CHAPTER 3

Educational Testing Policy: The Changing Federal Role

Contents

Table

CHAPTER 3

Educational Testing Policy: The Changing Federal Role

Highlights

- As the Federal financial commitment to education expanded during the 1960s and 1970s, new demands for test-based accountability emerged. Federal policymakers now rely on standardized tests to assess the effectiveness of several Federal programs.
- Evaluation requirements in the Federal Chapter 1 program for disadvantaged children, which result in more than 1.5 million children being tested every year, have helped escalate the amount of testing in American schools. Questions arise about whether results of Chapter 1 testing produce an accurate picture of the program's effectiveness, about the burden that the testing creates for schools, teachers, and children, and about the usefulness of the information provided by the test results.
- The National Assessment of Educational Progress (NAEP) is a unique Federal effort begun in the 1960s to provide long-term and continuous data on the achievement of American school children in many different subjects. NAEP has become a well-respected instrument to help gauge the Nation's educational health. Recent proposals to change NAEP to allow for comparisons in performance between States, to establish proficiency standards, or to use NAEP items as a basis for a system of national examinations raise questions about how much NAEP can be changed without compromising its original purposes.
- National testing is a critical issue before Congress today. Many questions remain about the objectives, content, format, cost, and administration of any national test.

The role of the Federal Government in educational testing policy has been limited but influential. Given the decentralized structure of American schooling, few decisions supported with test information are made at the Federal level. States and local school districts make most of the decisions about which tests to give, when to give them, and how to use the information. The Federal Government weighs in primarily by requiring test-based measures of effectiveness for some of the education programs it funds, operating its own testing program through the National Assessment of Educational Progress (NAEP), and affording some limited protections and rights to test takers and their parents (see ch. 2).

This circumscribed Federal role has nevertheless influenced the quantity and character of testing in American schools. As Federal funding has expanded over the past 25 years, so has the Federal appetite for test-based evidence aimed at ensuring accountability for those funds. This growth in Federal influence has evolved with no specific and deliberate Federal policy on testing. Most Federal decisions about testing have been made in the context of larger program reauthorization bills, with evaluation questions treated as program issues rather than testing policy issues. As discussed in the preceding chapter, Congress did consider several bills in the 1970s and 1980s related to test disclosure and the rights of test takers; only the Family Education Rights and Privacy Act of 1974 became law.

This picture is changing. Congress now faces several critical choices that could redefine the Federal role in educational testing. In three policy areas, Congress has already played an important role, and its decisions in the near term could have significant consequences for the quantity and quality of educational testing. Accountability for federally funded programs is the frost area. The tradition of achievement testing as a way to hold State- or district-level education authorities accountable is as old as public schooling itself. Continued spending on compensatory education has become increasingly dependent on evidence that these programs are working. Thus, for several decades now the single largest Federal education program--Chapter 1 (Compensatory Education)--has struggled with the need for evaluation data from States and districts that receive Federal monies. Increasing reliance on standardized norm-referenced achievement tests to

monitor Chapter 1 programs indicates an increasing Federal influence on the nature and quantity of testing. Congress has revised its accountability requirements on several occasions, and in today's atmosphere of test reform, the $6 billion Federal Chapter 1 program can hardly be ignored. The basic policy question is whether the Federal Government is well served by the information derived from the tests used today and whether modifications could provide improved information.

Second, Federal support for collection of educational data, traditionally intended to keep the Nation informed about overall educational progress, is now viewed by some as a lever to influence teaching and learning. Thus, the 20-year-old NAEP, widely acclaimed as an invaluable instrument to gauge the Nation's educational health, has, in the past few years, attracted the attention of some policymakers interested in using its tests to change the structure and content of schooling.

A third and related issue is national testing, In addition to various suggested changes to NAEP, a number of proposals have emerged recently– -from the White House, various agencies of the executive branch, and blue ribbon commissions-to implement nationwide tests. Although the purposes of these tests vary, it is clear they are intended to bring about improved achievement, not simply to estimate current levels of learning. The idea of national testing seems to have gained greater public acceptability. Proponents argue that "national" does not equal '' Federal," and that national education standards do not require Federal determination of curricula and design of tests. Others fear that national testing will lead inevitably to Federal control of education.

OTA analyzed the development and effects of the current Federal role in testing and examined pending proposals to change that role. This chapter discusses OTA's findings vis-á-vis Chapter 1, NAEP, and national testing.

Chapter 1, Elementary and Secondary Education Act: A Lever on Testing

The passage of the 1965 Federal Elementary and Secondary Education Act (ESEA) heralded a new era of broad-scale Federal involvement in education and established the principle that with Federal education funding comes Federal strings. The cornerstone of ESEA was Title I (renamed Chapter 1 in 1981), which is still the largest program of Federal aid to elementary and secondary schools.[1] The purpose of Title I/Chapter 1, both then and now, is to provide supplementary educational services, primarily in reading and mathematics, to low-achieving children living in poor neighborhoods. With an appropriation of $6.2 billion for fiscal year 1991,[2] Chapter 1 channels funds to almost every school district in the country. Some 51,000 schools, including over 75 percent of the Nation's elementary schools, receive Chapter 1 dollars, which are used to fund services to about 5 million children in preschool through grade 12. Given its 25-year history and broad reach, the effect of Chapter 1 on Federal testing policy is profound.

History of Chapter 1 Evaluation

From the beginning, the Title I/Chapter 1 law required participating school districts to periodically evaluate the effectiveness of the program in meeting the special educational needs of educationally disadvantaged children, using ''. . . appropriate objective measures of educational achievement' '3-interpreted to mean norm-referenced standardized tests. Congress has revised the evaluation requirements many times to reflect changing Federal priorities and address new State and local concerns.

During the 1960s and 1970s, the Title I evaluation provisions generally became more prescriptive and detailed. In 1981, a dramatic loosening of Federal requirements occurred: while evaluations were still required, Federal standards governing the format, frequency, and content of evaluations were deleted. In the absence of Federal guidance, confusion about just what was required ensued at the State and local

[1]The remainder of this section is from Nancy Kober, ''The Role and Impact of Chapter 1 Evaluation and Assessment Requirements, ' OTA contractor report, May 1991.

[2]Of this $6.2 billion, approximately $5.5 billion is distributed by formula to local school districts. The remainder is used for three State-administered programs for migrant students, students with disabilities, neglected and delinquent children, and for other specialized programs and activities, such as State administration and technical assistance.

[3]Public Law 89-10.

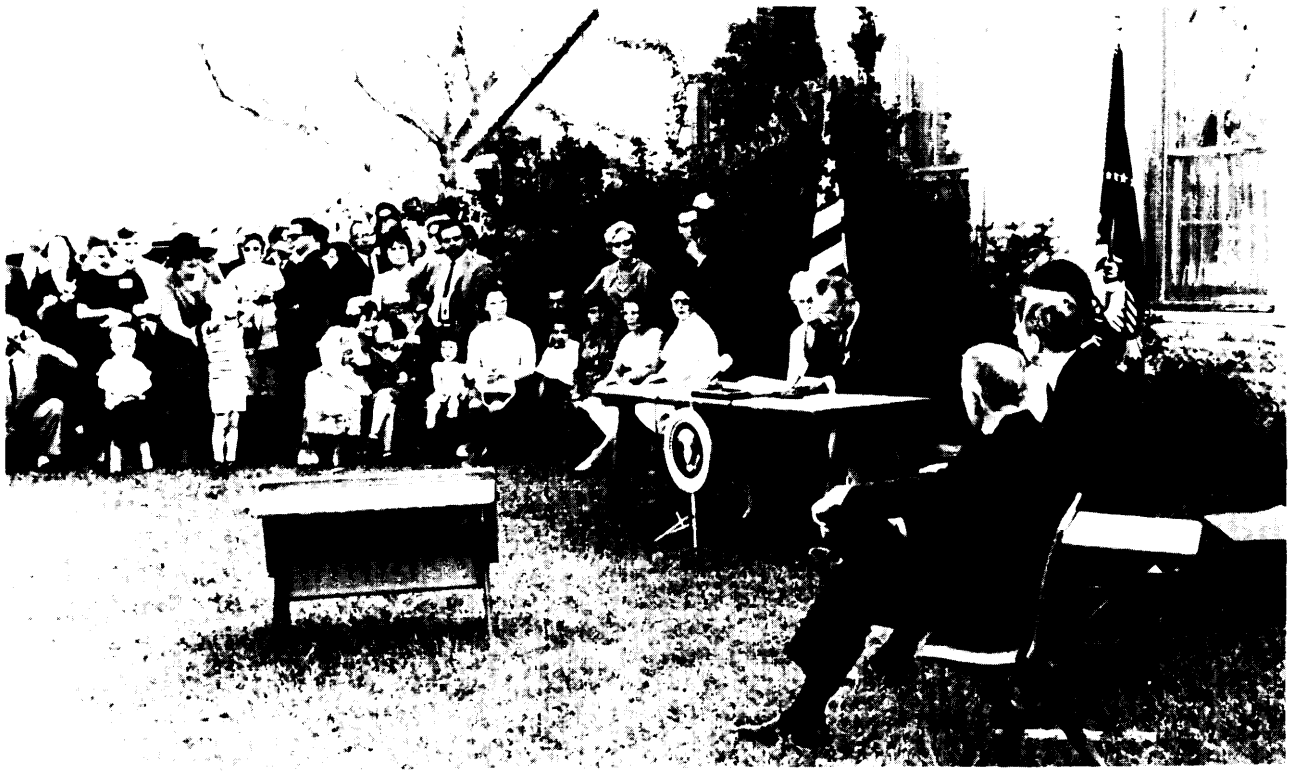

Photo reedit: UPI, Bettmann

President Johnson signing the Elementary and Secondary Education Act of 1965 at a school in Johnson City, Texas. The enactment of this law was a milestone in Federal education policy.

levels. Congress responded by gradually retightening the evaluation requirements. The most recent set of amendments, the 1988 reauthorization, made Chapter 1 assessment more consequential and controversial than ever before by requiring Chapter 1 schools to modify their programs if they could not demonstrate achievement gains among participating children-the so-called 'program improvement provisions.

Through all these revisions, the purposes of Title I/Chapter 1 evaluation have remained much the same: to determine the effectiveness of the program in improving the education of disadvantaged children; to instill local accountability for Federal funds; and to provide information that State and local decisionmakers can use to assess and alter programs.

Specific Requirements for Evaluating Programs

Title I/Chapter 1 is a partnership between Federal, State, and local governments, and the evaluation provisions reflect this division of responsibility. Evaluation of the effects of Chapter 1 on student achievement begins at the project level—usually the school. Test scores of participating children are collected from schools, analyzed, and summarized by the local education agency (LEA). Each LEA reports its findings to the State education agency (SEA), which aggregates the results in a report to the U.S. Department of Education. (States can, if they wish, institute additional requirements regarding the format, content, and frequency of Chapter 1 evaluations.) Congress, by statute, and the Department of Education, through regulations and other written guidance (particularly the guidance in the Department's Chapter 1 Policy Manua1[4]), set standards for SEAS and LEAs to follow in evaluating and measuring progress of Chapter 1 students. The Department also compiles the State data and sends Congress a report summarizing the national achievement results, along with demographic data for Chapter 1 participants.

[4]U.S. Department of Education, *Chapter] Policy Manual* (Washington, DC: April 1990).

Standardized Tests and Mandated Evaluations

Since the creation of the Title I/chapter 1 Evaluation and Reporting System (TIERS) in the mid-1970s, the Department has relied on norm-referenced standardized test scores as an available, straightforward, and economical way of depicting Chapter 1 effectiveness. The law, for its part, gives an imprimatur to standardized tests, through numerous references to "testing," "scores," "objective measures, '' ''measuring instruments,' and ' 'aggregate performance. " Chapter 1 evaluation has became nearly synonymous with norm-referenced standardized testing.

The purpose of TIERS has changed little since it became operative in 1979: to establish standards that will result in nationally aggregated data showing changes in Chapter 1 students' achievement in reading, mathematics, and language arts. To conform with TIERS, States and local districts must report gains and losses in student achievement in terms of Normal Curve Equivalents (NCEs), a statistic developed specifically for Title I. NCEs resemble percentile scores, but can be used to compute group statistics, combine data from different norm-referenced tests (NRTs), and evaluate gains over time. (Gains in scores, which can range from 1 to 99, with a mean of 50, reflect an improvement in position relative to other students.[5]) To produce NCE scores, local districts must use an NRT or another test whose scores can be equated with national norms and aggregated. Thus, although the Chapter 1 statute does not explicitly state that LEAs must use NRTs to measure Chapter 1 effectiveness, the law and regulations together have the effect of requiring NRTs because of their insistence on aggregatable data and their reliance on the NCE standard.

The 1988 law, as interpreted by the Department of Education, changed the basic evaluation provisions in ways that increased the frequency and significance of standardized testing in Chapter 1. Specifically, the law:

- through the new "program improvement" provisions, put teeth into the longstanding Title I/Chapter 1 requirement that LEAs use evaluation results to determine whether and how local programs should be modified. Schools with stagnant or declining aggregate Chapter 1 test scores must develop improvement plans, frost in conjunction with the district and then with the State, until test scores go up.
- gave the Department the authority to reinstate national guidelines for Chapter 1 evaluation (which had been eliminated in 1981) and required SEAS and LEAs to conform to these standards.
- focused greater attention on (and, through regulation, required measurement of) student achievement in higher order analytical, reasoning, and problem-solving skills.
- directed LEAs to develop 'desired outcomes,' or measurable goals, for their local Chapter 1 programs, which could include achievement outcomes to be assessed with standardized tests.
- expanded the option for high-poverty schools to operate schoolwide projects,[6] as long as they can demonstrate achievement gains (i.e., higher test scores) among Chapter l-eligible children.
- as interpreted by the Department, required LEAs to conduct a formal evaluation that met TIERS standards every year, rather than every 3 years. (In actual practice, most States required annual evaluations.)

Other Uses of Tests in Chapter 1

Producing data for national evaluations is only one of several uses of standardized tests in Chapter 1. Under the current law and regulations, LEAs are required, encouraged, or permitted to use tests for all the following decisions:

- identifying which children are eligible for Chapter 1 services and establishing a "cutoff score" to determine which children will actually be served;
- assessing the broad educational needs of Chapter 1 children in the school;

[5]Mary Kennedy, Beatrice F. B- and Randy E. Demaline, *of 1* (Washington DC: U.S. Department of Education, 1986), p. E-2.

[6]Under the schoolwide Project option, schools with 75 percent or more poor children may use their Chapter 1 funds fOr programs to upgrade the educational program for all children, without regard to Chapter 1 eligibility; in exchange for this greater flexibility, these schools must agree to increased accountability.

- determining the base level of achievement of individual Chapter 1 children before receiving services (the "pretest' ');
- assessing the level of achievement of Chapter 1 children after receiving services (the "post-test' '), in order to calculate the change data required for national evaluations;
- deciding whether schools with high proportions of low-achieving children should be selected for projects over schools with high poverty;[7]
- allocating funds to individual schools;
- establishing goals for schoolwide projects;
- determining whether schoolwide projects can be continued beyond their initial 3-year project period;
- annually reviewing the effectiveness of Chapter 1 programs at the school level for purposes of program improvement;
- deciding which schools must mod@ their programs under the "program improvement" requirements;
- determining when a school no longer needs program improvement;
- identifying which individual students have been in the program for more than 2 years without making sufficient progress; and
- assessing the individual program needs of students that have participated for more than 2 years.

In addition, Congress and the Department of Education use standardized test data accumulated from State and local evaluations for a variety of purposes:

- justifying continued appropriations and authorizations;
- weighing major policy changes in the program;
- targeting States and districts for Federal monitoring and audits; and
- contributing to congressionally mandated studies of the program.

Competing Tensions

Chapter 1 is a good example of how Congress must weigh competing tensions when making decisions about Federal accountability and testing. For example, in Chapter 1, as in other education programs, the need for Federal accountability must be weighed against the need for State and local flexibility in program decisions. The Federal appetite for statistics must be viewed in light of the undesirable consequences of too much Federal burden and paperwork-lost instructional time and declining political support for Federal programs, to name a few. The Federal desire for succinct, ''objective, and aggregatable data must be judged against the reality that test scores alone cannot provide a full and accurate picture of Chapter 1's other goals and accomplishments (e.g., redistributing resources to poor areas, mitigating the social effects of child poverty, building children's self esteem, and keeping students in school). Finally, the Federal need for summary evaluations on which to formulate national funding and policy decisions must be weighed against the local need for meaningful, child-centered information on which to base day-to-day decisions about instructional methods and student selection.

The number of times Congress has amended the Chapter 1 evaluation requirements suggests how difficult it is to balance these competing tensions.

Effects of Chapter 1 on Local Testing

Chapter 1 has helped create an enormous system of local testing. Almost every Chapter 1 child is tested every year, and in some cases twice a year, to meet national evaluation requirements. In school year 1987-88, over 1.6 million Chapter 1 participants were tested in reading and just under 1 million in mathematics. Sometimes this testing is combined with testing that fulfills State and local needs; other times Chapter 1 has caused districts to administer tests more frequently, or with different instruments, than they would in the absence of a Federal requirement.

Because SEAS and LEAs often use the same test instruments to fulfill both their own needs and Chapter 1 requirements, and because States and districts expanded their testing programs during roughly the period when Chapter 1 appropriations were growing, it is difficult, perhaps impossible, to sort out which entity is responsible for what degree of the total testing burden. Although States and districts often coordinate their Chapter 1 testing with other testing needs, many LEAs report that without

[7] A proposal to amend Title 1 *so* that all *funding* would be distributed on the basis of achievement test scores was put forth in the late 197@ by then-Congressman Albert Quie (R-MN). The proposal was not accepted, but a compromise provision was adopted, which remains in the law today, permitting school districts to allocate funds to schools based on test scores in certain limited situations.

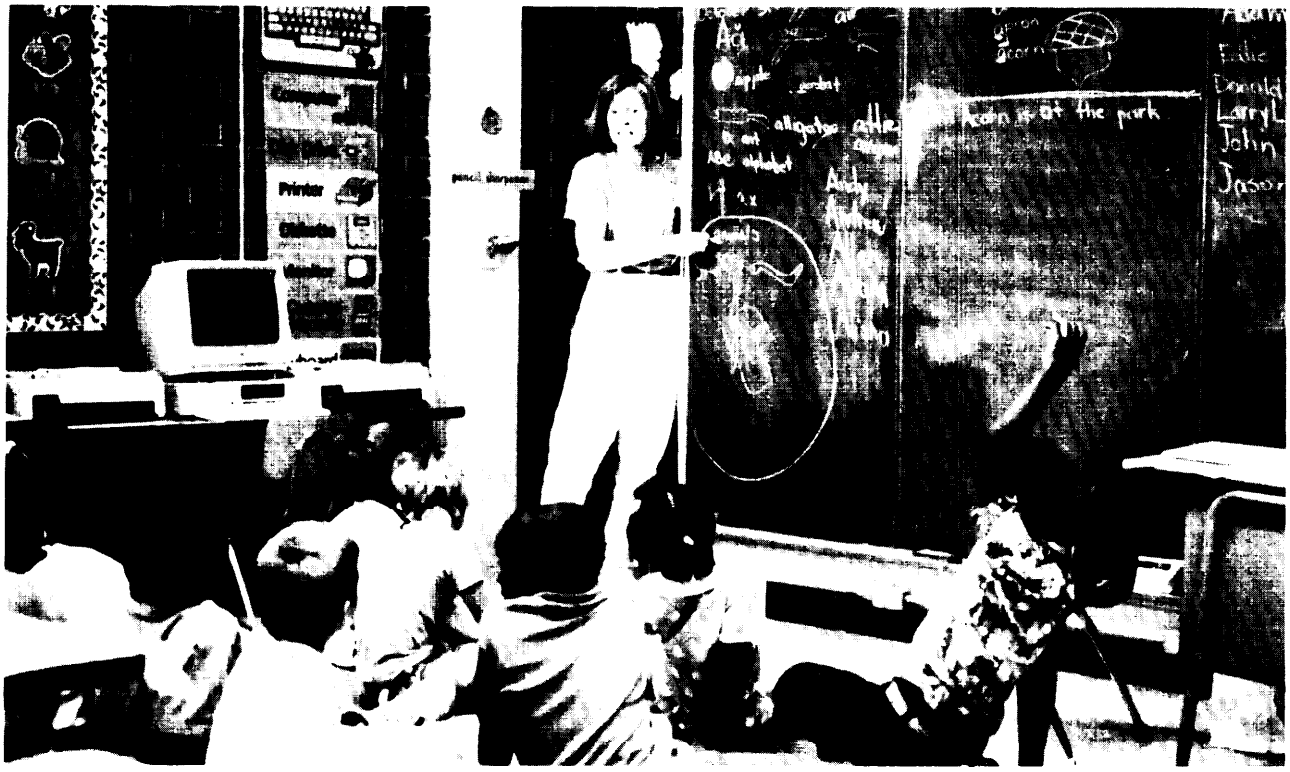

Photo credit: Julie Miller, Education Week

Classrooms like this in Jefferson Parish, Louisiana, benefit from the extra assistance for disadvantaged students provided by Chapter 1 of the Elementary and Secondary Education Act. Testing has always been a big part of Chapter 1 activity.

Chapter 1, they would do less testing. A district administrator from Detroit, for example, estimated that her school system conducts twice as much testing because of Chapter 1.[8] The research and evaluation staff of the Portland (Oregon) Public Schools noted that in the absence of a Chapter 1 requirement to test second graders, their district would begin standardized testing later, perhaps in the third or fourth grade.[9] (In school year 1987-88, about 22 percent of Chapter 1 public and private school participants were in grades pre-K through one and were already exempted from testing. Another 26 percent of the national Chapter 1 population were in grades two and three; these children must be tested under current requirements.) One State Chapter 1 coordinator said that without Chapter 1, his State would require only its State criterion-referenced instrument, and not NRTs. At the school level, principals and teachers express frustration with the amount of time spent on testing and tracking test data in Chapter 1 and the degree of disruption it causes in the academic schedule.

National studies of Chapter 1 and case studies of its impact in particular districts have uncovered some significant concerns about the appropriateness of using standardized tests to assess the program's overall effectiveness, make program improvement decisions, and determine the success of schoolwide projects. Over the years, Chapter 1 researchers and practitioners have raised a number of technical questions about the quality of Chapter 1 evaluation data and have expressed caveats about its limitations in assessing the full impact and long-term consequences of Chapter 1 participation. With the new requirements that raised the stakes of evaluation, debate over the data's validity and limitations has

[8] Sharon Johnson-Lewis, director, Office of Planning, Research and Evaluation, Detroit Public Schools, remarks at OTA Advisory Panel meeting, June 28, 1991.

[9] This and the other observations about the impact of Chapter 1 on testing practices are taken from Kober, op. cit., footnote 1. Case studies of the Philadelphia, PA, and Portland, OR, public schools helped inform OTA's analysis and are cited throughout this chapter.

become more heated. For example, there is evidence from Philadelphia, Portland, and other districts that because of measurement phenomena, test results do not always target for program improvement the schools with the lowest achievement or the weakest programs. Similarly, schools with schoolwide projects have argued that a 3-year snapshot based on test scores does not always provide adequate time or an accurate picture of the project's success compared with more traditional Chapter 1 programs.

State and local administrators have also expressed concerns about the effect of Chapter 1 testing on instruction. While administrators and teachers are loathe to admit to any practices that could be interpreted as ''teaching to the test, '' there is some evidence from case studies and national evaluations that teachers make a point to emphasize skills that are likely to be tested. In districts such as Philadelphia and Portland, where a citywide test tied to local curriculum is also the instrument for Chapter 1 evaluation, teachers can readily justify this practice. Discomfort arises, however, when local administrators and teachers feel they are being pressed by Federal requirements to spend too much time drilling students in the type of "lower order" skills frequently included on commercially published NRTs, or when teachers hesitate to try newer instructional approaches, such as cooperative learning and active learning, for fear their efforts will not **translate into** measurable gains.

Of more general concern is the broad feeling that for the amount of burden it entails, Chapter 1 test data is not very useful for informing local program decisions. According to case studies and other analyses, teachers and administrators use federally mandated evaluation results far less often than other more immediate and more student-centered evaluation methods--e.g., criterion-referenced tests (CRTs), book tests, teacher observations, and various forms of assessment—to determine students' real progress and make decisions about instructional practices. Frequently the mandated evaluations are viewed as a compliance exercise—a ''hoop' that States and local districts must jump through to obtain Federal funding.

Although Chapter 1 teachers, regular classroom teachers, and administrators do occasionally employ other types of assessment to make decisions about Chapter 1 students and projects, these alternative forms are not entrenched in the program in the same way that NRTs are, and are seldom considered part of the formal Chapter 1 evaluation process. While the Chapter 1 law contains some nods in the direction of alternative assessment—particularly for measuring progress toward desired outcomes and evaluating the effects of participation on children in preschool, kindergarten, and frost grade-the general requirements for evaluation cause local practitioners to feel that NCE scores are the only results that really matter. They believe that alternative assessment will not become a meaningful component of chapter 1 evaluation without explicit encouragement from Congress and the Department.

One bottom line question remains: what does the large volume of testing data generated by Chapter 1 evaluation tell Congress and other data users about the achievement of Chapter 1 children? To answer this question, it is useful to consider the data from a 10-year summary of Chapter 1 information, as shown in table 3-1.10 The first thing that is apparent from the summary data is how the millions of individual test scores required for Chapter 1 evaluation are aggregated into a single number for each grade for each year. Average annual percentile gains in achievement--comparing average student pretest scores and average post-test scores—have hovered in the range of 2 to 6 percentiles in reading, and 2 to 11 percentiles in mathematics. For some grade levels, in some years, there have been greater improvements, but in general the gains have been modest and the post-test scores have remained low. For example, in 1987-88 the average post-test score for Chapter 1 fourth graders was the 27th percentile in reading and the 33rd percentile in mathematics. In analyzing these data it is important to understand that Chapter 1 children, by definition, are the lowest achieving students in their schools, and that once a child's test scores exceed the cutoff score for the district that child is no longer eligible for Chapter 1 services. There has been some upward trend, more pronounced in mathematics than in reading, but overall closing of the gap has been slow. In addition, because there is no control group for Chapter 1 evaluation, it is difficult to assess what these post-test scores really mean, i.e., how well Chapter

[10]For the complete tables of data referred to in this discussion, see U.S. Department of Education, *A Summary of State Chapter 1* and (Washington DC: 1990).

Table 3-I—Achievement Percentiles for Chapter 1 Students Tested on an Annual Cycle, 1979-80 to 1987-88

Grade	1979-80	1980-81	1981-82	1982-83	1983-84	1984-85	1985-86	1986-87	1987-88
	Changes in percentile ranks for reading								
2	2	2	2	2	2	3	2	4	4
3	4	5	3	4	4	4	4	5	5
4	3	4	4	4	4	5	5	6	5
5	3	5	5	5	5	6	5	4	4
6	4	6	5	6	5	5	5	5	5
7	2	3	4	3	4	6	4	3	4
8	3	4	5	4	4	4	4	3	4
9	2	4	4	4	3	2	3	2	3
10	-1	2	1	2	1	2	2	2	2
11	-3	3	1	-1	0	2	3	3	2
12	2	0	2	0	1	0	0	2	0
	Changes in percentile ranks for mathematics								
2	2	5	5	3	6	6	9	10	11
3	1	3	5	5	6	4	6	7	7
4	3	6	5	4	5	6	6	8	8
5	4	4	6	8	7	7	9	7	7
6	6	8	6	8	7	6	7	7	7
7	4	3	5	7	5	6	6	5	4
8	4	5	5	6	5	5	4	4	5
9	1	1	2	3	1	2	2	5	4
10	-2	1	0	2	1	2	4	3	4
11	1	2	1	1	2	3	4	3	3
12	2	0	1	0	3	2	2	4	-1

SOURCE: Beth Sinclair and Babette Gutmann, *A Summary of State Chapter 1 Participation and Achievement Information for 1987-88* (Washington, DC: U.S. Department of Education, 1990), pp. 49-50.

1 children would achieve in the absence of any intervention.[11]

For purposes of this analysis, *the real question is whether the information from these test scores is necessary or sufficient to answer the accountability questions of interest to Congress.* For the disadvantaged population targeted by the program, the achievement score gains are evidence of improvement. Thus, when taken together with other evaluative evidence about the program's impact, the test scores support continued funding. But whether the test scores reveal anything significant about what and how Chapter 1 children are learning remains ambiguous. And in the light of unanticipated effects of the extensive testing, it is not clear that the information gleaned from the tests warrants the continuation of an enormous and quite costly evaluation system in its present form.

Ripple Effects of Chapter 1 Requirements

Title I/Chapter 1 established a precedent for achievement-based accountability requirements adopted in many subsequent Federal education programs. In the migrant education program added in 1966, the bilingual education program added in 1967, the Head Start program enacted in the Economic Opportunity Amendments of 1967, and programs that followed, Congress required recipients of Federal funds to evaluate the effectiveness of the programs funded.[12] As a result of Federal requirements, State and local agencies administer a whole range of tests—to place students, assess the level of participants' needs, and determine progress. Even when NRTs are not explicitly required, they are often the preferred mode of measurement for Federal accountability because they can be applied consistently, are relatively inexpensive, and leave a clearly under-

[11] One of the more vexing evaluation problems has been to infer "treatment effects" from studies with no control group. For discussion and analysis of methods designed to correct for 'regression to the mean' and other statistical constraints, see Anand Desai, Technical Issues in Measuring Scholastic Improvement Due to Compensatory Education Programs,' *Socio-Economic* vol. pp. 143-153.

[12] For discussion of outcome-based performance measures in vocational education and job training programs see, e.g., U.S. Congress, Office of Technology Assessment, "Performance Standards for Secondary School Vocational Education," background paper of the Science, Education and Transportation Program, April 1989.

stood and justifiable trail for Federal monitors and auditors.

The 1965 ESEA had another, less widely recognized impact on State testing practices. Title V of the original legislation provided Federal money to strengthen State departments of education, so that they could assume all the administrative functions bestowed on them by the new Federal education programs. This program helped usher in an era of increased State involvement in education and would have a significant impact down the road as States assumed functions and responsibilities far beyond those required by Federal programs or envisioned by Congress in 1965.

Photo credit: The Jenks Studio of Photography

Research has shown that early intervention is important, and many schools like this one in Danville, Vermont, use Chapter 1 funds for preschool and kindergarten programs.

Chapter 1 Testing in Transition

OTA finds that because of its size, breadth, and influence on State and local testing practices, Chapter 1 of ESEA provides a powerful lever by which the Federal Government can affect testing policies, innovative test development, and test use throughout the Nation.

OTA's analysis brings to light several reasons why Congress ought to reexamine and consider significant changes to the Federal requirements for Chapter 1 evaluation and assessment.

- National policymakers and State and local program administrators have different data needs, not all of which are well served by NRTs.
- The implementation of the 1988 program improvement and schoolwide project requirements has underscored some of the inadequacies and limitations of using NRTs for local program decisions, while simultaneously increasing the consequences attached to these tests.
- While the uses and importance of evaluation data have changed substantially as a result of the 1988 amendments, the methods and instruments for collecting this data have remained essentially the same since the late 1970s. A better match is needed between the new goals of the law, particularly the goal to improve the quality of local projects, and the tools used to measure progress toward those goals.

As Congress approaches Chapter 1 reauthorization, it should examine how all the pieces that affect testing under the umbrella of Chapter 1 fit together. Many pieces are interrelated, but they do not always work harmoniously. For example, the timing and evaluation cycles for Federal, State, and local testing in existing law are not well coordinated. As part of this review, Congress should pay particular attention to the need to revise language that inadvertently endorses norm-referenced testing in situations where that type of testing may be inappropriate. Options such as data sampling may meet congressional needs. Clearer legislative language could help maintain and improve accountability, because States and local districts would know better what was expected.

The following questions can guide congressional deliberations regarding changes in Chapter 1:

- What information does Congress need to make policy and funding decisions about Chapter 1? Is Congress getting that information, and is it timely and useful?
- What information does the Department of Education need to administer the program?
- How do the data needs of State and local agencies differ from those of the Federal Government and each other?
- Is it realistic to serve national, State, and local needs with the same information system based on the same measurement tool?
- How well do NRTs measure what Chapter 1 children know and can do?
- Is the nationally aggregated evaluation data that is currently generated accomplishing what

Congress intended? Specifically, do **aggregates** of **aggregates** of **averages** of NCE gains and negative gains present a meaningful and valid national picture of how well Chapter 1 children are achieving?

- To what extent is the value of cumulative data symbolic rather than substantive? For example, is being able to point to **a rising line on a chart as important as having accurate, meaningful data about what** Chapter 1 children know and can do? Can symbolic or oversight needs be fulfilled with less burdensome types of testing?
- What other types of data, beyond test scores, might meet Federal policy makers' criteria for objectivity?

In summary, OTA finds that Congress should revisit the Chapter 1 assessment and evaluation requirements in the attempt to lessen reliance on NRTs, reduce the testing burden, and stimulate the development of new methods of assessment more suited to the students and the program goals of Chapter 1. A careful reworking of the requirements could have widespread salutary effects on the use of educational tests nationwide. Congressional options for achieving these ends are identified in chapter 1 of this report.

National Assessment of Educational Progress

By the late 1960s, Title I/chapter 1 and other Federal programs had produced a substantial amount of data concerning the achievement of disadvantaged children and other special groups of students. State and local testing told SEAS and LEAs how their students stacked up against national norms on specific test instruments. What was missing, however, was a context—a nationally representative database about the educational achievement of elementary and secondary school children as a group, against which to confirm or challenge inferences drawn from State, local, or other nationwide testing programs.

Although policymakers and the public could draw from a wide variety of statistics to make informed decisions on such issues as health and labor, they were operating in a vacuum when it came to education. The Department of Education produced a range of quantitative statistics on school facilities, teachers, students, and resources, but had never collected sound and adequate data on what American students knew and could do in key subject areas.

Francis Keppel, U.S. Commissioner of Education from 1962 to 1965, became troubled by this dearth of information and initiated a series of conferences to explore the issue.[13] In 1964, as a result of these discussions, the Carnegie Corp. of New York, a private foundation, appointed an exploratory committee and charged it with examining the feasibility of conducting a national assessment of educational attainments. By 1966, the committee had concluded that a new battery of tests-carefully constructed according to the highest psychometric standards and with the consensus of those who would use it—would have to be developed.[14]

The vision became a reality in 1969, when the U.S. Office of Education began to conduct periodic national surveys of the educational attainments of young Americans. The resulting effort, NAEP, sometimes called ''the Nation's report card," has the primary goal of obtaining reliable data on the status of student achievement and on changes in achievement in order to help educators, legislators, and others improve education in the United States.

Purpose

Today, NAEP remains the only regularly conducted national survey of educational achievement at the elementary, middle, and high school levels.[15] To date it has assessed the achievement of some 1.7 million young Americans. Although not every subject is tested during every administration of the program, the core subjects of reading, writing, mathematics, science, civics, geography, and U.S.

[13] In 1%3, Keppel is reported to have lamented the fact that: ''Congress is continually asking me about how bad or how good the schools are and we have no dependable information. They give different tests at schools for different purposes, but we have no idea generally about the subjects that educators value. . . .' OTA interview with Ralph W. Tyler, Apr. 5, 1991.

[14] This early history of the National Assessment of Educational Progress (NAEP) is taken from the National Assessment Of Educational Progress, *General Information* (Washington, DC: National Center for Education Statistics, 1974); and George Madaus and Dan Stufflebeam (eds.), *Educational Evaluation: Works of Ralph Tyler* (Boston, MA: Kluwer Academic Publishers, 1989). Conversations with Frank Womer, Edward Roeber, and Ralph Tyler, all involved in different capacities in the original design and implementation of NAEP, enriched the material found in published sources.

[15] National Assessment of Educational Progress, (Princeton, NJ: Educational Testing Service, 1986).

Photo credit: Office of Technology Assessment, 1992

Known as the Nation's Report Card, the National Assessment of Educational Progress issues summary reports for assessments conducted in" a number of academic subject areas. These reports also analyze trends in achievement levels over the past 20 years.

history have been assessed more than once to determine trends over time. Occasional assessments have also examined student achievement in citizenship, literature, art, music, computer competence, and career development.

Safeguards and Strengths

The designers of the NAEP project took extreme care and built in many safeguards to ensure that a national assessment would not, in the worst fears of its critics, become any of the following: a stepping stone to a national individual testing program, a tool for Federal control of curriculum, a weapon to ''blast' the schools, a deterrent to curricular change, or a vehicle for student selection or funds allocation decisions.[16] An understanding of NAEP's design safeguards is crucial in order to comprehend what NAEP was and was not intended to do and why it is unique in the American ecology of student assessment. NAEP has seven distinguishing characteristics.

NAEP reports group data only, not individual scores. NAEP results cannot be used to infer how particular students, teachers, schools, or districts are achieving or to diagnose individual strengths and weaknesses. Prevention of these levels of score reporting was a prerequisite to gaining approval for

[16]Tyler, **op. cit., footnote** 13.

the original development and implementation of NAEP.[17]

NAEP is essentially a battery of criterion-referenced tests in various subject areas (although its developers prefer the term "objective-referenced," since NAEP tests are not tied to any specific curriculum but measure the educational attainment of young Americans relative to broadly defined bodies of knowledge). Unlike many commercially published NRTs, NAEP scores cannot be used to rank an individual's performance relative to other students. This emphasis on criterion-referenced testing represents an important shift toward outlining how children are doing on broad educational goals rather than how they are doing relative to other students. NAEP is the only test to provide this kind of information on a national scale.

NAEP has pioneered a survey methodology known as "matrix sampling." This approach grew out of item-response theory, and has been hailed as an important contribution to the philosophy and practice of student testing.[18] Under this method, a sample of students across the country is tested, rather than testing all students (which would be considered a' 'census' design). Furthermore, the students in the matrix sample do not take a "whole' test, or even the same subject area tests, nor are they all given the same test items. Rather, each student takes a l-hour test that includes a mix of easy, medium, and difficult questions. Thus, NAEP uses a method of sampling, not only of the students, but also of the content that appears on the test. Any student taking a NAEP test only takes one-seventh of the test in a l-hour testing session. *Because of matrix sampling, a much wider range of content and goals can be covered by the test than most other tests can allow. This* broad coverage of content is the essential foundation of a nationally relevant test, as well as a test that is relatively well protected against the negative side effects that can occur with teaching to a narrow test. It is probable that these important strengths of NAEP, which make it a robust and nationally credible test, would be difficult to incorporate into a test designed to be administered to individuals (unless it were a prohibitively long test). In addition, because no individual students can be assigned scores, the matrix sampling approach imposes an important technological barrier against the use of NAEP results for making student, school, district, or State comparisons, or for sorting or selecting students.

NAEP provides comparisons over time, by testing nationally representative samples of 4th, 8th, and 12th graders on a biennial cycle. (Prior to 1980, NAEP tested on an annual cycle.) This form of sampling deters the kinds of interpretation problems that can arise when different populations of test takers are compared.[19] Due to cost constraints, the out-of-school population of students that had been sampled in early NAEP administrations was eliminated.

NAEP strives for consensus about educational goals. NAEP's governing board employs a consensus-building process for establishing content frameworks and educational objectives that are broadly accepted, relevant, and forward looking. Panels of teachers, professors, parents, community leaders, and experts in the various disciplines meet in different locales and work toward agreement on a common set of objectives for each subject area. These objectives are then given to item writers, who come up with the test questions. Before the items are administered to students, they undergo careful scrutiny by specialists in measurement and the subject matter being tested and are closely reviewed in the effort to eliminate racial, ethnic, gender, and other biases or insensitivities.[20]

Recognizing that changing educational objectives over time can complicate its mandate to plot trends in achievement, NAEP has developed a valuable process for updating test instruments. Using this process, NAEP revises test instruments to reflect new developments in curricular objectives, at the same time maintaining links between current and past levels of achievement of certain freed objec-

[17]See, e.g., James Hazlett, University of Kansas, ''A History of the National Assessment of Educational Progress, 1%3-1973," unpublished doctoral dissertation, December 1973.

[18]The principles of matrix sampling are now used in many State assessment programs, as well as in other countries. See *ChS.* 6 and 7 for additional discussion.

[19]For example, this was a major problem in using the decline in Scholastic Aptitude Test scores as a basis for the inference that overall achievement had fallen. See Robert Linn and Stephen Dunbar, "The Nation's Report Card Goes Home: Good News and Bad About Trends in Achievement" *Delta Kappan*, October 1990, pp. 127-133.

[20]National Assessment of Educational Progress, op. cit., footnote 15.

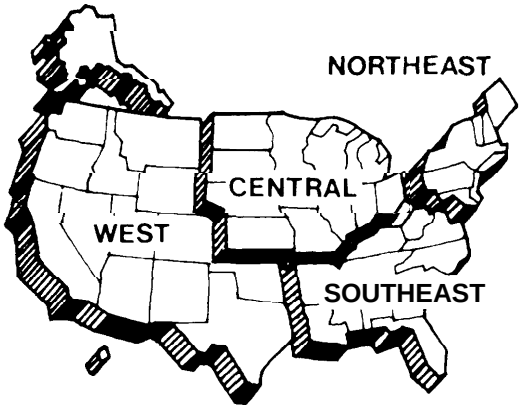

Photo credit: National Assessment of Educational Progress

In addition to information about the Nation as a whole, the National Assessment of Educational Progress (NAEP) reports for four regions of the country as well as by sex, race/ethnicity, and size and type of community. NAEP does not report results for individual students, but generates information by sampling techniques.

tives. In mathematics and reading, for example, representative samples of students are assessed using methods that have remained stable over the past 20 years, while additional samples of students are tested using instruments that reflect newer methods or changed definitions of learning objectives. Thus, the 1990 mathematics assessment allowed some students to use calculators, a decision generally praised by the mathematics teaching community. The NAEP authors took care to note, however, that the results of these samples were not commensurate with the mathematics achievement results from prior years.

Although NAEP is predominantly a paper-and-pencil test relying heavily on multiple-choice items, certain assessments include open-ended questions or nontraditional formats. For example: the writing assessment requires students to produce writing samples of many different kinds, such as a persuasive piece or an imaginative piece; the 1990 assessment also included a national 'writing portfolio' of works produced in classrooms; the science assessment combines multiple-choice questions with essays and graphs on which students fill in a response; and the 1990 mathematics assessment included several questions assessing complex problem-solving and estimation skills, as recommended by the mathematics teaching profession.

During its early years, NAEP experimented with even more varied test formats and technologies, conducting performance assessments in music and art that were administered by trained school personnel and scored by trained teachers and graduate students. Although many of its more innovative approaches were suspended due to Federal funding constraints,[21] many State testing programs continue to use the performance assessment technologies pioneered by NAEP. Moreover, NAEP continues to be a pioneer in developing open-ended test items that can be used for large scale testing; this is possible largely due to matrix sampling.

Accomplishments

All of these strengths have lent NAEP a degree of respect that is exceptional among federally sponsored evaluation and data collection efforts. NAEP has produced 20 years of unparalleled data and is considered an exemplar of careful and innovative test design. NAEP reports are eagerly awaited before publication and widely quoted afterward. In addition, NAEP collects background data about students' family attributes, school characteristics, and student attitudes and preferences that can be analyzed to help understand achievement trends, such as the relationship between television and reading achievement.

Because of NAEP, the Nation now knows, among other trends, that Black students have been narrowing the achievement gap during the past decade, 9-year-olds in general read better now than they did 10 years ago, able 13-year-olds do less well on higher order mathematics skills than they did 5 years ago, and children who do homework read better than those who do not.

Caveats

A relatively recent issue has emerged with potential consequences for NAEP administration and for interpretation of NAEP results. Researchers have begun to question whether NAEP scores tend to *underestimate* knowledge and skills of American students, precisely because NAEP is perceived as a low-stakes test. The question is whether students perform at less than their full ability in the absence

[21] For discussion of the 1974 funding crisis, see Hazlett, op. cit., footnote 17, pp. 297-299.

of extrinsic motivation to do well. It is not purely an academic question: much of today's debate over the future of American education and educational testing turns on public perceptions of the state of American schooling, perceptions based at least in part on NAEP.

Some empirical research on the general question of motivation and test performance has already demonstrated that the issue may be more important than originally believed. For example, one study found that students who received ". . . special instructions to do as well as possible for the sake of themselves, their parents, and their teachers. . ." did significantly better on the Iowa Tests of Basic Skills than students in the control group who received ordinary instructions.[22] This result supports the general findings in research discussed in the preceding chapter;[23] and another analyst's observation that ". . . when a serious incentive is present (high school graduation), scores are usually higher.' '[24]

Prompted by these and other findings, several researchers are conducting empirical studies to determine the specific motivational explanations of performance-on NAEP. One study involves experimental manipulation of instructions to NAEP test takers; the other involves embedding NAEP items in an otherwise high-stakes State accountability test.[25] Data are to be collected in spring 1992. The results of these studies will shed light on an important aspect of how NAEP scores should be interpreted.[26]

The 1988 Amendments

The original vision of NAEP has been diminished by years of budget cuts and financial constraints. Some of what NAEP once had to offer the Nation has been lost as a result. Concomitantly, over the past few years, new pressures have arisen in the attempt to adapt NAEP to serve purposes for which it was never intended. Some of this pressure has come from policymakers illustrated with the lack of effect of NAEP results in shaping educational policy and the relatively "low profile" of the test and the results. Responding in part to this pressure, Congress took some cautious steps in 1988 to amend NAEP to provide new types of information.

One dilemma that surfaced during NAEP's first two decades was that its results did not appear to have much impact on education policy decisions, especially at the State and local levels. While theoretically NAEP could provide benchmarks against which State and local education authorities could measure their own progress, many educators argued that the information was too general to be of much help when they made decisions about resource allocations. Others observed that since NAEP carried no explicit or implicit system for rewards or sanctions, there was simply no incentive for States and localities to pay much attention to its results.

Had NAEP not been so highly respected, criticisms about its negligible influence on policy might have been considered minor, but given NAEP's reputation, its lack of clout was viewed as a major lost opportunity. Pressure mounted to change NAEP to make State and local education authorities take greater heed of its message. These voices for change were quickly met by experts who reissued warnings from the past: that any attempts to use NAEP for purposes other than analyzing aggregate national trends would compromise the value of its information and ultimately the integrity of the entire NAEP program.[27] The principal concerns were:

1. that turning NAEP into a high-stakes test would lead to the kinds of score 'inflation' or ''pollution' that have undermined the credibility of other standardized tests as indicators of achievement (see ch. 2); and
2. that using NAEP to compare student attainment across States would induce States to change their curricula or instruction for the

[22]Steven M. Brown and Herbert J. Walberg, University of Illinois at Chicago, "Motivational Effects on Test Scores of Elementary School Students: An Experimental Study," monograph 1991.

[23]See Daniel Koretz, Robert Linn, Stephen Dunbar, and Lorrie Shepard, ''The Effects of High Stakes Testing on Achievement: Preliminary Findings About Generalization Across Tests,'' paper presented at the annual meeting of the American Educational Research Association, Chicago, IL, April 1991.

[24]See Paul Burke, "You Can Lead Adolescents to a Test But You Can't Make Them Try,' OTA contractor report, Aug. 14, 1991, p. 4.

[25]Robert Linn, University of Colorado at Boulder, personal communication, November 1991.

[26]For discussion of general issues regarding the public's understanding of National Assessment of Educational Progress scores, see Robert Forsyth, "Do NAEP Scales Yield Valid Criterion-Referenced Interpretations?" *Issues and Practice*, 10, No. 3, fall 1991, pp. 3-9; and Burke, op. cit., footnote 24.

[27]The strongest early warnings about NAEP were found in Harold Hand, "National Assessment Viewed as the Camel's Nose, " *Kappan*, 1, 1965, pp. 8-12; and Harold Hand, "Recipe for Control by the Few, " vol. 30, No. 3, 1966, pp. 263-272,

sake of showing up better on the next test, rather than as a result of careful deliberations over what should be taught to which students and under what teaching methods.

When NAEP came up for congressional reauthorization in 1988, it was amid a climate of growing public demands for accountability at all levels of education (fueled in part, ironically, by NAEP's own reports of mediocre student achievement in critical subjects). Almost a decade of serious education reform efforts had made little visible impact on American students' test scores, especially relative to those of international competitors.

Trial State Assessment

Congress responded by authorizing, for the first time, State-level assessments, to be conducted on a voluntary, trial basis. Beginning with the 1990 eighth grade mathematics assessment and the 1992 fourth grade mathematics and reading assessments, NAEP results were to be published on a State-by-State basis for those States that chose to participate. Congress considered this amendment a trial, to be followed up with careful evaluation, before the establishment of a full-scale, State-level NAEP program could be considered.

While proponents believed that the experiment would yield useful information for SEAS, critics worried that a State-by-State assessment would invite fruitless comparisons among States that did not take into account other factors influencing achievement; would put pressure on States to teach to the test or find other ways to artificially inflate scores; or would lead to general ''education bashing." Most importantly, critics cautioned that with the State assessment Congress would eventually succumb to pressure to allow assessments and comparative reporting by district, by school, or even by student-a travesty of NAEP's original purpose and design.

Thirty-seven States, the District of Columbia, Guam, and the Virgin Islands participated in the first trial State assessment of mathematics, conducted in 1990. Results were released in June 1991.[28] As expected, some media reports focused on the inevitable question of: ''Where does *your* State rank?' In general, however, the consequences of the trial will not be apparent for some time. In addition to analyzing the effects of the trial on the quality and validity of NAEP data and on State and local policy decisions, observers are likely to focus on whether the information will be worth the high cost of administering the State assessments, and whether the cost of the State programs will crowd out other necessary expenditures or improvements in the basic NAEP program.

Standard Setting

The 1988 reauthorization made another fundamental revision in the original concept of NAEP. From its inception, NAEP had reported results in terms of proficiency scales, pegged to everyday descriptions of what children at that performance level could do. For example, a 200 score in reading meant that students . . have learned basic comprehension skills and strategies and can locate and identify facts from simple informational paragraphs, stories, and news articles."[29] NAEP has been commended for its accuracy in describing how things are. In the late 1980s, however, it came under criticism because it was silent on how things *ought " to be. Those* who saw NAEP as a potential tool for reforming schools or measuring progress toward the President's and the Governors' National Goals for the year 2000 thought that NAEP should set proficiency standards-benchmarks of what students should be able to do. As with the statewide assessment proposal, the recommendation for proficiency standards raised the hackles of many educators, researchers, and policy makers. Opponents of the proposal said it would undermine local control of education; increase student labeling, tracking, and sorting; and compromise NAEP's original purpose and validity.

The 1988 amendments created a new governing body, the National Assessment Governing Board (NAGB), and charged it with identifying ". . . appropriate achievement goals for each age and grade in each subject area." NAGB has completed the standard-setting process for mathematics in 4th, 8th, and 12th grades, and in doing so, generated considerable controversy. Many observers felt that the

[28] See Ina V.S. Mullis, John A. Dossey, Eugene H. Owen, and Gary W. Phillips, Educational Testing Service, *The State of Mathematics* Achievement, prepared for the National Center for Education Statistics (Washington DC: U.S. Department of Education, Education Information Branch, June 1991).

[29] For analysis of National Assessment of Educational Progress' definitions of literacy see John B. Carroll, ''The National Assessments in Reading: Are We Misreading the Findings?" *Delta Kappan*, vol. 68, No. 6, February 1987, pp. 424-430.

mathematics standards were hammered out too quickly, before true consensus was achieved.

Adding the trial State assessment and standard-setting activity increased NAEP funding from about $9.3 million in fiscal year 1989 to over $17 million in fiscal year 1990 (nominal dollars).

NAEP in Transition

When authorization for the trial State assessments and standard-setting processes expires, Congress will face the issue of whether to continue and expand these efforts. As of now, Congress has authorized planning for the 1994 trial, but has not appropriated funds for the implementation of the trial itself. The Administration's "America 2000 Excellence in Education Act' recommends authorization of State-by-State comparisons in five core subject areas (mathematics, science, English, history, and geography) beginning in 1994 as a means of monitoring (and stimulating) progress toward the National Goals. The Administration's bill also suggests that tests used in NAEP be made available to States that wish to use them for testing at school or district levels at their own expense.

In conclusion, the basic question facing Congress is whether to make NAEP even more effective at what it was originally intended to do, or to explore ways that NAEP could serve new purposes. OTA finds that any major changes in NAEP should be carefully evaluated with respect to potential effects on NAEP's capacity to serve its original purpose.

National Testing

Overview

Perhaps the proposals with the most far-reaching implications for the Federal role in testing are those calling for the creation and implementation of a national testing program. Although the objectives of the various national testing proposals are somewhat unclear, they appear to rest on two basic assumptions: first, that the skills and knowledge of most American schoolchildren do not meet the needs of a changing global economy; and second, that new tests can create incentives for the teaching and learning of the appropriate knowledge and skills. Momentum for these efforts has built rapidly, fueled by numerous governmental and commission reports on the state of the economy and of the educational system; by the National Goals initiative of the President and Governors; by casual references to the superiority of examination systems in other countries; and most recently by the President's 'America 2000" plan.

Photo credit: National *Assessment of Educational Progress*

***The* National Assessment of Educational Progress has developed and pilot tested a variety of hands-on science and mathematics tasks. In this example, students watch an administrator's demonstration of centrifugal force and then respond to written questions about what occurred in the demonstration.**

Taken together, the questions of purpose and balance between local control and national interest frame the debate regarding the desirability of national testing. This debate must reflect both the needs of the Nation and the well being of individual students.

Congress provides the best forum for review of this question. Commitment to such a test represents a major change in education policy and should not be undertaken lightly. A number of issues must be considered in weighing the concept.

Will testing create incentives that motivate students to work harder? **What are the** effects of tests on the motivation of students? Tests should reward classroom effort, rather than undermine it. Tests built on comparing students to one another, for example, may reinforce the notion that effort does not matter, since the bell curve design of norm referencing always places some students at the top, some at the bottom, and most in the middle.

Furthermore, if the test is of no consequence to the students, they may not be motivated to try hard or to study to prepare for it. The motivation of those who do poorly on tests must be carefully considered. Those students who repeatedly experience failure on tests (starting in the earliest years of schooling), without any assistance or guidance to help them master test content, are unlikely to be motivated by a high-stakes test. Positive motivational effects are likely only if students perceive they have a good chance of achieving the rewards attached to strong test performance.

How broad will the content and skills covered be? Can just one test be offered to all students at a particular grade level, or will there need to be a range of tests at various levels and disciplines? This affects the testing burden on any one student and the range of levels at which testing can be focused. In some European countries, for example, students take subject-specific examinations at a choice of levels. Some examinations take many hours or are administered over several days, with combinations of testing items and formats that call on a range of performance by the student.

Would the test be voluntary or mandatory? Voluntary tests sound appealing. However, if a test becomes very widely used or needed for access to important resources, it will no longer be truly voluntary. Choosing not to take a test may not be a neutral option; negative consequences may result for those who choose not to be tested. This is especially true if a test is used for selection or credentialing; without a test result in hand, what chance does the student have? Furthermore, voluntary tests do not provide an accurate picture if the goal is school accountability. If only those students, schools, districts, or States that feel they can do well on a test participate in it, the results give an inaccurate picture of achievement. The claim that an important test can be voluntary should be taken with a grain of salt.

What happens to those who fail? Are there resources provided to help them? If consequences for failure are high and a student has no recourse once the examination has been taken, the wisest choice for a student who is having difficulty in school is to skip the examination altogether. The negative effects of examinations on students who do not do well have been a matter of serious concern in many European countries. Some countries have been dismayed to find that some students leave school before required high-stakes examinations are offered, rather than face the indignity and stigma that accompanies failure. This has also occurred with high school graduation examinations in some parts of this country. Rather than punishing those who do not succeed at standards that seem unattainable, tests can be designed to make standards more explicit and the path to their acquisition more clear. However, if it is certain that low scores do not mean failure but that additional or refocused resources will be provided to the student, testing can have positive outcomes.

Who will design the tests and set performance standards? In the decentralized U.S. educational system, national testing proposals raise questions of State and local responsibility for determining what is taught and how it is taught. Can any test content be valid for the entire Nation? Who shall be charged with determining test content? It is important to recall that achievement tests by definition must assess material taught in the classroom. As the content of a test edges away from the specifics of what is delivered in classrooms, based on State-defined curricular goals, and searches instead for common elements, it can become either a test of "basic skills" or of more general skills and understandings. In the latter case, however, the test risks becoming more a measure of aptitude than one of achievement. (See also, ch. 6, box 6-A.) Similarly, setting performance standards on a national basis assumes the feasibility of consensus not only on what is taught and measured, but also on what constitutes acceptable performance, and on procedures to distinguish among levels of performance.

Will the content and grading standards be visible or invisible? Will the examinations be secret or disclosed? Experience from the classroom and other countries suggests that students are more motivated and will learn better when they understand what is expected of them and when they know what competent performance looks like. It is important to note that in Europe the impact of examinations on teaching and learning—what is taught and learned and how it is taught and learned-is mediated through the availability of past examination papers. The tradition in this country is just the opposite. Most high-stakes examinations are kept secret, in part because of high development costs. For a national examination to have salutary effects on learning, the additional costs of item disclosure

should be weighed against the larger impact of the examination on teaching and learning.

Would the examination be administered at a single setting or several times, perhaps when students feel ready? This question affects students' control over the opportunity to study and prepare for an examination. If students can schedule a test when they feel they have mastered the material, they are more likely to be motivated by a realistic expectation of success. Conversely, accountability examinations are more likely to require single-sitting administration if they measure achievement within a common timeframe.

Do students have a chance to retake an examination to do better? Allowing retakes suggests a mastery model in which effort is rewarded and students can try again if they do not master the material the frost time. It reinforces the idea that students can learn what they need to know.

Would the tests be administered to samples of students or individuals? If a test is intended to increase student motivation, then it will have to be an individual test. However, tests administered to individuals need safeguards to meet high technical standards if they will affect the future opportunities of individuals.

At what age are students to be tested? American elementary schoolchildren are tested far more often than their European counterparts, especially with standardized examinations. Much of the rationale for this testing is related to the selection of children for Chapter 1 services and for identification of progress within those programs. This testing has had a spill-over effect greatly influencing overall elementary school testing practice. However, the use of multiple-choice, standardized norm-referenced testing of elementary school children in general, and young (prior to grade three) children in particular, is under attack by those who see the negative consequences of early labeling. Thus, the suggestion of a new national examination at this age stands in contrast with efforts in many States to reduce early childhood standardized testing and to use instead teacher assessments, checklists, portfolios, and other forms of performance-based assessments.

What legal challenges might be raised? Legal challenges based on fairness have become a part of the American landscape. Public policy in this country is based on assurances of equal protection under the law; furthermore, cultural and racial diversity make equity issues far more significant in this country than in most others. Tests must meet these challenges by careful design that assures that the administration and scoring procedures are fair, the content measures what all participants have been taught, and the scores are used for the purposes understood and agreed to by the participants.

What test formats will be used? Tests send important signals to students about the kinds of skills and knowledge they need to learn. Tests that rely on a single format, such as multiple choice, are likely to send a limited message about necessary skills. As noted earlier, the United States and Japan are the only countries to rely almost exclusively on multiple-choice paper-and-pencil examinations for testing. Current proposals for national tests range from the use of multiple-choice norm-referenced standardized tests to the use of "state-of-the-art" assessment practices. Test format and procedures for scoring go hand in hand. Because performance assessments generally involve scoring by teachers or other experts, they are more expensive than machine-scorable tests. A diversity of formats in tasks and items may be the best means of balancing tradeoffs between the kinds of skills and understandings that any one test can measure and the costs of testing.

Conclusions

The answers to these questions will shed light on the larger questions of whether or not national testing is desirable. Goals must be clearly set to determine the kind of tests, content, costs, and potential linkages to curriculum. For example, if Congress sets as its goal increasing student effort for higher achievement by testing in specific subjects, one would expect mandatory tests, administered to all individuals, with the content made explicit through a common syllabus covering a broad scope of material, with past test items made public so students can study and practice for them. If other countries are to be a guide, this kind of examination is not used for testing children under the age of 16 or even 18. Some States are already using tests of this sort (e.g., New York Regents, California Golden State Examinations) for students as high school-leaving examinations. Congress should consider how the participation of these States would be affected, or how these tests could serve as models for use, or be calibrated to match some national standard.

Furthermore, if the goal is to encourage performance that includes direct measures of complex tasks, then written essays, portfolios of work over time, or oral presentations may be called for. These tests would be considerably more costly to develop, administer, and score than machine-scored norm-referenced examinations. Tests of this type are not as carefully researched and may be challenged if used prematurely for high-stakes outcomes like selection or certification.

At present, there is controversy over the use of many test results. The development and use of tests is complicated, both in terms of science and politics. If a test is placed into service at the national level before these important questions are answered, OTA finds that the test could easily become a barrier to many of the educational reforms that have been set into motion and become the next object of concern and frustration within the American educational system.

Congress should consider the questions of test desirability and use first, and then consider policy directions that emerge from these conclusions. This deliberation cannot be separated from a comprehensive look at the other issues discussed in this section, specifically, the role of NAEP in the national testing mosaic, the ways testing is used for Chapter 1 purposes, and how students' interests are to be protected. The policy implications of these choices are considered collectively in chapter 1.