TESTS, TESTS, TESTS

BY PAUL E. BARTON

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, 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 massive testing, who include 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.

Although there have been constructive attempts to improve the testing enterprise in the 1990s, most of the testing today is not much changed from what it was a dozen years ago. It is important that these improvements be made because testing has become, over the past twenty-five years, the approach of first resort of policymakers. Robert Linn, a scholar of testing, identifies several reasons for the attractiveness of testing:

1. Tests are relatively inexpensive, especially when you compare them with other more costly changes like increasing class time, decreasing class size, or providing substantial professional development.
2. Tests can be externally mandated by states or districts; it is very difficult to mandate anything that involves change inside the classroom.
3. Tests can be rapidly implemented, even within the term of elected officials.
4. Test results are visible. They can be reported to the press. Poor results in the beginning are desirable for policymakers who want to show that they have had an effect.[1]

Exposing the existence of substandard education has long been the objective of written examinations, but 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.[2] A sustained investigation revealed that students' scores almost everywhere were above average, a phenomenon that came to be dubbed the Lake Wobegon effect. Robert Linn, who studied the Lake Wobegon effect, summarized his conclusions in this way:

> There are many reasons for the Lake Wobegon effect... among [them] 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 norm-

ILLUSTRATED BY SONJA LAMUT

Paul E. Barton is director of the Educational Testing Service's Policy Information Center. This article is drawn from a report entitled "Too Much Testing of the Wrong Kind: Too Little of the Right Kind in K-12 Education." Copies of the full report are available online (http://www.ets.org/research/pic/testing/tmt.html) or for $9.50 (prepaid) from the Policy Information Center, ETS, Rosedale Road, Mail Stop 04-R, Princeton, NJ 08541-0001 (609/734-5694).


IGAP
MSPAP
MEA
ITBS
CASE
TAAS
CSAP
STAR
FCAT
SAT
SAT-8
TIMSS
SAT-9
NAEP
SOL
TCAP
CMT
MAT-7
MEAP
HSPT
MCAS
ACT

ing studies, and the narrow focusing of instruction on the skills and question types used on the test.[5]

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

Promising Trends

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

We can change the way we administer standardized tests for school/teacher control and accountability, with much less intrusion into the classroom. The National Assessment of Educational Progress (NAEP) provides a proven means of giving a test to a *sample* of students rather than testing *all* students. NAEP is mandated by Congress and administered by the National Center for Education Statistics, with the purpose of finding out what fourth, eighth, and twelfth grade students know and are able to do. Sample-based approaches will provide *better* information about schools, will be much less intrusive into instructional settings, and will require less frequent testing. If the objective is a report card on the schools, testing every couple of years will accomplish the purpose. 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.

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 determine 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 the policymakers who ordered their use intended? 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.

Aligning Standards and Assessments

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. Both the Clinton and Bush administrations have encouraged such efforts, and both 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; forty-two states had content standards in 1998. Science is second, with forty-one 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 forty states with social studies/history standards; English and language arts follow, with thirty-seven states having established standards. About half the states now have standards in foreign languages, health, and physical education.

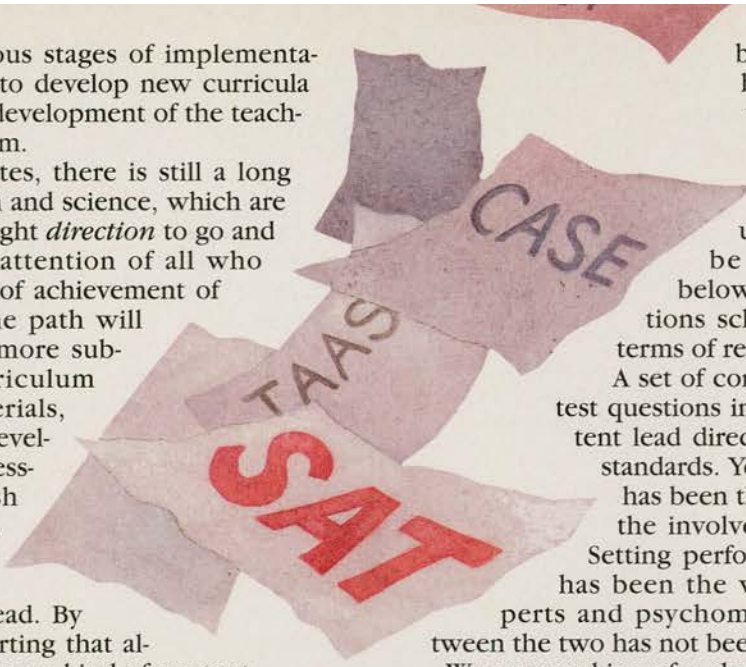
The Council of Chief State School Officers (CCSSO) reports that these states have "standards ready for implementation." The extent of actual implementation varies widely; such standards mean little until they are translated into curricula. This standard-setting has led to a constructive dialogue in the great majority of states about what should be taught in the schools, and at what level. 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. Pro-

posals 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.

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 twenty-nine 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* (Nov. 2, 1998). 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 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.

Setting Performance Standards

Even when assessment standards 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 that 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 probably 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.

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.

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 eleven 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 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—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.

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).

- 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.

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. It called for a number of measurement approaches to reward and penalize schools; two are described here.[4]

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:

> 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].[5]

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 1980s.

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 has been providing a continuous record of school achievement, for the nation and regions, for almost three decades, and more recently it has provided a record for states that have 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. 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. A 1998 report from the Educational Testing Service (ETS) explained it this way:

> While in most cases the average NAEP scores of today's students are slightly higher than those of students twenty or twenty-five 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 and Arkansas was the bottom-scoring state, both states had the same cohort growth, fifty-two 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?[6]

The comparison of trends in cohort growth and averages at a particular grade is shown in the accompanying table. The Maine/Arkansas comparison is shown in the figure. The ETS report urged that we be able both to measure changes in the levels of student knowledge in the same grade and changes in the knowledge of the same students between two points in time. The report also asked whether standards should be set for both kinds of change, if we are to

Table: 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.

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 value-added—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 to compare student achievement gains in twenty-five 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. And Chicago has developed what is called a "grade productivity profile" that enables judging schools on this basis, even though the testing system itself was not designed for this use.[7]

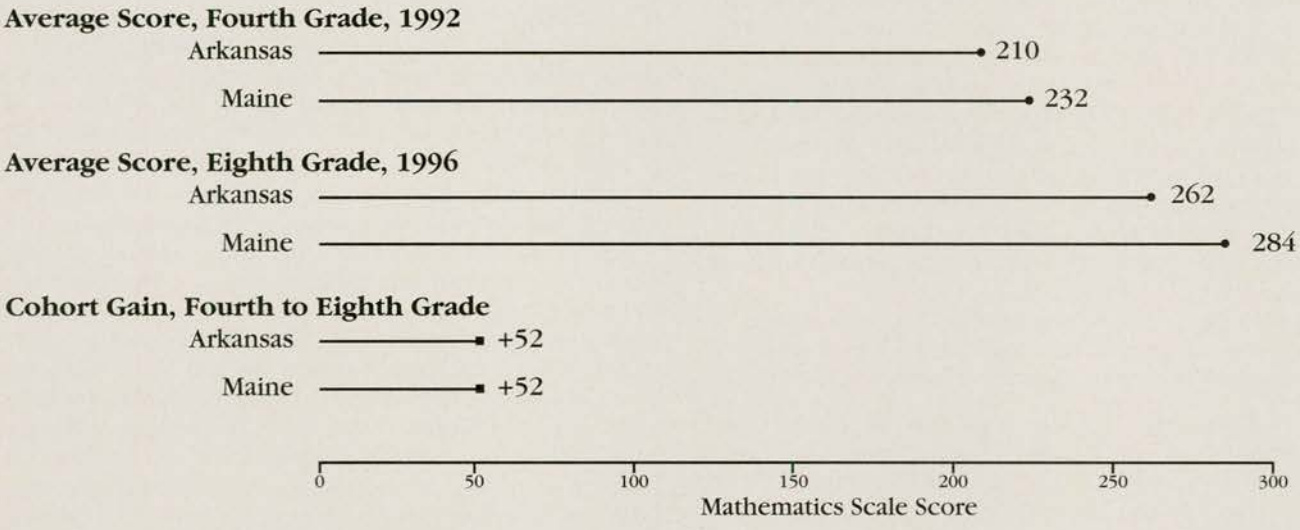
What all three of the efforts described above have in common is a measure of learning gain between two points in time for the same students (or the same cohort of students). These are exceptions in the vast day-to-day enterprise in using standardized assessments to hold schools and teachers accountable.

It Comes Back to Teachers

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

(Continued on page 44)

Figure: Average NAEP Mathematics Scores and Cohort Growth for Arkansas and Maine



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