, and how to "reform" education, comes — either quickly or eventually — to testing. The testing enterprise in K-12 education has mushroomed the last quarter-century; Americans want numbers when they look at students, schools, state education systems, and how America's students compare to those of other countries. Among political leaders, testing is turning into a *means* of reform, rather than just a way of finding out whether reforms have been effective.

I believe that there is much that is wrong in this system, that there are signs here and there of improvement, and that there are ways to make assessment much better in serving teaching and learning. We have more and more of these numbers, but they are too often not adding up to good information.

Paul E. Barton
Director
Policy Information Center

Acknowledgements

I am indebted to a number of people for providing thorough, thoughtful, and often critical reviews: Robert Linn at the University of Colorado, Robert Stake at the University of Illinois, Emerson Elliott, former Commissioner of the National Center for Education Statistics in the U.S. Department of Education, and at Educational Testing Service, Henry Braun, Ted Chittenden, and Charlotte Solomon. While there was substantial agreement on the general thesis of the manuscript, not all would agree with all that is in the final report.

Carla Cooper provided desktop publishing, Kirsty Brown was the editor, Ricardo Bruce was the cover designer, and Jim Chewning was the production coordinator.

Introduction

"We need less frequent but far better testing."
Albert Shanker, President, American Federation of Teachers, 1993.

There is just too much standardized testing going on in our schools, lamented an unlikely source, the late Gregory Anrig, then-president of the world's largest testing organization, Educational Testing Service. (Before that, he was Chief State School Officer for Massachusetts, and had been an educator throughout his career.)

The testing enterprise has mushroomed in the United States. To show you mean business in dealing with crime, you call for more prisons and mandatory sentencing. To show you are tough on welfare reform, you ask for time limits. To show seriousness in raising educational achievement in the U.S., you call for more frequent and more rigorous testing. Those who oppose testing are accused of protecting teachers and the educational system, and not putting children first.

The critics of such massive testing, including many in educational measurement, offer the following complaints. Tests have been composed mostly of multiple-choice questions, which cannot assess a student's ability to come up with his or her own answers. Commercial or state tests may not test what local schools are actually teaching. Some critics argue that teachers are pushed in the direction of narrowing instruction to what they think is on the test. Further, test preparation sometimes *becomes* the instruction, with instructional materials mimicking the formats and exercises that appear on such tests.

In the 1990s, there have been constructive attempts to improve the testing enterprise. Serious efforts have been made to broaden tests beyond multiple-choice questions, and to include open-ended questions, "performance" assessments, and portfolios. (However, the assessment reform movement has been slowed over issues of reliability and measurement error.) The National Assessment of Educational Progress (NAEP) has been expanded. The large effort of the New Standards Consortium of states and school districts has tried to construct more educationally useful tests, and has involved teachers extensively in building tests from the ground up. And, in case tests are being "taught to," the New Standards Consortium wants to turn this into a positive, rather than a negative. There is no intention here of reviewing this decade of reform in standardized testing — there have been improvements — but we are still left with some major challenges to fully harness assessment to the purposes of education reform. *Most* of the testing today is not much changed from what it was a dozen years ago.

This report starts with a quick review of the beginning of standardized testing in the schools, and the reasons for growing reliance on testing. It summarizes the recent promising trends and suggests how testing

for accountability could be less intrusive and provide better information about achievement in schools. The proposal for a voluntary national test is examined, and an alternative is offered in view of the political stalemate in which the proposal is mired. The most promising development on the horizon — setting content standards and aligning curriculum and assessment to them — is described as what I call the "patient approach." The challenge in setting performance standards is also set forth. The purposes of accountability assessment are recognized; the alternative of measuring "value added" for these assessments is described, and examples of its use are provided; and exit examinations are discussed. Finally, the critical role of the teacher in assessing students is examined, as is the need to equip teachers with the knowledge and tools to use assessment in day-to-day instruction.

The Reason Why

Improving testing is important because testing has become, over the last 25 years, the approach of first resort of policymakers.

Robert Linn, in his 1995 Angoff Lecture at ETS, explains why:

> *1. Tests and assessments are relatively inexpensive. Compared to changes that involve increasing instructional time, reducing class size, attracting more able people to teaching, hiring teacher aides, or enacting programmatic change involving substantial professional development for teachers, assessment is cheap.*
>
> *2. Testing and assessments can be externally mandated. It is far easier to mandate testing and assessment requirements at the state or district level than to mandate anything that involves change in what happens inside the classroom.*
>
> *3. Testing and assessment changes can be rapidly implemented. New test or assessment requirements can be implemented within the term of elected officials.*
>
> *4. Results are visible. Test results can be reported to the press. Poor results in the beginning are desirable for policymakers who want to show they have had an effect ....*

Exposing the existence of substandard education has long been the objective of written examinations, and while the boom has come in the last quarter century, the popularity of testing is long-standing. In *Testing in American Schools: Asking the Right Questions* (1992), the Office of Technology Assessment provides some early history of testing in American schools from the mid-19th century.

> *"The idea underlying the implementation of written examinations... was born in the minds of individuals already convinced that education was substandard in quality. This sequence — perception of failure followed by the collection of data designed to document failure (or success) — offers early evidence of what has become a tradition of school reform and a truism of student testing: tests are often administered not just to discover how well schools or kids are doing, but to obtain external confirmation — validation — of the hypothesis that they are not doing well at all."*

Robert Stake, of the University of Illinois, gives a succinct summary of the different waves of education reform in recent American history (in "Some Comments on Assessment in U.S. Education").

> *"Earlier, in the Century's third quarter, the impetus for changing American schooling was the appearance of Sputnik.*

It was reasoned that the American schools were unsuccessful if the Soviets could be first to launch spacecraft. College professors at the National Science Foundation stepped forward to redefine mathematics education, and the rest of the curriculum, creating a new math, inquiry teaching, and many courses strange to the taste of most teachers and parents. According to Gallup polls year after year, citizens expressed confidence in the local school but increasingly worried about the national system. In the 1960s, curriculum redevelopment was the main instrument of reform, but in the 1970s state-level politicians, reading the public as unhappy both with traditional and federalized reform, created a reform of their own. Their reform spotlighted assessment of student performance."

The mushrooming of standardized testing started in earnest in the early 1970s with the "minimal competency" testing movement, which, at best, helped achieve more minimal competency. It continued to grow in the 1980s, as a response to *A Nation at Risk*. Such statewide testing probably misinformed more than it informed. By 1987, John Cannell, a physician in West Virginia, had noticed that many states or schools were claiming that their students were above average. A sustained investigation revealed that students' scores almost everywhere were above average, a phenomenon that came to be dubbed the Lake Wobegon effect. He concluded that "… standardized, nationally normed achievement tests give children, parents, school systems, legislatures, and the press inflated and misleading reports on achievement levels."[1]

Robert Linn, in *Assessment-Based Reform: Challenges to Educational Measurement* (ETS Policy Information Center, 1995) was a leader in assessing Cannell's complaints, and summarized his conclusions this way:

"There are many reasons for the Lake Wobegon effect …among the many are the use of old norms, the repeated use of the same test form year after year, the exclusion of students from participation in accountability testing programs at a higher rate than they are excluded from norming studies, and the narrow focusing of instruction on the skills and question types used on the test …."

Whatever the reason for the Lake Wobegon effect, it is clear that the standardized test results widely reported as part of accountability systems in the 1980s were giving an inflated impression of student achievement.

[1] Cannell, J. J. (1987). *Nationally normed elementary achievement testing in America's public schools: How all 50 states are above the national average* (2nd edition). Daniels. West Virginia: Friends of Education.

Promising Trends and Reducing Intrusion

In the 1980's and 1990's it was elected officials — governors and state legislators — who continued to press for more testing.

While the bulk of it was mass use of standardized testing in ways that are deplored in this report, there have been promising developments. For example, in the 1980s the Southern Regional Education Board (SREB) began to use NAEP to get state-level results among member states, and led the way for the expansion of NAEP to do this. SREB was a leader both in using data to track policy implementation — testing data included — and in setting goals for education, years ahead of the National Goals and the National Education Goals Panel.

Of course, in the 1990's, tests are also expected to somehow be a means of reform, and too often, to be the principal means. *How* this is to work is not clear. However, it is perfectly clear that standardized testing is here to stay. The question is whether it can be made to play a more constructive role, or will continue to be used as a shortcut across quicksand.

Testing has been improving during the 1990s, and is slowly being aligned to new and higher content standards. However, pitfalls still exist: testing is often an instrument of public policy to affect schools, to grade schools, to scold schools, and to judge whether other improvements in the education system are having the desired effect. Most of these tests have not been validated for these purposes. By and large, tests are not used within the classroom by teachers as *their* means of assessment; rather, teachers know the tests are used to grade them. Surveys have shown that teachers use the tests they make themselves, or the tests that accompany the instructional materials provided by private publishers.

We can change the way we administer standardized tests for school/teacher control and accountability, with much less intrusion into the classroom. Sampling, as is done in NAEP, is more effective than testing every student frequently, with the same test, providing individual scores for all. Sample-based approaches will provide *better* information about schools (see later discussion), and will be much less intrusive into instructional settings and require less frequent testing. If the objective is a report card on the schools, testing every couple of years will accomplish the purpose.[2] Changes in education cannot be accomplished abruptly; a meaningful reordering of an important phase of the instructional process takes time. There is an impatience at work here that is typically American; it is like pulling up the carrots to see how they are growing.

[2] How often tests are needed may depend partially on school size. Robert Linn points out that testing has to be more frequent in small schools "because of instability in the scores as a function of cohort" and "having results every year enables one to smooth out extreme fluctuations or reduce them by taking two-year averages, as has been done in Kentucky" (personal correspondence).

While I am here advocating sample-based assessment as less intrusive, and capable of broader coverage of subject matter than continuous mass testing of all students, I do so in the hopes of limiting the harm done to instruction, and improving the measure of what students have learned over a period of time. Many questions remain, however. Most tests are constructed to measure the knowledge a student has acquired. They have not been designed for the accountability purposes for which they are now regularly used; they are not designed, for example, as measures of teachers' capabilities. They have not been validated in this use to assess whether they have the intended consequences. Have the results based on testing, for example, been compared to results of other rigorous efforts to evaluate teacher and school performance? Have the results been useful in changing teacher behavior in desired ways? Do the tests actually measure what it is that the policy makers who ordered their use intended? I have pointed out elsewhere the misuses of standardized testing; the use of such tests for accountability without meeting standard and well-known methods of validation amounts to testing malpractice.

What we want from standardized testing is *better information* for teachers, administrators, policymakers, and the public. Testing used presently too rarely results in better information to aid instruction and achievement.

"National" Testing

Amidst the testing explosion at the state and district level, the Clinton Administration is attempting to launch a new National Test, something President Bush had also proposed.

It was first discussed in the State of the Union Address of 1997 as a "test." Later, the President referred to the need for "national standards." The test would be used to determine the extent to which national standards were being met. It is to be a test in which students receive individual scores, to establish how close students are coming to meeting the standards, and how they compare to other students. This proposal taps into a desire on the part of many parents and policy makers for some measures of student performance that are external to the school, a desire that can be met in a variety of ways, and at different levels of government.

At the present time, the *development* of such a test is proceeding, under a $45 million contract with the American Institutes of Research and a host of subcontractors. The contractor is the National Assessment Governing Board (NAGB), the statutory policy board for NAEP. There is little question about the quality of the work which will go into developing such a test; the best people available are involved. The debate concerns the issue of doing it at all, and what benefit to American education will result.

Many statements have been made about what the test would tell us, and what it would do to help education. The discussion outside government revolves around the information the test would make available. It is worth being clear about this, because much of this discussion is not well informed. Even the word "national" means different things to different people.

In fact, we already *have* a national test; it is called the National Assessment of Educational Progress (NAEP), or the Nation's Report Card. Seldom does an editorial or news article commenting on the National Test proposal refer to NAEP, or point out how NAEP and the National Test differ. While the information from NAEP filters through the newspapers, it has not achieved wide identity for a number of reasons. Although people know about the SAT, few would say they know NAEP, although they have probably read the results from NAEP in news stories. We already learn from NAEP a whole lot of what is claimed that a new test will tell us. The results of NAEP tell us:

- *How well students are doing in the U.S. and regions of the U.S. in all major subjects, and how that has changed over time.*

- *How minority students are doing, how students from different socioeconomic classes are doing, how well inner- and outer-city students are doing, and how males and females are doing.*

- *How individual states are doing, and how one state compares with another. Through linking with international assessments, each state can be compared with other countries.*

- *How a school district is doing; the law now permits districts to enter the system, and some have.*

NAEP itself does not extend to the level of the school. But NAEP is constructed so as to *permit* school level assessments if it is desired, as long as schools recognize that the assessment is based on NAEP content frameworks. In NAEP, half the test exercises are released to the public; any capable testing organization can construct a NAEP-linked assessment and render scores on the NAEP achievement scale. This has been done regularly in the schools participating in the High Schools That Work Consortium of the Southern Regional Education Board for the last 10 years, under a contract with Educational Testing Service.

Using these released NAEP items to construct an assessment linked to the NAEP scale is quite feasible now. All it takes is desire and money. It has been done in the 21-state SREB Consortium referred to above. And about 10 years ago it was done throughout Florida.

The defining difference between the National Test now under development and the existing NAEP is that the new test is to be used to provide *individual student scores*. The matter of how these individual scores would be used is unclear, and this might depend on what a school, or a district, or a state might want to use them for. The program is "voluntary," and presumably the testing entity could use them as it pleases. The more the scores come to mean in terms of consequences to the students and to the teachers of those students, the more stringent the criteria must be with respect to the validity and appropriateness of the test for the purposes for which it is used.

The most oft heard example of a good use is that a parent can look at their child's score and know how good an education the school is providing — or perhaps, how well the teacher is teaching. We can compare *student achievement* better by looking at *average* scores for a subject matter in a school, or for a class, which can be done now as described above, without yielding individual student scores. With this information, parents can compare achievement in their school with that of students across the state, or the nation. They already know, from grades and class rankings, how well the student is doing within the school. Comparing achievement is *not* the same thing as comparing the quality of instruction; evaluating teachers and schools is a much more complex matter.

To give an individual student a score raises the bar for judging quality of the testing instrument a whole lot, for consequences for the individual begin to be attached to the score. Even with high quality standards in constructing a test, decisions about individuals should not depend solely on an individual test score.

The reason the NAEP approach will better enable parents to know how achievement in their school compares, is that each student

is not asked to answer the same questions. By using a sampling system in which students answer different questions, and combining test results into a composite score, tests can reveal proficiency across a broad scope of subject matter. The scores can represent the results of several hours of testing instead of the results of a one-hour individual student test. So we can learn very well how a *group* of students is doing in eighth-grade math, rather than how well an individual student does on a relatively small number of questions that can represent only a fraction of the subject matter taught.

It is the limited range of subject matter in an individual student test that makes its use suspect for any important purpose. Lyle Jones in *National Tests and Education Reform: Are They Compatible?* (ETS Policy Information Center, 1997), summarizes the case for viewing such a simple standardized test with care. He quotes Robert Stake, from the University of Illinois:

> *"Mathematics test scores — that do a good job of indicating which students are doing best and which are doing relatively poorly, do not necessarily provide a valid indication of subject-matter mastery. One test alone will not provide valid measurement of the mathematics achievement of individual students or of a group as a whole. Test content is almost always too narrow. Just as … a few books do not represent a library, 20 or 30 test items do not represent the broad range of mathematics skills and knowledge that teachers are teaching. For measurement of subject matter attained, the simplicity of testing is at odds with the complexity of teaching and learning."*

All this is true of testing in a local school district, even where there is an agreement on the content of instruction within the district and a common curriculum. The problem of having a test that measures mastery of a subject area, and enables comparison of scores among students, is greatly magnified when the test is the same but the students are studying different content in differing curricula — which is exactly what a national test does. There is some commonality in American instruction, but there is variation also, and this variation coupled with the pitfalls in the use of a single test, even when there *is* uniformity in instruction, makes the meaning of an individual score in a National Test very problematic.

There is variation and there will be more of it as states and localities struggle with raising achievement. While the National Council of Teachers of Mathematics (NCTM) math standards are used in developing state standards they are not copied *verbatim,* and there are rebellions. California has recently injected more math of the older style, rejecting the NCTM emphasis on so much problem-solving. A study by the American Federation of Teachers found large differences in new state standards, using a number of criteria, as did the Council on Basic Education and the Fordham Foundation. The National Education Goals Panel has compared the state-by-state evaluations made by different organizations. Even within states there will be variation. Fairfax County, VA has recently permitted four schools to switch to the so-called core knowledge curriculum created by

E. D. Hirsch, Jr., which focuses more on basic knowledge and less on "thinking skills," which tend to get heavy emphasis in the emerging content standards (*Washington Post*, March 2, 1998, p. B1).

The Administration intends to link individual scores on the new National Test to the NAEP proficiency scale so that scores on the new test will be comparable to those on NAEP. Given what is known to be possible, this will be at best a very "rough and ready" link. A recent review of the matter by the National Research Council concludes that a linkage of acceptable quality is doubtful.

There is also a characterization of the whole effort as "national standards," as well as a national test. While national "content" standards have been developed in several subjects, with assistance or prodding from both this administration and the prior administration, these standards *describe what students should be taught*. The Administration is referring *not* to these standards, but to the "achievement levels" set for NAEP scores by the National Assessment Governing Board.[3] Through a fairly complex process called the Modified Angoff Method, scores for each of the three grade levels tested by NAEP at the fourth, eighth, and twelfth grades, are set to represent Basic, Proficient, or Advanced levels. NAEP then reports the percentage of students below, at, or above these three levels of student performance.

This process involves a panel of judges who look at the test questions and decide which ones a student has to answer correctly in order to reach one of these three levels. The methods used to do this have been roundly criticized by most of the members of the educational measurement community who have examined the process. NAGB, however, vigorously defends its procedures, while accepting that its judgment is involved. There is no intent here to judge the matter. The Administration proposes linking the scores of the new national voluntary test, so that results determine whether a student is below the Basic level, or at or above any one of these levels: These levels are the "standards" being referred to. Doing so will put a considerable strain on this level-setting process, and, as consequences are attached to these individuals' scores,[4] this process will come under very close scrutiny by many groups beyond educational measurement experts. The question goes not only to the result, but also to the legitimacy of the process by which standards are set.

[3] Such state content standards are most nearly comparable to the NAEP content frameworks that specify the content that is to be assessed, rather than either to NAEP achievement levels, or to "performance standards" generally, discussed later in this report.

[4] The matter of the extent to which "consequences" are to be attached, and to whom, has not been clear; the fact that it is to be voluntary suggests this might vary among those entities that use the process.

An Alternative: National Help For Local Action

The National Test is under development, so whether Congress blocks implementation or not, the resources will be spent on developing test exercises. One constructive approach, I believe, would be to use this pool of items to better enable schools, districts, and states to have a NAEP-linked test they could use to measure achievement for whole districts, and possibly whole schools. The contractors now involved, or other intermediaries, could provide such a service to schools to assist in doing this, providing a "School Assessment and Comparison Service." The state and local assessments such a service would help to create would not be NAEP per se, since that really requires very sophisticated survey research to assure comparability and link scores to a host of student, teachers, and school characteristics. But such services would be a much more constructive use of the $45 million to be expended than to continue the pursuit of individual student scores on a National Test. It would clearly be more sensible than developing this large volume of test items, only to be blocked by Congress from making a National Test operational.

Such a School Assessment and Comparison Service could be a way out of the impasse that has occurred and the polarization in viewpoints that have developed. The service could do things such as the following:

- *Supply NAEP-released items and those newly developed under the National Test contract, to states, districts, and schools, and provide technical assistance for their use in developing and using tests that enable estimation of NAEP scale scores (as is now done by the SREB Schools that Work Consortium).*

- *Help states and localities develop valid statistical links between their own accountability assessments and the NAEP scale showing them the choices they can make in embedding NAEP assessment items, or blocks of items, into their own tests and alignments of various kinds.*

- *Help states and localities make use — if they so choose — of NAEP frameworks and scoring guides, perhaps helping align curriculum frameworks to the NAEP frameworks.*

- *Help get NAEP knowledge and tools down to the school level, through the liaisons the service would develop with those who want to make greater use of NAEP. The National Center for Education Statistics (NCES) competently "reports out" the NAEP results, but does not transform them into a variety of forms and levels of detail for more specific applications.*

The work done by NAGB and NCES on NAEP and in development of a NAEP-linked National Test can be made valuable in improving achievement at all levels, over and above the routine release of NAEP reports. And the new items developed for the National Test can be used at state and local levels for a variety of purposes. Examples of *existing* uses of the present NAEP are instructive:[5]

- *West Virginia is aligning its content standards with NAEP frameworks, according to Henry Marockie, the state superintendent of schools. And the state's own tests included "NAEP-like items."*

- *"Staff members at the North Carolina Department of Public Instruction are studying the feasibility of tying NAEP achievement levels to the levels the state uses to gauge student performances," according to Michael Ward, state superintendent of public instruction.*

- *In New York, BOCES (Board of Cooperative Educational Services) used "NAEP results in designing pre-assessment and post-assessment tasks that we embedded in instruction," according to Phyllis Aldrich, coordinator of gifted education at an upstate BOCES.*

- *In designing a mathematics assessment in Minnesota, a group of teachers doing the design "found NAEP's released mathematics items to be a valuable source of ideas for their work," according to James E. Ellingson, who served on the National Assessment Governing Board from 1995 to 1998.*

For the *proponents* of a National Test, such a School Assessment and Comparison Service could result in NAEP and NAEP-based assessments being used in a wide variety of ways, at district, school, and teacher levels. Also, achievement comparisons could be made with national, regional, and state NAEP results (although not with the rigor of regular NAEP assessments, which involve sophisticated research).

For the *opponents* of the National Test, who object to intrusion in state and local affairs that might come about from such a test, teachers, schools, districts, and states would become *clients* of such a service, getting help in fashioning what *they* want to do, with different patterns emerging.

[5] Culled from *Standards Count*, a volume of papers prepared for the Tenth Anniversary Conference of the National Assessment Governing Board, November 19, 1998.

The Patient Approach: Content Standards and Aligned Assessment

The greatest promise continues to be in intensifying efforts to establish strong standards for the content of instruction, developing curricula reflecting this content, and aligning assessments to the curricula actually being taught.

This approach does require more patience. Both the Clinton and Bush administrations have encouraged such efforts, and both administrations have played a role in encouraging national (*not* federal) content standards. These national standards have led states to develop their own modifications. The math standards led the way, emerging from the work of the National Council of Teachers of Mathematics, begun in the early 1980s; 42 states had content standards in 1998. Science is second, with 41 states, and emerged from the work of the National Science Teachers Association, the American Association for the Advancement of Science, and the National Research Council. There are now 40 states with social studies/history standards; English and Language Arts follow, with 37 states having established standards. About half the states now have standards in foreign languages, health, and physical education.

The Council of Chief State School Officers (CCSSO) reports that these states have "standards ready for implementation." The extent of actual implementation varies widely; such standards mean little until they are translated into curricula.

This standard setting has led to a constructive dialogue in the great majority of states about *what should be taught in the schools, and at what level*. What better place is there to begin a process of reforming the schools and raising achievement? The 1997 review of these developments by the Council of Chief State School Officers summed it up this way:

> *"State initiatives in the 1990s to develop state standards and framework documents differ from earlier state efforts in several ways. First, the pattern across states is widespread involvement of local educators, community leaders, business groups, and political leaders; a dialogue and review concerning what should be taught and learned in mathematics and science.*
>
> *… a second development in the 1990s is active involvement of classroom teachers in writing and editing content standards and frameworks …. A common practice for states in producing standards documents is to convene a large steering committee or task force which represents educators, administrators, subject specialists, and community leaders from across the state …. [The process also] developed new alliances among educators and the public, as they jointly defined the directions for mathematics and science education for children."*

These content standards vary in a number of respects. Some just spell out content. Others go well beyond to give more detailed "benchmarks" concerning what students

should accomplish, describe what is expected of students, give examples of approaches to teachers, give guidance on how to assess students' accomplishments, and also address professional development. And some fall in between. They vary in rigor and quality, and they are often a work in progress. Proposals are also in various stages of implementation, with much to do to develop new curricula and begin professional development of the teachers who have to use them.

A comprehensive review of the state of standards-setting in math and science is included in the CCSSO report of 1997, *Mathematics and Science Content Standards and Curriculum Frameworks, State Progress on Development and Implementation* (now updated in 1998). The American Federation of Teachers, the Council of Basic Education, and the Fordham Foundation have all looked at these standards with a critical eye, and have often reached strikingly different conclusions. For a great many states there is still a long way to go, even in math and science, which are far ahead. But it is the right *direction* to go, and deserves the focused attention of all who want to raise the level of achievement of American students. The path will be difficult to assess more subjects, to develop curriculum and instructional materials, to encourage teacher development and proper assessments, and to establish *performance* standards.

For most states, the alignment of assessments is a big task ahead. By 1998, CCSSO was reporting that almost all the states had some kind of content standards in place. But 29 of those states also reported in 1997 that their assessments were not yet aligned with standards. So, frequently, the system is divided against itself — *new* content standards with *old* tests that do not reflect the new content and the curriculum. What counts for students and schools, still, are the results on the old tests.

One example of what is required is what Pennsylvania is doing, beginning in the fall of 1998, as reported by *Education Daily* (11/2/98). In a move to help teachers align classroom instruction to the standards, state officials have mailed 50,000 resource kits to schools across the state. Developed by more than 100 teachers, the new Classrooms Connection's Resource Kit contains an overview of the standards; assessment tips and instruction strategies; resources for parents; sample lesson plans; and professional development ideas. All this is also available on CD-ROM and, by January 1999, all the materials will be available on the state education department's Web site. What alignment means, however, will vary among the states, depending on how much local variation the state tolerates, and its views concerning desirable levels of decision-making. In general, activity has occurred at the state level. The process must devolve to the community level, and educators in inner cities, who often feel left out of the process, must participate.

A dialogue on what should be taught in school seems healthy. Once "content standards" are established, they mean nothing unless they affect the curriculum that is in use. However, whether these must be statewide standards or localized standards, and to what extent there is benefit in completely standardizing the curriculum are open questions. The

benefit in any particular school or locality depends on the circumstance of its schools, its history, and its current dynamics. In the U.S., the responsibility for education is given to the individual state. Exercising that responsibility, states have varied widely in how much local discretion they have permitted, and how much uniformity they have required.

The Challenge of Setting Performance Standards

Even when assessments reflect content standards, the task of establishing performance standards remains.

States must assess *how much* of that content a student needs to master, and whether an assessment will show that students have learned the content standards. The question becomes: what score is necessary for performance to be judged acceptable, or advanced? Teachers do it by judgment when they assign an A or a C to students who have all studied the same material. Setting these "cut points" on assessments means confronting the wide dispersion of achievement among students in any one grade. A standard the bottom third of students can reasonably be expected to reach under higher content standards will be no incentive for the students higher up the scale. A standard high enough to challenge those up the scale will likely be out of reach for those below, at least given the limitations schools are likely to have in terms of resources.

A set of content standards and a set of test questions intended to reflect that content lead directly to setting performance standards. Yet setting *content* standards has been the work of educators (with the involvement of various publics). Setting *performance* standards on tests has been the work of measurement experts and psychometricians. The bridge between the two has not been constructed. A review of the various means used to set such performance standards was recently provided, in a form for a more general audience, in a 20-page report called *Setting Performance Standards: Contents, Goals, and Individual Differences*, by Bert F. Green of Johns Hopkins University, and published by the ETS Policy Information Center in 1996. He sums up the situation as follows:[6]

> *"The performance standards have to reflect the content standards. The bridge from the content standards to the performance standards depends on the test specifications, the item writers and test editors, and on the resulting performance measurement scale. 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 mainly 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 the bridge from content to performance. This would seem a more straightforward task than imagining the test behavior of marginally competent test-takers ....*

[6] In other countries, there is greater reliance on the judgment of panels created to set standards, and much less use of the psychometric procedures that have developed in the U.S.

In summary, the psychometric problem of determining just where a cut-point should be placed on a scale seems not to be a central feature of standard setting …. And finally, finding a way to map content standards onto performance standards is a challenge.

Beyond the performance standards reflecting the content standards as discussed above, there is the issue that the ***form*** *of the assessment tasks be appropriate to the standards. For example, if a standard calls for the student "to know" something, a short constructed-response item might be appropriate, but more might be required if the standard calls for a student to "be able to analyze the results of …."*

CCSSO reported that in November of 1998, 21 states had established performance standards that met the review criteria established by the U.S. Department of Education. This means those states went through the prescribed steps, but the Department has not ruled on the quality of the work, or the appropriateness of the cut points set by these states.

We are speaking of a challenge in setting cut points on a standardized instrument used for large-scale assessment, used for accountability, or possibly for promotion or graduation. At the *classroom* level these test results are not determinants of *teachers'* judgments of student performance. Once content standards have evolved into curriculum, and into pedagogical approaches, *teachers* will be the judges in the classroom. They give the tests and assign the grades. They will do it as professionals, not as psychometricians using statistical methodologies. (At the end of this report we say more about the critical role that teachers play, and the need to help them use assessment in service of learning.)

Here then is the situation we find ourselves in at the end of about two decades of education reform. Most states have content standards established in at least some subjects. A minority of these have assessments that they say are aligned to these standards; and only 11 states have trend data on student achievement for two or three years. In some key subjects, just half the states have content standards. Where performance standards have been established, we do not know how directly the standards are linked to the content standards, and whether or how these states overcame the challenges Green says they face.

The whole content-assessment-performance approach is incomplete, and to the extent that this approach is the linchpin of "educational reform," we don't have it adequately in place as we approach the year 2000. But steady progress is being made.

Accountability, But For The Right Things

If the standardized tests used for school, district, and state accountability were switched from the intrusive testing of every student to sample-based assessments, and assessments were aligned to content standards, would we be on the right track in standardized testing for accountability?

No, there would still be some work to do. In many respects, standardized testing is at its zenith, and reaches elegance in such things as its refined principles, standards of validity and reliability, latent trait analysis, equating, and techniques of spotting biased test items.

But the way tests are used *in practice* in elementary and secondary education — of rewarding and punishing schools, closing schools, and judging educational progress — is often appallingly primitive. Frequently,

- *Commercial standardized tests are used that measure a blend of what is being taught across the nation — not what is taught in a school or district (and not what is* ***supposed*** *to be taught).*[7]

- *The test content changes from time to time to reflect changing views of what should be taught. Yet the scores from year to year are used to judge whether progress is being made.*

- *In many cases, norm-referenced tests designed to show how one school's students compare with those in the entire nation are used to track change in the school's performance over time, a task they are not designed to do.*

- *While the tests are presumed to judge the quality of what the school does, a large part of an individual's score is attributable to family background and opportunities before school and outside the classroom. Current tests that measure both the quality of current in-school instruction* ***and*** *out-of-school development are used to unfairly reward or punish schools, or close them down entirely.*

- *While tests are presumably used to determine how well the school instructs from the beginning of one grade to the beginning of the next grade, the tests actually determine the cumulative level of knowledge of eighth graders, for example —* ***not what knowledge was added during the eighth grade.*** *It is rare to have a measure of "value added," a measure of the change in the levels of knowledge between two points in time.*[8]

[7] Testing for what is actually taught has been given extended discussion and debate under the rubric of "opportunity to learn," and encompasses resource levels and adequacies as well as curriculum content and instruction.

[8] Robert Stake points out the challenge to value added measures from the fact that single scores are more reliable than the change in scores from, say, one year to the next (personal correspondence).

This summary covers a very wide territory. Each point deserves elaboration. A number of scholars have examined these matters and the impact such practices have on instruction and student achievement.[9]

Measuring and comparing what students have learned in school in a given time period is quite different from measuring and comparing the total of what they know. One early recognition of the difference was reflected in the 1984 South Carolina Education Improvement Act, a broad measure to improve schools in the state. It called for a number of measurement approaches to reward and penalize schools; two are described here.[10]

First, the act dealt with the different levels of students' socioeconomic backgrounds by grouping the state's schools into five comparison groups based on certain context variables. These included the percentage of free-lunch eligible students and, for elementary schools, the percentage of first-grade students meeting the state readiness standards. Schools within each of the five groups were compared on achievement results.

Second, it dealt with the matter of how much is learned *within* a school year, as compared to total knowledge accumulated. Kaagan and Coley describe it this way:

> *"… The report cards present a matched longitudinal analysis of reading and mathematics test scores for the two most recent test administrations. Put simply, this procedure allows the calculation of score gains (or losses) of the* ***same students from one year to the next*** *[emphasis supplied]."*

Thus school accomplishments were not to be judged simply in terms of background that students brought to school with them; nor teachers in terms of what students had already been taught (or not taught) when they entered their classrooms. Instead, students would be judged on what they had learned *in* the classroom. This was a huge departure in the use of standardized testing as it had developed in the 1970s and 80s. While other states have used regression approaches to sort out school and non-school accomplishments, they have not used *gains* in scores as the measure of achievement.

For the nation, regions, and for state data on a comparable basis, we have relied on the reports of the National Assessment of Educational Progress (NAEP). For the nation and regions NAEP has been providing a continuous record of school achievement for almost three decades, and more recently has provided a record for states that have

[9] A recent survey of such work is "the Political Legacy of School Accountability Systems," by Sherman Dorn of the University of South Florida (In *Education Policy Analysis Archives, 6,* (1), January 2, 1998). Also, see "The Adverse Impact of High Stakes Testing on Minority Students: Evidence From 100 Years of Test Data," by George Madaus and Marguerite Clarke, at the National Board on Educational Testing and Public Policy at Boston College, December, 1998. Also see Robert Stake, "The Invalidity of Standardized Testing for Measuring Mathematics Achievement" in Thomas A. Romberg, Editor, *Reform in School Mathematics and Authentic Assessment*. Albany: SUNY Press, 1995.

[10] The South Carolina indicator system is described in *State Education Indicators: Measured Strides, Missing Steps*, by Steven Kaagan and Richard Coley, published by Rutgers University and Educational Testing Service in 1989.

participated in the program. These reports have all been about levels of achievement at ages 9, 13, and 17 or grades 4, 8, and 12. Thus, we can compare the scores in mathematics for students in grade 4 in 1996 with scores of fourth-graders in earlier years. Again, when we look at trends in these scores of fourth-graders, we know whether they now know more. We can't tell whether it is because they were better developed by the time they were in the first grade, had learned more in grades 1 through 3, or had learned more in grade 4 — the year in which they were being tested. Have the schools performed better? Or is it the family? If it is the schools, was the change due to better teaching in the second grade? Or the fourth grade? Or both? Change over time may be influenced by any one of these, or by a combination of factors.

A redesign of NAEP in the early 1980s led to a provision for tracking a *cohort* of the same students, in addition to measuring the level of fourth graders at a given time, compared to some previous time. The data has been examined from this standpoint; the ETS Policy Information Center published a report in 1998 describing achievement in these terms of "value added" (*Growth in School: Achievement Gains from the Fourth to the Eighth Grade,* by Paul E. Barton and Richard J. Coley).

What emerged was quite a different picture from that given by the NAEP reports based on the levels of student knowledge in a particular grade (or at a particular age), compared with the levels of their counterparts in earlier years. The report explained it this way:

> *While in most cases the average NAEP scores of today's students are slightly higher than those of students 20 or 25 years ago, the cohort growth between the fourth and the eighth grade is not. In fact, cohort growth is the same as, or lower than, it was during the earliest period for which we have data.*
>
> *And when we compare states, there is little difference in the cohort growth between the fourth and eighth grade. While Maine was the top-scoring state in the nation[11] and Arkansas was the bottom-scoring state, both states had the same cohort growth, 52 points on the NAEP scale (in mathematics) between the fourth and eighth grade.*
>
> *How do we, and how should we, look at NAEP scores in reaching a judgment as to whether the education system is performing better or worse over time? Are Maine and Arkansas at the two ends of the school quality continuum, or are they actually equal?*

[11] Among the 37 states participating in NAEP in both 1992 and 1996.

The comparison of trends in cohort growth and averages at a particular grade is shown in Table 1. The Maine/Arkansas comparison is shown in Figure 1.

The *Growth in School* report urged that we be able to measure both changes in the levels of same grade student knowledge, *and* changes in the knowledge of the same students between two points in time. And we asked whether standards should be set for both kinds of change, if we are to have a standards-based assessment system.

From NAEP, to state, to district, to school standardized testing, it is *levels* of achievement that are measured — not growth in what students know and can do. The exception of South Carolina in the early 1980s was noted above. Also, since 1992, Tennessee has used the Value-Added Assessment System. Recently, Memphis City Schools used this assessment (TCAP) to compare student achievement gains in 25 elementary schools that began implementing national school redesign models in 1995-96 with a comparable group of schools that were not redesigned. The comparison measured year-to-year *gains* in achievement, and redesigned schools showed greater gains. Chicago has also created a system that enables judging schools on this basis, even though the testing system itself was not designed for this use.

Chicago's changes were set in motion by the 1988 Chicago School Reform Act. That Act decentralized control to the individual school level, and created a need to examine resulting improvement in achievement. The test used then was the Iowa Test of Basic Skills (ITBS). Because of the use of different forms of the test, changes in the content from time to time, and its norm-referenced characteristics, ITBS was not an accurate measure of trends in achievement over time for individual schools, or even for levels of achievement at individual grades. The Consortium on Chicago School Research, working under the Chicago Panel on School Policy, has spent years creating a system to measure the productivity of individual schools, and is now using it to do so.

Researchers equated the different forms of the ITBS, the different tests used in different years, and the tests used at different grade levels. This enabled them to place all students who took the test on the same achievement scale. They called this scale the "Measurement Ruler." Test questions are placed at different intervals along this ruler, to illustrate the level of difficulty. The result is a developmental scale similar to what NAEP uses. The difference is that NAEP assessments are designed to enable creation of a scale on which students in all three grades assessed can be arrayed. Making such a scale out of the ITBS norm-referenced test created a far-from-perfect result. Its creators (and others) point out that the data limitations are considerable, and call for a better measure of achievement. This approach is described here because of the principle it has put into operation in evaluating schools in Chicago, despite the handicaps of the tests that are now available.

What the Consortium has created is called a "Grade Productivity Profile." The consortium describes it as follows:

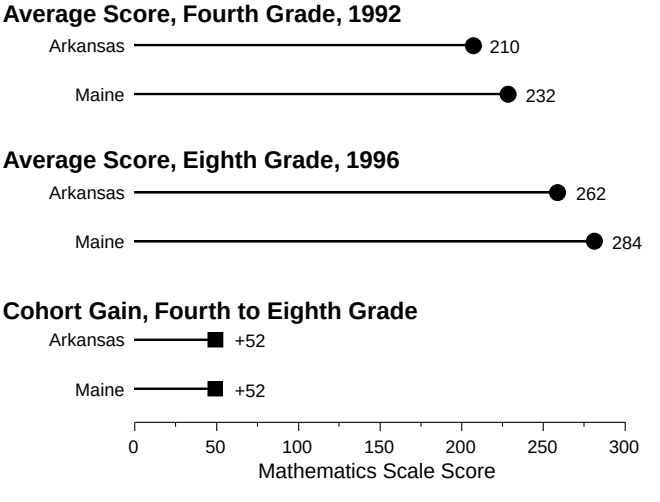
Table 1: Trends in Cohort Growth Compared to Average Score Trends for 9- and 13-year-olds*			
	Cohort Growth, Age 9 to 13	Average Score Trend, Age 9	Average Score Trend, Age 13
Science	Level	Up	Up
Mathematics	Down	Up	Up
Reading	Level	Up	Up
Writing**	Level	Level	Level

Source: National Assessment of Educational Progress data analyzed by the ETS Policy Information Center. See http://nces.ed.gov/naep. "False Discovery Rate" procedure used to test for significance.

* Science cohort changes are from 1973-77 to 1992-96. Average science score trends are from 1973 to 1996. Mathematics cohort changes are from 1973-77 to 1992-96. Average mathematics score trends are from 1973 to 1996. Reading cohort changes are from 1971-75 to 1992-96. Average reading score trends are from 1971 to 1996. Writing cohort changes are from 1984-88 to 1992-96. Average writing score trends are from 1984 to 1996.

** Writing was administered to fourth- and eighth-graders.

Figure 1: Average NAEP Mathematics Scores and Cohort Growth, Arkansas and Maine



Source: National Assessment of Educational Progress data analyzed by the ETS Policy Information Center. See http://nces.ed.gov/naep.

"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 ITBS test scores from the previous spring …*

… As for the ***learning gain*** *for each school grade, this is simply how much the end-of-year ITBS results have improved over the input status for this same group of students."*

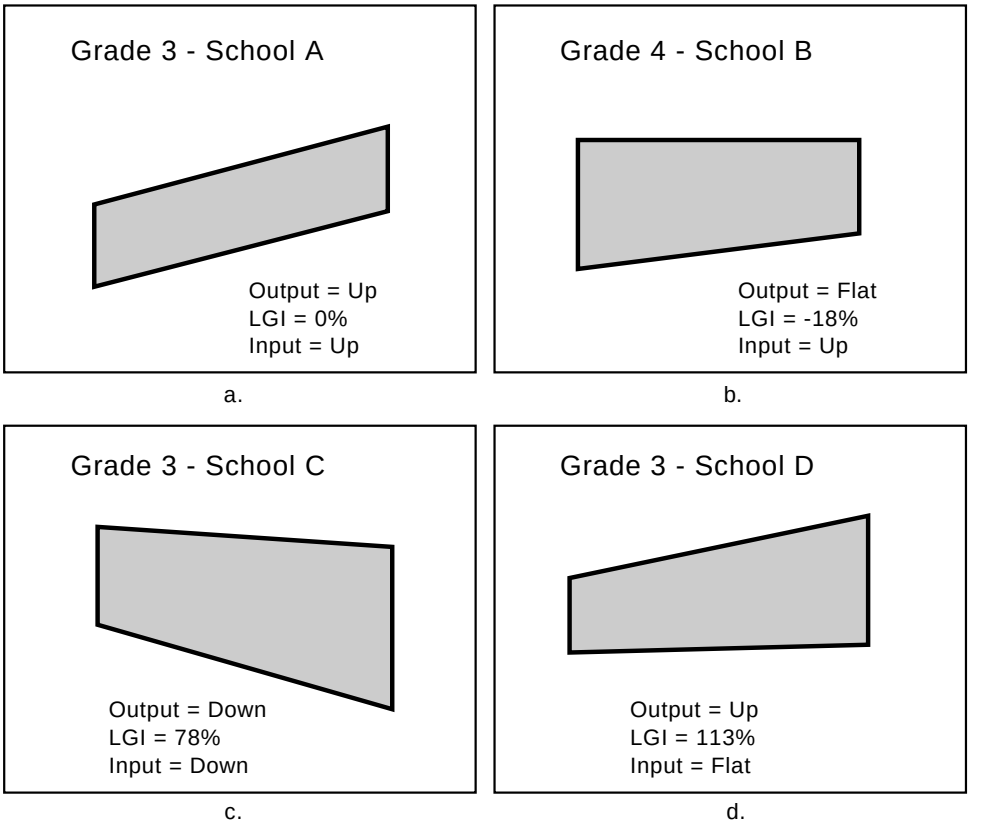
The principle operational meaning here is that: "A school should be held responsible for the learning that occurs among students actually taught in the school." In Figure 2 (see p. 27), examples of grade productivity profiles are displayed, using the Learning Gain Index (LGI). A school with its output up may have an LGI of 0%, because the input was up by an equal amount (School A). A school with its output down had a positive LGI, because its inputs had dropped more than the output (School B). Other combinations are also shown.[12] The experience could be instructive for others wishing to measure school productivity using assessments designed for this purpose.

What all three of the efforts described above have in common is a learning gain measure between two points in time for the same students (or the same cohort of students).[13] These are exceptions in the vast day-to-day enterprise in using standardized assessments to hold schools and teachers accountable. A related development is the work of David Grissmer of RAND, analyzing NAEP data. He distinguishes the changes due to the status of the family and non-school influences, and the change that results directly from schooling.

[12] A full description of all this can be found in *Academic Productivity of Chicago Public Elementary Schools*, by Anthony S. Bryk, Yeow Meng Thum, John Q. Easton, and Stuart Luppescu, Consortium on Chicago School Research, March, 1998.

[13] While I am advocating use of such measures of gain, I recognize that this measurement approach has its own set of measurement challenges. For example, Robert Linn points out to me that in South Carolina and Tennessee the frequent testing required resulted in using simple multiple-choice testing (personal correspondence).

Figure 2. Grade Productivity Profiles



Note: LGI = Learning Gain Index, computed for 1992-1996.

Source: Academic Productivity of Chicago Public Elementary Schools, by Anthony S. Bryk, et al., Consortium on Chicago School Research, March, 1998.

Exit Examinations?

A discussion entitled "Too Much Testing" would be incomplete without pointing out that there is one area where there is very little testing whatsoever in the U.S., and there is a lot in other developed countries.

Most countries require extensive examinations at the exit point for secondary education. The U.S. does not have high school exit tests, except for students taking Regents courses in New York State. (We are not speaking of basic skills tests students must pass to graduate, requiring minimal abilities in reading and mathematics.)

In 1993, a very comprehensive study was published of the exit examination approaches of the U.S., China, Japan, Germany, England and Wales, France, Sweden, and the former Soviet Union.[14] The contrast is stark. Eckstein and Noah put it this way: "The United States is unique among the countries we have studied in having no coordinated, public, national system for assessing student achievement at the end of secondary school." (p. 238)

The examinations in these countries are very closely related to the curriculum. In the U.S., it is hard to conceive of any *national* exam being closely related to the actual curriculum for such a high stakes examination, because of the decentralized control over the curriculum. Eckstein and Noah observe that "governmental control of the school curriculum in the United States and England and Wales has been extraordinarily weak, sometimes even absent," and further that "Decentralization of school control has been even greater in the United States than in England/Wales."

While a centralized exit examination system in the U.S. may be out of the question given the decentralized control and resource allocation decisions, that does not limit the introduction of decentralized exit examination systems. Eckstein and Noah conclude with an examination of "The Persistent Dilemmas of Examination Policy;" and try to answer the question "How can the United States secure [the] advantages, while avoiding, or at least minimizing, the disadvantages that may accompany them?"

A more recent look at the international scene was reported in 1997 in *International Comparisons of Entrance and Exit Examinations*, by Harold W. Stevenson and Shin-ying Lee, in collaboration with five of their colleagues at the University of Michigan. From their study of Japan, the United Kingdom, France, and Germany they observe:

> *"Entrance and exit examinations in these countries are based on a curriculum established by ministries of education at the local, regional, or national level. Rather than imposing some arbitrarily defined standard of achievement, the examinations*

[14] The book is entitled *Secondary School Examinations: International Perspectives on Policies and Practices*, by Max A. Eckstein and Harold J. Noah (published by the Yale University Press). I recommend it for anyone wanting an understanding of practice abroad and how it contrasts to practice in the U.S.

are closely tied to what the students have studied in high school. Because teachers are aware of what students are expected to know in examinations, it becomes their responsibility to equip students with the information and skills needed to pass the examination." (p. 47)

And on the nature of the examinations themselves:

"These examinations typically include open-ended questions that require organization and application of knowledge, and oral examinations that require students to express themselves verbally" (p. 47)

This is all quite different from the many tests we have — tests that are cheap, all or heavily multiple-choice; used to establish how much students *don't* know and *haven't* been taught; and used to grade teachers and schools rather than as a constructive tool of instruction.

But we do have some experience with tests that are designed to reflect curriculum. The Advanced Placement examinations do that and represent an external verification of whether standards were met. And we do, at present, have one set of examinations with similarities — the Regents Examinations in New York State. The New York Regents are low- to medium-stakes tests taken in different subjects in Regents courses at different high school grades, at the student's discretion. Regents courses or Regents tests are not required to get a high school diploma (although that is changing).[15] And the results of the tests are only a fraction of what determines a grade. The results of Regents tests, in terms of their effect on achievement, have been investigated by John Bishop and reported in a monograph.[16]

New York State is now phasing in a requirement that *all* student pass Regents examinations in five core subject areas, to be fully effective with students graduating in 2003. This transforms these tests into high-stakes tests with widespread ramifications. John Bishop's analysis leads him to conclude that:

"Requiring that all students reach the Regents standard in 5 Core Subjects will significantly increase student achievement, college attendance and completion, and the quality of jobs that students get after high school. The biggest beneficiaries of the policy will be the students, often from disadvantaged backgrounds, who have been allowed to avoid rigorous courses in the past. In the All-Regents high schools,[17] there was a massive reallocation of teacher time and resources toward struggling students. It was these students whose achievement rose the most. Their probability of going to and completing college rose significantly." (p. 4)

[15] But passing has been required for a "Regents Diploma." This is a very old system and practices have varied.

[16] *Diplomas For Learning, Not Seat Time: The Impacts of New York Regents Examinations*, published by the Cornell University Center for Advanced Human Resources (with Joan Moriarty and Ferran Mane).

[17] Ten schools have already moved to all Regents courses.

John Bishop and his colleagues offer a number of ways to avoid adverse effects.

Maryland also is now in the process of installing high school exit tests. Of course, not all will agree with the New York or Maryland claims for their approach, or with John Bishop's conclusions, but, nobody agrees on much of anything in American education. For example, concerns have been raised about the effect on lower-achieving students, and whether all students will have the opportunity to learn what is required in the examinations. John F. Jennings, director of the Center on Education Policy, voiced this concern strongly in his article in *Education Week* entitled "Opportunity to Learn or Lose?", referring to the general movement to raise standards in the schools. Students must have an opportunity to learn, and students who are not challenged by high standards and expectations will be shortchanged by the school they attend.

Obviously, I have not stated a clear position on the value of exit examinations, yet I do think the matter is worth attention and examination. Why is the U.S. unique in not having such examinations? What can we learn from the experience of other countries? Such examinations can take many forms, and can even be created by the teachers in an individual school. Unfortunately, there are also opportunities for a new arena for the misuse of testing, of the kind described in this report.

My criticisms of massive testing in the U.S. are not based on the philosophical debate about local control vs. higher control. Above, I have argued on other bases, and I argue for higher standards. It would be consistent with those arguments to have rigorous exit examinations.[18] But examinations must be formulated at the same level as the curriculum, and must involve teachers. They must cover the full curriculum studied. They should be only one factor in deciding whether a student graduates, rather than the sole factor.

[18] A similar set of issues and concerns arises in the use of standardized tests to determine promotion from one grade to the other, and I have not examined current experience and the evaluation of it for this paper. This use has been advanced to put an end to "social promotion." But good teacher assessment and grading can do that, and a single test should not be the sole basis for a promotion.

It Comes Back to Teachers

This report began with a discussion of the current excess of standardized testing.

While we need to complete the content-assessment-performance triad, we do not need this ever-larger volume of standardized testing of individual students to render individual scores. Aligned assessments can examine whether educational achievement is progressing, and for what kinds of students. Teachers should be the judges of performance, give out the grades, and pass or fail students. Aligned standardized instruments can be used on a sampling basis, or without assigning individual scores, for school accountability purposes and tracking achievement changes, as they have been in the past.[19]

This position will leave a lot of people concerned that while testing and grading is left up to the teachers, they have not been well prepared to conduct quality assessments. They are taught little about day-to-day classroom assessment approaches in school. Nor is much professional development offered. Assessment is part of teaching and instruction, and teachers must learn to adequately assess students. Given continued emphasis on standardized testing to hold teachers and schools accountable, the alternative of equipping teachers to do their jobs will continue to be neglected. Teachers and teaching need help. We can have external verification of how well the students in a class or school are doing through sample-based standardized assessments that are properly designed and aligned.

We are in danger of focusing too much on highly structured systems — largely for outside control — and not on the teacher as a professional. We give doctors the professional competencies to treat patients; all patients with infections are not given standardized examinations by third parties to see at what rate their infections receded. It is, and will remain, the teacher who delivers the "content," who aligns his or her assessment methods to this content, and who judges performance. The elevation of the teaching practice to a teaching profession that has our confidence cannot be avoided through these formal exercises taking place outside the classroom, as important as they may be when properly used. If we examine this problem realistically, for all the rhetoric and activity of the 1990s, we have not begun to remake the profession. A reading or re-reading of John Goodlad's *Teachers for Our Nation's Schools* would be a good place to start.

There are many today who believe that American education was better, 30, 40, or 50 years ago. Some have pointed to McGuffey's readers and made comparisons

[19] If tests are used to judge teachers and schools, they should measure *gains* in achievement, not just levels of knowledge.

with Dick and Jane. People remember demanding teachers who took no nonsense in the classroom, and meted out punishment surely and swiftly. Examples of outstanding teaching abound, such as the one-room school in Kentucky, that the writer Jesse Stuart described in *The Thread That Runs So True*. He describes how he taught his charges, who won a contest in the city schools. If there *was* such superior teaching and learning in the old days, it was done without the standardized testing we know today. That is something worth thinking about.

We can move toward more professionalism in teaching and toward respecting the judgments teachers make about their students' learning. At the same time, we can move toward "less frequent but far better testing," in the words of Albert Shanker's report in 1993. Shanker was a proponent of good testing with consequences to the student and to schools. Americans must demand higher standards in testing, as they are demanding higher standards in education generally. Standardized testing, used properly, may tell us whether the standards-based reforms are working. In and of itself, testing is not the treatment.

* * * *

There are worrisome trends in the American testing enterprise. Standardized testing has produced more and more numbers, and has fed a quantitative approach to managing the education system. But we are short-changed in terms of the *information* that we are getting to help teachers and schools improve student performance. At the same time though, there are some hopeful signs that the situation will improve. And there are prospects for harnessing assessment in the service of learning if we are willing to face squarely the situation we have created.

15992-04202 • S49M6 • Printed in U.S.A.
I.N. 204928