CHAPTER 1

Summary and Policy Options

Contents

Boxes

Figures

CHAPTER 1
Summary and Policy Options

The American educational system is unique. Among the first in the world to establish a commitment to public elementary and secondary schooling for all children, it has achieved an extraordinary record: enrollment rates of school-age children in the United States are among the highest in the world, and over 80 percent finish high school in some form between the ages of 18 and 24.[1] This tradition of education for the masses was nurtured in a system that, by all outward appearances, is complex and fragmented: 40 million children enrolled in some 83,000 schools scattered across some 15,000 school districts. Pluralism, diversity, and local control—hallmarks of American democracy-distinguish the American educational experiment from others in the world.

Student testing has always played a pivotal role in this experiment. Every day millions of school children take tests. Most are devised by teachers to see how well their pupils are learning and to signal to pupils what they should be studying. Surprise quizzes, take-home written assignments, oral presentations, pretests, retests, and end-of-year comprehensive examinations are all in the teacher's toolbox.

It is another category of test, however-originating outside the classroom, usually with standardized rules for scoring and administration-that has garnered the most attention, discussion, and controversy. From the earliest days of the public school movement, American educators, parents, policymakers, and taxpayers have turned to these tests as multipurpose tools: yardstick of individual progress in classrooms, agent of school reform, falter of educational opportunity, and barometer of the national educational condition.

Commonly referred to as "standardized tests,"[2] these instruments usually serve management functions; they are intended to inform decisions made by people other than the classroom teacher. They are used to monitor the achievement of children in school systems and guide decisions, such as students' eligibility for special resources or their qualification for admission to special school programs. Children's scores on such tests are often aggregated to describe the performance of classrooms, schools, districts, or States. With technological advances, these tests have become more reliable and more precise, and their popularity has grown. Today they are a fixture in American schools, as common as books and classrooms; standardized test results have become a major force in shaping public attitudes about the quality of American schools and the capabilities of American students.

Testing at a Crossroads

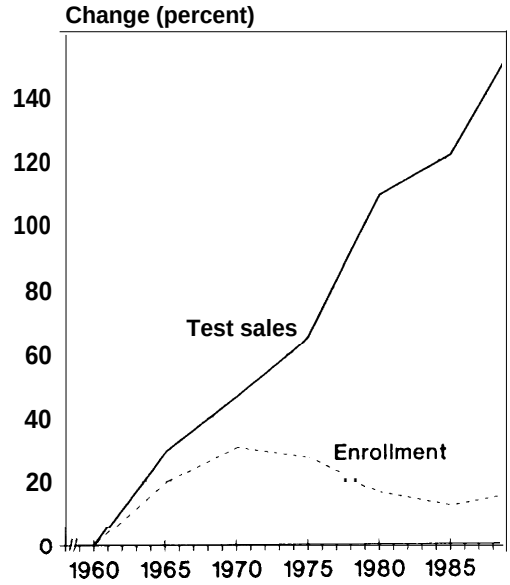
Tests designed and administered outside the classroom are given less frequently than teacher-made tests, but they are thoroughly entrenched in the American school scene and their use has been on the rise. One indicator of growth is sales of commercially produced standardized tests. Revenues from sales of tests used in elementary and secondary schools more than doubled (in constant dollars) between 1960 and 1989 (see figure l-l), a period during which student enrollments grew by only 15 Percents The rise in testing reflects a heightened demand from legislators at all levels-and their constituents-for evidence that education dollars

[1]For current data comparing primary and secondary school enrollment rates in the United States and other countries, see U.S. Department of Education, National Center for Education Statistics, *Digest of 1990* February 1991), p. 380; and George Madaus, Boston College, and Thomas Kellaghan, St. Patricks College, Dublin, "Student Examination Systems in the European Community: Lessons for the United States, ' OTA contractor report, June 1991. For a thorough analysis of completion and dropout data, see U.S. Department of Education, National Center for Education Statistics, *Rates in the 1989* (Washington DC: September 1990). With respect to postsecondary education, as well, participation rates of American high school graduates are the highest in the world: close to 60 percent of persons of college-going age were enrolled in postsecondary institutions in 1985, compared to 30 percent in France, Germany, and Japan, 21 percent in the United Kingdom, and 55 percent in Canada. For details see Kenneth Redd and Wayne Riddle, Congressional Research Service, "Comparative Education: Statistics on Education in the U.S. and Selected Foreign Nations, ' 88-764 EPW, Nov. 14, 1988.

[2]Testing terms have both technical and common meanings, and often cause confusion. Box 1-A is a glossary of words used in this report, and will help the reader understand the precise meanings of these words.

[3]U.S. Department of Education, *Digest of Education Statistics, 1990*, op. cit., footnote 1, p. 12. The fact that testing grew proportionally more rapidly than the student population suggests that policymakers may have responded to increased enrollments by attempting to institute greater administrative efficiency in the schools. As discussed in ch. 4, this is a familiar historical trend.

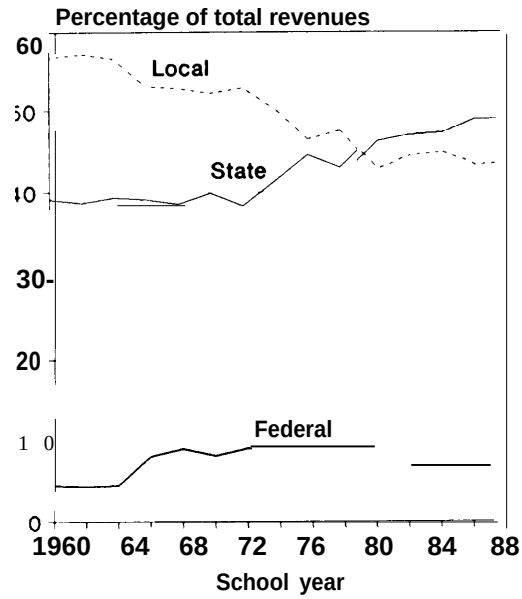
Figure 1-1--Growth in Revenues From Test Sales and in Public School Enrollments, 1960-89



NOTE: Revenues from test sales are in constant 1982 dollars. Tests are commercially produced standardized tests for grades K-1 2. Enrollments are total students in public schools, grades K-12. Percent change is computed over 1960 base year (not over prior year level).

SOURCE: Office of Technology Assessment, 1992. Test sales data from Filomena Simora (cd.), *The Bowker Annual* (New York, NY: Reed Publishing, 1970-1 990). Enrollment data from U.S. Department of Education, National Center for Educational Statistics, *Digest of Education Statistics,* 1990(Washington, DC: February 1991), p. 12.

Figure 1-2-Shifts in Federal, State, and Local Funding Patterns for Public Elementary and Secondary Schools, Selected Years



SOURCE: U.S. Department of Education, National Center for Education Statistics, *Digest of Educational Statistics 1990* (Washington, DC: February 1991).

are spent effectively. Holding schools and teachers "accountable" has increasingly become synonymous with increased standardized testing.

State and local governments have traditionally assumed the greatest share of elementary and secondary education tiding, as shown in figure 1-2. State funding began to exceed local funding as a percentage of the total starting in the mid-1970s, and State-mandated testing grew accordingly; 46 States had mandated testing programs in 1990 as compared to 29 in 1980.[4] Similarly, increases in Federal education spending during the 1960s and 1970s spurred increases in testing as Congress sought data to evaluate Federal programs and monitor national educational progress. The Federal Government currently spends over $20 billion per year on elementary and secondary education in programs administered by over a dozen Federal agencies.[5] Outcome-based measures of the effectiveness of educational programs-generally achievement test scores—have become key elements in the congressional appropriations and authorization process.

Contradictory demands for reevaluation of testing have been caught up in recent school reform initiatives. On the one hand, many teachers, administrators, and others attempting to redesign curricula, reform instruction, and improve learning feel stymied by tests that do not accurately reflect new educational goals. On the other hand, most leading educational measurement experts emphasize that conventional standardized tests are useful tools in gauging the strengths, weaknesses, and progress of American students.

Motivated in part by changing visions of classroom learning and by frustration with tests that many critics claim can hinder children's progress toward higher levels of achievement, many educators are turning to changed methods of testing. Some of these methods are modifications of conventional written tests; others are bolder innovations, requiring stu-

[4]OTA data on State testing practices, 1985 and 1991.

[5]U.S. Department of Education, *Digest of* *1990*, cit., footnote 1, p. 337.

Box 1-A—A Glossary of Testing Terminology

A test score is an estimate. It is based on sampling what the test taker knows or can do. For example, by asking a sample of questions (drawn from all the material that has been taught), a biology testis used to estimate how much biology the student has learned. Tests can provide valuable information about an individual's competence, knowledge, skills, or behavior. *Achievement tests are* intended to estimate what a student knows and can do in a specific subject as a result of schooling. Achievement tests and *aptitude tests are both* instruments that estimate aspects of an individual's developed abilities; they exist on a continuum, with the former being more closely tied to specific curricula and school programs and the latter intended to capture knowledge acquired both in and out of school.

Standardized tests are administered and scored under conditions uniform to all students. Although most people associate standardized tests with the multiple-choice format, it is important to emphasize that standardization is a generic concept that can apply to any testing format-from written essays to oral examinations to producing a portfolio. Standardization is needed to make test scores comparable and to assure as much as possible that test takers have equal chances to demonstrate what they know.

The word *standards* applied to tests has at least two different meanings. In the more general context it denotes goals, desirable behaviors, or models to which students, teachers, or schools should aspire. Such standards describe what optimal performance looks like and what is desirable for students to know. For example, the National Council of Teachers of Mathematics has determined that a standard for mathematics instruction is to emphasize mathematics as problem solving. The word standards, in its more technical meaning, denotes the specific levels of *proficiency* that students are expected to attain. Thus, setting a passing score for a test is equivalent to setting a standard of performance on that test.

they are based on samples of behavior, tests are necessarily imprecise: scores can vary for reasons unrelated to the individual's actual achievement. Test scores can only *describe* what skills have been mastered, but they cannot, alone, *explain* why 1earning has occurred, or prescribe ways to improve it. The fact that achievement is affected by schools, parents, home background, and other factors constrains the inferences that can be drawn about schools and programs. Test scores must be interpreted carefully.

Reliability refers to the consistency and generalizability of test data Will a student's score today be close (if not identical) to her score tomorrow? Do the questions covering a subset of skills generalize to the broader universe of skills? If tests are scored by human judges, to what extent do different judges agree in their estimations of student achievement? A test needs to demonstrate a high degree of reliability before it is used to make decisions, particularly those with high stakes attached.

Validity refers to whether or not a test measures what it is supposed to measure, and whether appropriate inferences can be drawn from test results. Validity is judged from many types of evidence, including, in the views of some experts, the consequences of translating test-based inferences into decisions or policies that can affect individuals or institutions. An acceptable level of validity must be demonstrated before a test is used to make decisions.

There are two basic ways of interpreting student performance on tests. One is to describe a student's test performance as it compares to that of other students (e.g., he typed better than 90 percent of his classmates). *Norm-referenced tests are designed to make this type* of comparison. The other method is to describe the skills or performance that the student demonstrates (e.g., he typed45 words per minute without errors). *Criterion-referenced tests* are designed to compare a student's test performance to clearly defined learning tasks or skill levels.

Performance assessment refers to testing methods that require students to create an answer or product that demonstrates their knowledge or skills. Performance assessment can take many different forms including writing short answers, doing mathematical computations, writing an extended essay, conducting an experiment, presenting an oral argument, or assembling a portfolio of representative work.

Constructed-response items are one kind of performance assessment consisting of open-ended written items on a conventional test. However, they require students to produce the solution to a question rather than to select from an array of possible answers (as multiple-choice items do).

Computer-administered testing is a generic term covering any test that is taken by a student seated at a computer. A special type of computer-administered testing is *computer-adaptive testing, which applies the* computer's memory and branching capabilities in order to adapt the test to the skill levels shown by the individual test taker as the test is taken.

SOURCE: Office of Technology Assessment 1992.

Photo credit: Bob Daemmrich

Most children in the United States take standardized achievement tests several times during elementary and secondary school, Standardized test results have become a major force in shaping public attitudes about the quality of American schools and the capabilities of American students.

dents to demonstrate their knowledge and skills through methods known as ''performance assessment. Computer technologies, video, and integrated multimedia systems add capabilities and richness not usually attainable from conventional tests, and are gaining ground in assessment as well as instruction.

These new approaches to testing have been fueled by some cognitive scientists who claim that complex thinking involves processes not easily reduced to the routinized tasks required on conventional tests. A recent report on science education, for example, argued that:

> Rather than mastering concepts, students believe that recognizing terms in a multiple-choice format is the appropriate educational goal, In the long run the impact of current modes of testing on enduring skills and strategies for learning will be inimical to reform.[6]

In contrast, many testing professionals maintain that school improvement efforts must be constructed on a solid foundation of information about what

[6]National Research Council, *Fulfilling the Promise: Biology* (Washington, DC: 1990), p. 44. Another recent report concluded that: ". . . to direct testing along a more constructive course, we must draw on richer direct evidence of knowledge and skill from information sources beyond multiple choice tests.' See National Commission on Testing and Public Policy, *Transforming America* (Chestnut Hill, MA: Boston College, 1990), p. xi; also Walter Haney and George Madaus, ''Searching for Alternatives to Standardized Tests: Whys, Whats, and Whithers,' vol.

students are learning; well-designed tests, they say, if used and interpreted properly, can provide invaluable information in a reliable, consistent, and efficient fashion. For example, standardized tests can inform policy makers by supplying trend data on the skill levels of American students. Recent analysis of data from the Iowa Tests of Basic Skills revealed that student performance improved between 1979 and 1985, even on test items designed to assess certain higher order skills, contradicting findings from other test data that improvements were limited to mechanical tasks.[7]

Measurement experts contend that these standardized tests are also useful to teachers, as tools to calibrate classroom impressions of student progress; they are viewed as one relatively efficient, albeit inexact indicator of how a given child or school system is progressing relative to students nationwide. One test author expressed a view shared by many others in the testing community:

> . . . comprehensive, survey-type standardized achievement tests have served a useful function in monitoring the achievement levels of individual pupils and the aggregate groupings of these students in terms of classrooms, buildings, and the district. . . .[8]

Common Ground

To outsiders listening in on this debate, it may appear that proponents of conventional and new forms of assessment are adversaries locked in an intractable stalemate. Closer inspection, however, reveals that testing policy is not a zero-sum game in which either existing testing or new methods win, but an arena with multiple and mutually compatible choices.

The trick is using the kind of test that is best suited to providing the desired type of information. Thus, although some activists in the debate have carved out extreme positions, most others **agree on at** least these two fundamental points:

- different forms of testing can, if used correctly, enrich our understanding of student achievement; and
- tests of any kind should be used only to serve the functions for which they were designed and validated.

On this common ground it may be possible to build genuine reform. One prominent psychologist and long-time participant in the politics and science of testing, commenting on what appears to be a rare opportunity, observed that: ''. . . our testing ecology is entirely manmade; what we made we can change.

Lessons of History

But history tempers the optimism. Since the birth of mass public education in America some 150 years ago, innovation in tests and testing has been most attractive during periods of heightened public anxiety about the state of the schools. During these periods, however, legislators and school officials feel the greatest pressure to act, and are most prone to rely on *existing* **tests as** levers of policy. Thus, researchers and policy makers involved in the painstaking process of curricular reform and new test design often find themselves at odds with those who demand quicker and more immediately noticeable action. Hence (as described in detail in ch. 4), tests have too often been used to serve functions for which they were not designed or adequately validated. Within the education policy and research community, therefore, there is an undercurrent of concern that new tests will, as in the past, be implemented before they have been validated and before their effects on learning can be understood.

For some educators the principal concern is that new tests will raise new barriers-to women, people of color, other minorities, and the economically disadvantaged, On these issues, too, caution flags are up: precisely because testing has historically been viewed as a means to achieve educational equity, tests themselves have always been scrutinized on the question of whether they do more to alleviate or exacerbate social, economic, and educational disparities (see box l-B).

---

[7]See Elizabeth Witt, Myunghee Han, and H.D. Hoover, ''Recent Trends in Achievement Tests Scores: Which Students arc Improving and on What Levels of Skill Complexity?" paper presented at the annual meeting of the National Council on Measurement in Education Boston, MA, 1990. See also Robert Linn and Stephen Dunbar, ' 'The Nation's Report Card Goes Home: Good News and Bad About Trends in Achievement, " *Kappan*, vol. **72, No. 2, October 1990, p. 132.** For a thorough analysis of trends in achievement that illustrates the importance of using multiple measures of performance, see Daniel Koretz, *Educational* (Washington, DC: Congressional Budget Office, 1986).

[8]Herbert Rudman, ' 'The Future of Testing is Now, ' vol. fall 1987, p. 6.

[9]Sheldon White, professor of psychology, Harvard University, personal communication, June 1991.

Box 1-B—Equity, Fairness, and Educational Testing

Steven Jay Gould's seminal treatise on the history of intelligence testing is dedicated to". . . the memory of Grammy and Papa Joe, who came, struggled, and prospered, Mr. Goddard notwithstanding."[1] From his very first pages, then, Gould telegraphs the deeply emotional chords struck by concepts of psychological measurement and testing. As Gould explains midway through the book, Goddard had been one of a handful of prominent American psychologists who used test data to advance racist, xenophobia, and eugenicist ideologies. Although Goddard himself later recanted,[2] in one of the more impressive turnarounds in the history of science, the atmosphere of the 1920s and 1930s gave tests ". . . the rather happy property of being a conservative social innovation. They could be perceived as justifying the richness of the rich and the poverty of the poor; they legitimized the existing social order.'

The historical misuse of intelligence tests and their achievement test" cousins-to bolster support for restrictive immigration laws, to limit college admissions, and to label children as uneducable-has left an indelible stain on the "science" of mental measurement.[4] It is no wonder that testing policy arouses the passions of Americans concerned with equal opportunity and social mobility. As in the past, those passions run in both directions: everyone may agree that testing can be a wedge, but some see the wedge forcing open the gates of opportunity while others see it as the doorstop keeping the gates tightly shut.

Consider, for example, the following excerpts, both from individuals deeply concerned with opportunities for minority and disadvantaged children:

> . . . minority youngsters who . . . are disproportionately among the poor, tend to be relegated to poor schools, or tracked out of academic courses, just as young women are not encouraged to take math and science. Therefore, the differences in the "group" scores [on the Scholastic Aptitude 'I&t].. . represent anything but "bias." Rather, the score is a faithful messenger of the unequal distribution in our country of educational resources and encouragement.[5]

> Test makers claim that the lower test scorns of racial and ethnic minorities and of students from low-income families simply reflect the biases and inequities that exist in American schools and American society. Biases and inequities certainly exist-but standardized tests do not merely reflect their impact; they compound them.[6]

---

[1] Steven Jay Gould, *The Mismeasure* Norton, 1981), dedication, p. 7.

[2] See, e.g., Carl Degler, In *of Human* (London, England: Oxford University Press,

[3] Sheldon White, "Social Implications of IQ," *Myth of Measurability,* Paul Houts (cd.) (New York, NY: Hart Publishing Co., 1977), p. 38. See also Clarence Karier, "Testing for Order and Control in the Liberal Corporate State," *IQ Controversy,* N. Block and G. Dworkin (eds.) (New York, NY: Random House, 1976), pp. 339-373. Karier's basic argument, as summarized by another historian of testing, was". . . the tests . . . were biased in terms of social class, economic, cultural, and racial background. Their use in schools served to block opportunity for the lower classes and immigrants .., [and fashion] a system of tracking in the schools tbat reinforced social inquality... ." Paul Chapman, *Schools as* York, University Press, 1988), p. 8. For opposing viewpoints see, e.g., Marl Snyderman and Stanley Rothman, *IQ* Brunswick, NJ: Transacti"on Books, 1988); Arthur Jensen, *Bias Mental* York, Free Press, 1980); or Richard Herrnstein, "IQ,' 'Atlantic *Monthly,* vol. 228, September 1971, pp. 43-64.

[4] For details on the history of achievement and intelligence testing, see ch. 4 of this report.

[5] Donald Stewart, president, College Entrance Examinati"on Board, "Thinking the Unthinkable: Standardized Testing and the Future of American Education," speech before the Columbus Metropolitan Club, Columbus, OH, Feb. 22, 1989.

[6] Monty Neill and Noe Medina, "Standardized Testing: Harmful to Educational Health," *Delta Kappan,* vol. 70, No. 9, May 1989, p. 691.

The Purpose of This Report

Federal policymakers are caught in an unenviable dilemma. On the one hand they must satisfy the growing demand for accountability, which is often expressed in terms of simple questions: Do the schools work? Are students learning? On the other hand, they must also be responsive to growing disaffection with the quality of data on which administrators rely for evaluations of programs: achievement scores are rough indicators, at best, of progress in attaining the many goals of federally funded programs. Not surprisingly, Federal evaluation requirements that place additional testing burdens on grantees and program participants often spur an interest in revising those very requirements .10 As the Federal Government has become a more prominent player in elementary and secondary education,

---

[10] For example, the Department of Education recently formed a task force to lo& into problems of testing and evaluation for the Chapter 1/Title I compensatory education program See ch. 3 of this report.

These excerpts make clear the need to specify and control the functions of testing. Both sides appear to agree that tests can be used to identify inequalities in educational opportunities.[7] But the question becomes how to use that information. Advocates of testing as a "gatekeeper" argue that ability and achievement, rather than family background, class, or the specific advantages that might accrue to students in wealthy school districts, should govern the distribution of opportunities and rewards in society. Moreover, they add, this system of distribution creates incentives for school systems to provide their students with the best possible chances for success.

On the other hand, opponents contend that ability and achievement scores are highly correlated with socioeconomic background factors[8] and with the quality of schooling children receive[9]; under these circumstances, ". . . no assessment can be considered equitable for students if there has been differential opportunity to access the material upon which the assessment is based."[10]

This debate will not be resolved easily or quickly; nor will it become moot with the advent of alternative methods of assessment. On the contrary, it could very well become even more heated and complex.[11] Educational testing policy in the United States is at a crossroads, and if history supplies any clues, the future of assessment will depend in large part on basic issues of equity, fairness, and the improvement of opportunities for minorities and the disadvantaged. The core questions are well summarized in a recent book on science assessment:

> Are we better off with the flawed system now in place or with an unknown examination system that could bring even greater problems? What differences in opportunity to learn and achieve will flow from assessment? Will it help students, teachers, or parents do something different to promote learning, for example, by moving the best teachers to the neediest students or providing summer instruction for students not at grade level at the end of the school year? And does better assessment increase our responsibility for intervention, as better technology in medicine has increased the demand and the ethical dilemmas we face in determining the use of that technology in treatment? If we are prepared to more, once we know more, perhaps the dangers of inequity possible in new assessment are worth the risk. But absent the resolve to intervene, one could argue that assessment becomes little mom than voyeurism.[12]

[7] For discussion of test bias and the effects of testing on minority students, see, e.g., Walter Haney, Boston College, "Testing and Minorities," draft monograph, January 1991, p. 24.

[8] See, e.g., Christopher Jencks et al., *Inequality* (New York, NY: Basic Books, 1972).

[9] See, e.g., Ronald Ferguson, "Paying for Public Education: New Evidence on How and Why Money Matters," *Journal on* vol. 28, No. 2, summer 1991, pp. 465-498.

[10] Shirley Malcom, "Equity and Excellence Through Authentic Science Assessment," *in the Service of Reform*, Gerald Kuhn and Shirley Malcom (eds.) (Washington, DC: American Association for the Advancement of Science+ 1991), p. 316. It is interesting to note that standardized test scores, viewed by some critics as blocking entry to education and work opportunities, have been used to justify major public programs to help minority and disadvantaged children :"... the preeminent example . . . was in the 1960s, when lower performance of minority and inner city children was used to bolster arguments for the war on poverty and to help propel passage of the Elementary and Secondary Education Act of 1%5. . . ." (Haney, op. cit., footnote p. 22.)

[11] Some minority educators, for example, fear that new assessment methods will stifle opportunities for minority students who have recently begun to do better on conventional tests. There is also uncertainty over whether or not tests should be used for placing children in remedial programs. Parents in California sued recently, not because their children were being tested but, on the contrary, because the State had followed the precedent set in the landmark *Hobson* and banned testing as a basis for diagnosing learning difficulties and placing children in remedial tracks. For further discussion of this and other legal issues, see ch. 2.

[12] Malcom, op. cit., footnote 10, p. 320.

and as the public's attitudes toward concepts of national educational goals and standards have evolved, Congress has become more involved in the testing debate.[11]

Congress has a stake in U.S. testing policy for three main reasons:

. to ensure that accurate and reliable data about American educational achievement are provided to lawmakers, program administrators, parents, teachers, test takers, and the general public;

. to ensure that the tests used to evaluate Federal education programs do not, in themselves,

[1] [1] A 1989 Gallup poll found that the majority of respondents supported the idea of national achievement standards and goals, but few supported either State or Federal intervention in the definition of those standards and goals. For discussion see George Madaus, Boston College, and Thomas Kellaghan, St. Patricks College, Dublin, "Examination Systems in the European Community: Implications for a National Examination System in the United States," OTA contractor report, April 1991.

impede progress toward program goals; and
- to ensure that tests are used fairly and do not infringe on individual rights or impose unacceptable social costs.

Congress faces a variety of decisions that could have significant and long-term effects on the scope, quantity, and quality of testing in the United States. Issues related to national testing and the role of tests in Federal education programs are already on the congressional agenda; issues regarding the rights of test takers may emerge, as they have in previous times, if new national and State tests are mandated or if the stakes attached to existing tests are raised.

This report is aimed at helping Congress:

. better understand the functions, history, capabilities, limitations, uses, and misuses of educational tests;
. learn more about the promises and pitfalls of new assessment methods and technologies; and
- identify and weigh policy options affecting educational testing.

To unravel the complexities of these topics, OTA examined technological and institutional aspects of educational testing. This summary and policy chapter synthesizes OTA's findings on tests and testing, and outlines options for congressional action. Chapter 2 examines recent changes in the uses of testing as an instrument of policy, chapter 3 covers current issues affecting the role of the Federal Government in educational testing, chapter 4 reviews the history of testing in the United States, and chapter 5 considers lessons from testing in selected European and Asian countries. The final three chapters focus on the tests themselves. Chapter 6 explains characteristics and purposes of existing educational tests, and examines the reasons new test designs seem warranted. Chapter 7 explores various approaches to performance assessment and how these methods are being implemented in schools, and chapter 8 examines the current and future roles of computers and other information technologies in assessment.

In this report, the analysis and discussion are framed in terms of the functions of testing. OTA concludes that examining the capability of various tests to meet specific objectives is the necessary first step in abating the seemingly endless controversy over the quantity and format of testing in American schools, and in laying the groundwork for new approaches.

The Functions of Testing

Educational tests have traditionally served many purposes that can be grouped into three basic functions:

. to aid teachers and students in the conduct of classroom learning;
. to monitor systemwide educational outcomes; and
- to inform decisions about the selection, placement, and credentialing of individual students.

These three functions have a common feature: they provide information to support decisionmaking. However, they differ in the kinds of information they seek and the types of decisions they can support, and test results appropriate for some decisions may be inappropriate for others.

Classroom Feedback for Students and Teachers

Teachers must constantly adapt to the behaviors, learning styles, and progress of the students in their classrooms.[12] Tests can help them organize and process the steady stream of data arising from classroom interactions. Just as physicians use body temperature, blood pressure, heart rate, x rays, and other data to form an image of the patient's health and to determine appropriate treatments, teachers can use data of various types to better manage their classes and, in some circumstances, to tailor lessons to the specific needs of individual students. Students can use information to gain sharper understanding of their strengths and weaknesses in different subjects and can adjust their study time accordingly.

Tests that can aid classroom instruction and learning need to:

- provide detailed information about specific skills, rather than global or general scores;
. be linked to content that is taught in the classroom;
. be administered frequently;
. give feedback to students and teachers as quickly as possible;

[12]For a recent analysis of the internal workings of classrooms and implications for education policy, see Edward Pauly, *What Works, What* Basic Books, 1991), especially ch.

Photo credit: Library of Congress

A student in 1943 takes her oral spelling examination after completing a written examination on the blackboard. Teachers have always used a variety of tests to help them manage their classes and evaluate student progress.

- be scored or graded to help students learn from their errors and misunderstandings, and help teachers intervene when students get stuck; and
- be based on clear and open criteria for scoring so **that** students know what to study and how they are being evaluated.

System Monitoring

How well is a school or school system performing? This is a question often **posed** from the outside, by parents, legislators, and others with particularly high stakes in the answer. As shown in chapters 2 and 4, the question is usually posed with more urgency when the impression is that the answer will be "not very well. "

Educational tests of various sorts have long been viewed as objective instruments capable of providing systematic and informed answers about the learning that takes place in schools. In an educational system as decentralized and diverse as the American one, there is a nearly insatiable appetite for evidence that all schools are providing children with a decent education. Since the mid-19th century, tests have been used to determine how much students in different schools or school districts were learning. Recent increases in Federal expenditures have stimulated new demands for system accountability.

Test scores alone cannot reveal how or why learning has occurred, or the degree to which schools, parents, the child's home background, or other factors have affected learning. When combined appropriately with other data, however, such as prior test results and children's socioeconomic status, test results can help explain-as well as describe-the outcomes of schooling.[13]

For tests to yield meaningful comparisons across schools and districts, they must:

- . be uniformly and impartially administered and scored; and
- meet reasonable standards of consistency, fairness, and validity.

In addition, to be useful system monitoring tools, these tests:

- should provide general information about achievement, rather than detailed information on specific skills;
- should describe the performance of groups of students--classrooms, schools, districts, or States-rather than individuals (thereby allowing the use of sampling methods that yield the desired information without the costly testing of every student); and
- can be administered infrequently (once or twice a year at the most).

Selection, Placement, Credentialing[14]

Tests designed to provide data about *individual* students' current achievement or predicted perform-

13 For example, recent analysis of data from close to 1,000 school districts in Texas found significant differences in student achievement scores that could be explained by variations in measures of teacher quality and other inputs. See Ronald Ferguson, "Paying for Public Education: New Evidence on How and Why Money Matters, " *Legislation*, ummer 1991, pp. 465-498; and Richard Murnane, "Interpreting the Evidence on 'Does Money Matter?' " vol. ummer 1991, pp. 457-464.

14 These three terms overlap. However, selection refers primarily to decisions about a student's qualifications for admission to schools; placement refers to decisions about qualifications of students to participate in programs within schools they attend; and credentialing (or certification) refers to decisions regarding proficiencies reached by students who have participated in programs or completed courses of study.

ance can be used for individual selection, placement, or credentialing decisions. This function of testing has a long historical tradition: the earliest recorded examples are Chinese civil service qualifying tests given in the 2nd century B.C. As discussed in greater detail in chapter 5, many European and Asian countries continue to use examinations primarily for professional and educational "gatekeeping" functions, such as certifying students as qualified to attend specialized or elite public education programs.

Placement and certification decisions are still quite commonly based on tests, even in elementary and secondary education. Minimum competency examinations are required in many States for high school graduation, for promotion from one grade to the next, or for placement in remedial or gifted programs;[15] Advanced Placement examinations are used to determine whether high school students will be given college credit and placed in advanced courses when they arrive at college; and the National Teacher's Examination is necessary for teacher licensing in 35 States.

In the United States, however, the use of tests for selective admissions decisions has been more limited than in most other countries.[16] It is rather at the end of high school, when students compete for admission to colleges and universities, that selection tests play a critical role.[17]

Some recent proposals to initiate new tests at the national level include provisions for placement and certification. One such proposal calls for a ''certificate of initial mastery, ' to be issued to graduating high school students who perform at prescribed levels on the test, and for examinations as certification criteria for completion of fourth and eighth grades.[18]

In contrast with tests used for system monitoring, tests used for selection, placement, or certification decisions must:

- provide individual student scores;
- meet particularly high standards of comparability, consistency, fairness, and validity;
- provide information that is demonstrably relevant to successful performance in future school or work situations (in the case of selection tests); and
- provide information that is demonstrably relevant to the identification of children with special needs (in the case of placement tests used for gifted and talented programs, remedial education, or other special K-12 situations).

These tests are similar to system monitoring tests with respect to the need for impartial scoring, standardized administration, generality of information, and frequency of testing.

Some proposals for a new national test or system of examinations have selection or certification as a principal function. Good tests for these purposes must undergo intensive and time-consuming development as well as careful empirical evaluation. They must be carefully and clearly validated for these intended purposes. Historically, tests used for these purposes have been the most subject to legal challenges and scrutiny (see chs. 2 and 4).

---

[15]There is widespread concern about tests being used as the principal basis for placement of children into special programs, such as ''gifted and talented" or remedial. "A major problem is getting students who obviously need it into either gifted or remedial programs when they do not meet the 'required' minimum or maximum score on the tests [to qualify for State funding]," said Jack Webber, a sixth grade teacher in Redmond, WA (personal communication September 199 1). Precise data on the numbers of schools or districts that rely on tests for these purposes, and on exactly how test data enter into those decisions, are difficult to find. Recently the New York State Commissioner of Education struck down the use of achievement tests as the sole screening criteria for placement of students in "enriched" programs. See also discussion in ch. 2.

[16]The situation has changed since the turn of century, when, e.g., ''. . . a student could not be admitted to Central [High School] without demonstrating academic competence on an entrance exam. . . .' See David Labaree, *School* (New Haven, CT: University Press, 1988), p. 50. This was not a phenomenon limited to the East Coast: rural students in Michigan and elsewhere in the Midwest needed to pass entrance examinations to gain admissions into urban high schools. Since that time, however, policies of selective admissions into public high schools have disappeared in all but a handful of special institutions, such as the Bronx High School of Science in New York.

[17]Over 3,000 colleges and universities use the Scholastic Aptitude Test (SAT) or American College Test (ACT) to aid in their selection from vast numbers of applicants, and recruits take the Armed Services Vocational Aptitude Battery (ASVAB) for placement within the military. Many private elementary and secondary schools use tests as a criterion for admission.

[18]For a summary of national testing proposals as of early 1991, see James Stedman, Congressional Research Service, 'Selected National Organizations Concerned With Educational Testing Policy,' memorandum, Feb. 8, 1991. For a more recent update and discussion of the central issues, see ''National Testing: An Overview, " *Policy*, vol. 13, Nos. 4-5, special issue, September 1991, pp. 29-35. For a critique of these proposals see also Madaus and Kellaghan, op. cit., footnote 11.

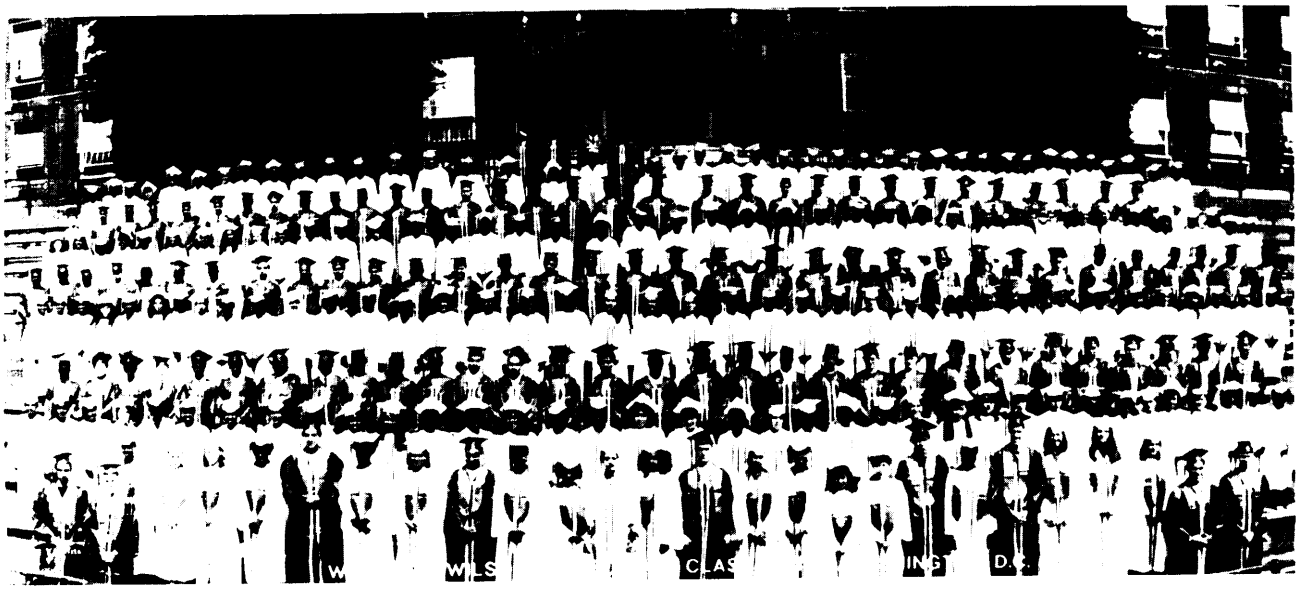


Photo credit: Panoramic Visions

The United States ranks high in the world in terms of the percentage of the population graduating from high school. These students were photographed during their 1991 graduation ceremony at Woodrow Wilson High School, a large public high school in the District of Columbia. During the 1970s and 1980s many States instituted. minimum competency testing as a criterion for graduation.

Raising the Stakes

In theory, educational tests are unobtrusive instruments of estimation. A major sticking point in any discussion of testing, however, is whether, in practice, testing affects the behavior it is intended to measure. In the current debate, advocates of new ways to test often argue that since tests can play a powerful role in influencing learning, they must be designed to support desired educational goals. These advocates disparage "teaching to the test" when a test calls for isolated facts from a multiple-choice format, but endorse the concept when the test consists of ''authentic" tasks. For these educators, one of the main criteria for a' 'good' test is whether it consists of tasks that students should practice.

More traditional measurement theorists, on the other hand, are skeptical about the value of teaching to the test because of the need to obtain valid and reliable information about the whole domain of knowledge, not just the sample of tasks that appears on the test. Thus, they argue that, regardless of a test's format, test scores are meaningless if students have practiced the tasks.

The core of the often shrill debate reflects positions on two central questions:

- Do conventional standardized tests designed to estimate student achievement negatively influence instruction and learning?
- Do new testing methods designed **to** guide instruction and learning accurately estimate student achievement?

Tests and Consequences

As the Nation's use of standardized tests has increased, the consequences attached to test results have become more serious. All but four States have standardized testing programs. Test scores are applied to a wide array of decisions affecting individual children, schools, and school systems. Students who have taken college entrance examinations, high school juniors who have failed State minimum competency tests, schools that have become lures in real estate advertisements, and States that have found themselves ranked in the national media by their average test scores are likely to remember the event—and its consequences—long afterwards.

Many educators, extrapolating from their experiences in classrooms as students or as teachers, contend that tests influence students and teachers only if they perceive that important consequences

are linked to test results.[19] But a fundamental problem arises when important consequences, or high stakes, are attached to test results; and not surprisingly, the increase in high-stakes testing over the past two decades has brought a concomitant rise in controversy. To understand the problems that can arise from high-stakes testing it is useful to consider a familiar medical metaphor.

Fever thermometers are used to measure body temperature without influencing that temperature; they provide information that could lead to treatment of the underlying conditions suspected of causing the fever. Similarly, well-designed educational tests can provide useful information to help students, teachers, or even school systems. Teachers can use tests to gauge their students' progress and decide how to "treat" children who are not doing well; students (in the upper grades especially) can review their test results to see whether they are learning the material and to determine how they might learn it more effectively; and State finding authorities can use information on the relative progress of students in different schools to develop responsive educational strategies. Thus, the information from tests can be used to choose appropriate educational ''treatments. '

Suppose, however, that patients were punished for running a high fever (or rewarded for a low one), or that doctors were rewarded for bringing down their patients' fever (or penalized if the fever remained high). They could easily take actions-cold showers, aspirin, a glass of cold beer-to "cure" the symptom but not necessarily the underlying illness. More comprehensive and appropriate treatment could be delayed or skipped. Just as temporary drops in body temperature could give misleading indications of changes in health status, fluctuations in scores from high-stakes educational tests may not reflect genuine changes in achievement. When stakes are high, a heavy emphasis is sometimes placed on specific test results, and especially on increasing scores. The symptom-low test scores—is treated without affecting the underlying condition—low achievement.

An instructive lesson about the mixed effects of high-stakes testing comes from the minimum competency testing (MCT) movement of the 1970s and 1980s (see box l-C). As described also in greater detail in chapter 2, many State legislatures pegged promotion, placement, and graduation requirements to performance on criterion-referenced tests. The underlying rationale was that extrinsic rewards and sanctions would induce students to learn the relevant material more diligently and heighten teachers' motivation to ensure that all students learned the basics before moving them ahead. It now appears that the use of these tests misled policymakers and the public about the progress of students, and in many places hindered the implementation of genuine school reforms.

More recent research seems to confirm that high-stakes testing can mislead policymakers.[20] Complicating this picture, however, is other preliminary research evidence suggesting that students may underperform on tests that bear no individual consequences at all.[21] If such distortions are occurring, they may be misleading policymakers and the general public into believing the schools are in worse shape than they really are (and into blaming the school system for a long list of social and economic problems[22]). The free-tuning knob that could adjust tests to provide just the right degree of incentive to students-enough to elicit their best *genuine* performance-has not been invented.

Test Use

One of the most vexing problems in testing policy is how to prevent test misuse, principally the

---

[19]See, for example, Lauren Resnick, professor, University of Pittsburgh, testimony before the U.S. Congress, Senate Committee on Labor and Human Resources, Subcommittee on Education, Arts, and Humani"ties, Mar.

[20]See, e.g., 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; and Thomas Haladyna, Susan B. Nolan, and Nancy S. Hass, "Raising Standardized Achievement Test Scores and the Origins of Test Score Pollution," vol. June-July 1991.

[21]See, e.g., Steven Brown and Herbert Walberg, University of Illinois at Chicago, "Motivational Effects on Test Scores of Elementary School Students, " monograph, n.d.; and Paul Burke, "You Can Lead Adolescents to a Test But You Can't Make Them Try," OTA contractor report, Aug. 14, 1991.

[22]See, e.g., Clark Kerr, "Is Education Really All That Guilty?" 10, No. 3, Feb. 27, 1991, p, 30.

Box 1 -C—The Minimum Competency Debate

The American public school system is often accused of being resistant to change. It is common to hear rhetoric accusing classrooms of being virtually indistinguishable from those of 50 years ago. In fact, though, American schools have been changing since the very inception of the common school in the early 19th century. [1] One education historian and policy analyst, citing the multiple waves of reform of curriculum, instructional methods, and classroom technology, argues that American schools are 'awash with innovation. But he questions whether these technological and institutional innovations affect the". . . core technology of the enterprise--processes of teaching and learning in classrooms and schools. "[3]

The question of whether innovation is always a good thing for schools helps frame a discussion of minimum competency testing (MCT), clearly an institutional innovation of major proportion. Its "key demand," as one commentator has written, ''. . . was that no student be given a high school diploma without first passing a test showing that he could read everyday English and do simple arithmetic. From its beginnings in a handful of school districts in the late 1970s (Denver's program actually began in 1962), MCI' spread rapidly, with the biggest expansion occurring between 1975 and 1979. By 1980,29 States had implemented legislation that required students to pass criterion-referenced examinations, and 8 more had such legislation pending.[5] Some States used the examinations to determine eligibility for remedial programs and promotions and some required it for graduation. By 1985, growth in such programs had leveled off, although 33 States were still mandating statewide MCT; 11 of these States required the test as a prerequisite for graduation.[6]

Although there is vehement debate about the effects of MCT (and of high-stakes testing in general), there is general agreement on the origins of MCT As one of its more ardent proponents has written:

> . . . this movement . . . was, in essence, a popular uprising . . . demand[ed] mainly by parents who were anguished about the fact that millions of their children were graduating from high school without the competence to go to the grocery store with a shopping list and come back with the right items and the right change. They were determined to change that, and convinced that a required exit test would produce the result they demanded.[7]

---

[1] The transition of the school system from one servicing the elites to one aspiring to universal access is described in many histories of American education. See, e.g., Ira Katznelson and Margaret Weir, *All* York, NY: Basic Books, 1985); David Tyack, *System: A* (Cambridge, MA: Harvard University Press, 1975); Michael B. Katz, (Cambridge, MA: Harvard University Press, 1%8); or Lawrence Cremin, *Progressivism* York, NY: Vintage Books, 1964).

[2] Richard Elmore, "Paradox of Innovation in Education: Cycles of Reform and the Resilience of 'leaching, " paper presented at the Conference on Fundamental Questions of Innovation Governors Center, Duke University, May 1991.

[3] Ibid. Other analysts have also addressed the innovation question in education. See, e.g., Richard Nelson and Richard Murnane, "production and Innovation When Techniques are Tacit: The Case of Education," *Economic Organization,* pp. 353-373; or Larry Cuban, *of* York, NY: Teachers College Press, 1986).

[4] Barbara Lerner, "Good News About American Education," March 1991, p. 21.

[5] Ronald A. Berk, "Minimum Competency Testing: Status and Potential," *Future Testing,* Barbara S. Plake and Joseph C. Witt (eds.) (Hillsdale, NJ: L. Erlbaum Associates, 1986), pp. 88-144.

[6] U.S. Congress, Office of Technology Assessment, "State Educational Testing Practices,' background paper of the Science, Education and Transportation Program, December 1987.

[7] Lerner, op.cit., footnote 4, p. 21. See also Douglas A. Archbald, University of Delaware, and An&w C. Porter, University of Wisconsin, Madison, "A Retrospective and an Analysis of the Roles of Mandated Testing in Education Reform," OTA contractor report, Jan. 6, 1990.

Continued on next page

application of a test to purposes for which it was not designed.[23] A familiar case of test misuse is the ranking of State school systems on a ''wall chart' displaying average scores on the Scholastic Aptitude Test (SAT) along with other data.[24] Why was this a case of test misuse? First, the SAT is designed to rank applicants from diverse educational backgrounds with respect to their likely *individual* performance as college freshmen. It is designed specifically to override differences in curricula,

---

[23] See also Burke, op. cit., footnote 21; Larry Cuban, ''The Misuse of Tests in Education, ' OTA contractor report, Sept. 9, 1991; Robert L. Linn, ''Test Misuse: Why is it so Prevalent,'' OTA contractor report, September 1991; and Nelson L. Noggle, "The Misuses of Educational Achievement Tests for Grades K-12: A Perspective," OTA contractor report, October 1991.

[24] The wall chart, now defunct, was initiated in 1984 by then Secretary of Education Terrell Bell.

Box 1-C—The Minimum Competency Debate-Continued

As with every other surge of testing in American education history,[8] MCI' was quickly shrouded in controversy. Educators and measurement specialists warned against the quick-fix mentality that exit tests could solve the problems stemming from a complex web of home, school, and societal decay; teachers lamented this new intrusion in their classrooms; and minority advocates challenged the legal and ethical basis for what appeared to be the latest obstacle to the educational and economic well-being of their children.

What have been the effects of MCI'? The research community remains divided: there is common ground that MCI' influenced education, but disagreement over whether it influenced education for the better.

Challenged to show that MCI' worked, its supporters like to point to trends in achievement test scores: the apparent improvement in literacy and numeracy among students generally, the shrinking of the gap between white and minority students, and the upturn in Scholastic Aptitude Test (SAT) scores that began in 1979. Although MCT had its most direct effects on high school juniors and seniors, proponents claim that the effect trickled down to the lower grades too, where students heard the message that they would need to work harder in order to be promoted and eventually graduate. Thus, they credit MCT even with the upturn in standardized test scores in the elementary grades.

Other analysts dismiss these conclusions. First, test scores went up even in States without MCT programs, undermininggg the causal relation between MCI' and achievement.[9] Second, even in States with MCI' where scores did go up, the timing of these events raises important questions. A **1987** congressional study noted that: ".. . most of the increase in competency testing occurred . . . several years after the upturn in achievement first became apparent in the lower grades."[10] Thereport showed that achievement scores probably began to climb beginning with fifth graders in 1975. Thus, unless one is willing to believe that tests can have virtually instantaneous effects on achievement, the timing of the rise in scores cannot be attributed to MCT. Third, the change in SAT scores beginning in 1979 reflects the general improvement in performance recorded by that cohort of test takers all through their school years, and not the advent of MCT. As one analyst put it: ". . . the higher scores rolled through the grades like a rippling wave as the elementary schoolchildren got older."[11]

Finally, what about the observed improvements in National Assessment of Educational Progress (NAEP) scores? First, NAEP scores did rise in the 1970s and 1980s, but the rise actually began as early as the 1974 assessment, well before MCI' was in operation in all but one or two States. Second, analysts point out that while test performance among Black and Hispanic 17-year-olds improved markedly during the 1970s and 1980s, it would be misleading to infer that the gap between white and Black students had disappeared: "... white students constituted the great majority of students in the two highest categories [suggesting] that there is still a substantial

[8] See ch.

[9] See Gerald Bracey, rejoinder to Barbara Lerner, *Commentary*, vo1. 92, No. 2, August 1991, p. 10.

[10] Daniel Koretz, *Educational Achievement: Explanations* *Trends* (Washington, DC: Congressional Budget August 1987), p. 84.

[11] Bracey, op. cit., footnote 9.

instruction, and academic rigor that may exist in the thousands of high schools from which applicants have graduated; by design, therefore, it does not measure a student's mastery of any given curriculum, and therefore should not be used to gauge a school's effectiveness at delivering its curriculum. Second, the SAT is taken only by about one-third of all students nationwide (with considerable regional variation), so it provides a very inadequate measure of the quality of education offered to *all the* students in a State.[25]

There is considerable professional agreement about a number of principles of good test development and appropriate test use. The primary vehicle for enforcing these principles is self-regulation by

[25] For discussion of these and other problems in using the Scholastic Aptitude Test as an indicator of State educational programs, see Cuban, op. Cit., footnote 23; and Harold Hodgkinson, "Schools are Awful-Aren't They?" *Education* vol. 11, No. 9, Oct. 30, 1991, p. A32.

gap between the reading proficiency of the average Black or Hispanic 17-year-old and the average white 17-year-old." [12] Third, there is a widespread fear that with its emphasis on basic skills, MCI' forced many schools to cut back on instruction in so-called "higher order" skills.[13]

But the debate over the effects of MCI' goes well beyond trends in test scores, which are always difficult to attribute to any single policy or intervention. Proponents look at the test scores and see a glass half full: it is, to them, a reform policy that worked for basic skills and could now be successfully applied toward the goal of teaching more children higher order skills. By and large, though, there is considerable agreement that State-mandated testing, and MCI' in particular, had unintended effects on classroom behavior of teachers and students, and that these effects should serve as a warning for any future anticipated uses of high-stakes tests.

For example, one study combined analysis of survey data and intensive interviews with teachers and school administrators, and concluded that the testing reinforced an excessive emphasis on basic skills and stymied local efforts to upgrade the content of education being delivered to all students.[14] Other studies have bemoaned the narrowing effect that MCT seems to have had on instructional strategies, content coverage, and course offerings. [15] Still other studies focus on the potentially misleading information derived from high-stakes tests: recent research suggests that improvements on high-stakes tests do not generalize well to other measures of achievement in the same domain; [16] and studies that focus in particular on teachers in districts with high-stakes testing conditions-such as minimum competency tests, school evaluation tests, or externally developed course-end tests-demonstrate a greater influence of testing on curriculum and instruction.[17]

In the end, then, there appears to be consensus that innovation in school testing policies can have profound effects-the disagreement is over the desirability of those effects. Although some of the evidence is contradictory, at times even confusing, one thing is clear: *test-based accountability is no panacea. Specific* proposals for tests intended to catalyze school improvement must be scrutinized on their individual merits.

@o& Linn and Stephen Dunbar, ''The Nation's Report Card Goes Home: Good News and Bad About Trends in Achievement' *Kappan*, vol. October 1990, p. 130. For discussion of trends in reading scores, see also John Carroll, "TheNational Assessments in Reading: Are We Misreading the Findings?" *Delta Kappan*, vol. February 1987, pp. 424430.

[13]It should be noted, however, that the empirical data on this issue are ambiguous. While the National Assessment of Educational Progress reports generally conclude that American students' higher order abilities have remained stagnant, other studies have challenged that finding. See, e.g., Elizabeth Witt, Myunghee Han, and H.D. Hoover, "Recent Trends in Achievement Tests Scores: Which Students are Improving and on What Levels of Skill Complexity?" paper presented at the annual meeting of the National Council on Measurement in Education, Bestow MA, 1990.

[14]H. D. Corbett and B. Wilson, ''Unintended and Unwelcome: The Local Impact of State Testing,'' paper presented at the annual meeting of the American Educational Research Association Boston, MA, April 1990.

[15]For review and discussion, see Archbald and Porter, op. Cit., footnote 7.

[16]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, p. 20.

[17]Claire Rottenberg and Mary Lee Smith, ''Unintended Effects of External Testing in Elementary Schools,' paper presented at the annual meeting of the American Educational Research Association, Boston, MA, April 1990.

test developers and other trained professionals.[26] Standards and codes developed by professional associations, critical reviews of tests, and individual professional codes of ethics all contribute to better testing. But, in general, few safeguards exist to prevent misuse and misinterpretation of scores, especially once they reach the public domain. Many professionals in the testing community also believe the codes lack enforcement mechanisms. Moreover, there has recently been heightened concern among test authors and publishers that market forces may interfere with good testing practice. As one test

---

[26]An example of self regulation often cited in the testing community is a decision taken by the Educational Testing Service (ETS) concerning the National Teachers Examination (NTE), which is designed to certify new teachers. When the Governor of Arkansas signed a bill in 1983 requiring teachers to pass the test in order to keep their jobs, ETS President Gregory Anrig protested: ''It is morally and educationally wrong to tell someone who has been judged a satisfactory teacher for many years that passing a certain test on a certain day is necessary to keep his or her job." ETS announced it would no longer sell the NTE to States or school boards that used it to determine the futures of practicing teachers. See Edward Fiske, 'Test Misuse is Charged, ' *York* p. Cl; also David Owen, (Boston, MA: Houghton Mifflin, 1985), pp. 243-260.

author has warned: ". . . new corporate managers . . . [are] rushing to produce tests that will ostensibly meet purposes for which the tests have never been intended. '[27]

New Testing Technologies

Educators dedicated to the proposition that testing can be an integral part of instruction and a tool for assessing the full range of knowledge and skills have given impetus to new efforts to expand the technologies, modes, formats, and content of testing. Test developers and educators are experimenting with:

- performance assessment, a broad category of testing methods that require students to create answers or products that demonstrate what they are learning, and
- computer and video technologies for developing test items, administering tests, and structuring whole new modes of content and format.

This section of the summary begins with an overview of the characteristics of these new approaches to assessment, and then considers their potential role in advancing the three basic functions of testing. It is important to remember that:

- new assessment methods alone cannot ensure consensus on what children should learn or the levels of skills children should acquire,
- curriculum goals and standards of student achievement need to be determined before appropriate assessment methods can be designed, and
- new assessment methods alone do not necessarily equip teachers with the skills necessary to change instruction and achieve new curricular goals.

Performance Assessment

The move toward new methods of student testing has been motivated by new understandings of how children learn as well as by changing views of curriculum. These views of learning, which challenge traditional concepts of curricula and teaching, also challenge existing methods of evaluating student competence. For example, it is argued that if instruction ought to be individualized, adaptive, and interactive, then assessment should share these characteristics. In general, educators who advocate

Photo credit: Educational Test/rig Service

Performance assessment covers a broad range of testing methods that require students to create answers or products to demonstrate what they are learning. In this art assessment, students record their observations as they sculpt with day; the finished product and their notes will become part of their portfolio for the year,

performance assessment believe testing can be made an integral and effective part of learning.

One type of performance assessment uses paper-and-pencil methods such as ''constructed-response items, for which students produce their own answers rather than select from a set of choices, Other approaches take performance assessment further along the continuum----from short-answers at one extreme to live demonstrations of student work at the other (see box 1-D). Under ideal circumstances, these methods share the following characteristics:

- they require students to construct responses, rather than select from a set of answers;
- they assess behaviors of interest as directly as possible;
- they are in some cases aimed at assessing group performance rather than individual performance;
- they are criterion-referenced, meaning they provide a basis for evaluating a student's work with reference to criteria for excellence rather than with reference to other students' work;
- in general, they focus on the process of problem solving rather than just on the end result;
- carefully trained teachers or other qualified judges are involved in most of the evaluation and scoring; and

[27]Rudman, op. cit., footnote 8, p. 6.

Box 1 -D—The Many Faces of Performance Assessment

Performance assessment is a broad term. It covers many different types of testing methods that require students to demonstrate their competencies or knowledge by creating an answer or product. It is best understood as a continuum of formats that range from the simplest student-constructed responses to comprehensive demonstrations or collections of large bodies of work over time. This box describes some common forms of performance assessment.

Constructed-response questions require students to produce an answer to a question rather than to select from an array of possible answers (as multiple-choice items do). In constructed-response items, questions may have just one correct answer or may be more open ended, allowing a range of responses. The form can also vary: examples include answers supplied by filling in a blank; solving a mathematics problem; writing short answers; completing figural responses (drawing on a figure like a graph, illustration, or diagram); or writing out all the steps in a geometry proof.

Essays have long been used to assess a student's understanding of a subject by having the student write a description, analysis, explanation, or summary in one or more paragraphs. Essays are used to demonstrate how well a student can use facts in context and structure a coherent discussion. Answering essay questions effectively requires analysis, synthesis, and critical thinking. Grading can be systematized by having subject matter specialists develop guidelines for responses and set quality standards. Scorers can then compare each student's essays against models that represent various levels of quality.

Writing is the most common subject tested by performance assessment methods. Although multiple-choice tests can assess some of the components necessary for good writing (spelling, grammar, and word usage), having students write is considered a more comprehensive method of assessing composition skills. Writing enables students to demonstrate composition skills-inventing, revising, and clearly stating one's ideas to fit the purpose and the audience--as well as their knowledge of language, syntax, and grammar. There has been considerable research on the standardized and objective scoring of writing assessments.

Oral discourse was the earliest form of performance assessment. Before paper and pencil, chalk, and slate became affordable, school children rehearsed their lessons, recited their sums, and rendered their poems and prose aloud. At the university level, rhetoric was interdisciplinary: reading, writing, and speaking were the media of public affairs. Today graduate students are tested at the Master's and Ph.D. levels with an oral defense of dissertations. But oral interviews can also be used in assessments of young children, where written testing is inappropriate. An obvious example of oral assessment is in foreign languages: fluency can only be assessed by hearing the student speak. As video and audio make it possible to record performance, the use of oral presentations is likely to expand.

Exhibitions are designed as comprehensive demonstrations of skills or competence. They often require students to produce a demonstration or Live performance in class or before other audiences. Teachers or trained judges score performance against standards of excellence known to all participants ahead of time. Exhibitions require a broad range of competencies, are often interdisciplinary in focus, and require student initiative and creativity. They can take the form of competitions between individual students or groups, or may be collaborative projects that students work on over time.

Experiments are used to test how well a student understands scientific concepts and can carry out scientific processes. As educators emphasize increased hands-on laboratory work in the science curriculum, they have advocated the development of assessments to test those skills more directly than conventional paper-and-pencil tests. A few States are developing standardized scientific tasks or experiments that all students must conduct to demonstrate understanding and skills. Developing hypotheses, planning and carrying out experiments, writing up findings, using the skills of measurement and estimation, and applying knowledge of scientific facts and underlying concepts—in a word, "doing science'—are at the heart of these assessment activities.

Portfolios are usually files or folders that contain collections of a student's work. They furnish a broad portrait of individual performance, assembled overtime. As students assemble their portfolios, they must evaluate their own work, a key feature of performance assessment. Portfolios are most common in writing and language arts-showing drafts, revisions, and works in progress. A few States and districts use portfolios for science, mathematics, and the arts; others are planning to use them for demonstrations of workplace readiness.

SOURCE: Office of Technology Assessment, **1992.**

. students understand clearly the criteria on which they are judged.

Computer and Video Technologies

Data processing technologies have played a significant role in shaping testing as we know it today, and could be important tools for the development of innovative tests. Computers have most commonly been used for the creation of test items and the scoring and reporting of test results. New computer and video technologies, however, used alone or in conjunction with certain types of performance assessment, offer possibilities for enhancing testing in the classroom. As computers have become more available in schools, their use for testing has become more feasible. Research in this field is showing promise in the following areas:

- questions presented and answered on computers can go beyond the traditional multiple-choice format, allowing test takers to create answers rather than select from alternatives presented to them;
- video, audio, and multimedia can make more realistic and engaging questions and tasks available;
- computer-adaptive testing can establish an individual test taker's level of skill more quickly and, under ideal conditions, more accurately than conventional paper-and-pencil testing; and
- integrated learning systems, already found in some classrooms, often come with testing embedded in the instruction and provide ongoing analysis of student progress.

Continued research combining computing power, principles of artificial intelligence, learning theory, and test design could yield significant advances in the form and content of assessment. But a set of impressive technological and economic barriers need to be surmounted: for example, the limited availability (and relatively higher cost) of hardware, compared to paper-and-pencil tests, has prevented more rapid innovation and adoption. And even with more hardware, there is no guarantee that the capacity of that hardware will be adequate to meet constantly increasing software requirements. An even greater barrier is the lack of communication between educators, test developers, and technologists in achieving a consensus on the goals of testing and in shaping a vision for technology in the service of those goals.

Using New Testing Technologies Inside Classrooms

Performance assessment is not new to teachers or students; many techniques have long been used by teachers as a basis for making judgments about student achievement within the classroom. The form and complexity can vary:

- Imagine yourself a rebel at the Boston Tea Party and write a letter describing what occurred and why.
- Complete the following five geometry proofs.
- Describe both the dramatic and situational irony in Dickens' *Hard Times,* specifically using the characters of the Teacher, Mr. McChoakumchild, and the boss businessman in Coketown, Thomas Gradgrind.

As illustrated in box l-E, what students produce in response to these testing tasks can reveal to the teacher more than just what facts they have learned; they reveal how well the student can put knowledge in context. *Well-crafted classroom performance tasks are useful diagnostic tools that can reveal where a student may be having problems with the material. They can also* help the teacher gauge the pacing and level of instruction to student responses. At their best, these tasks can be exciting learning experiences in themselves, as when a student, required to create a product or answer that puts knowledge into context, is blessed with that flash of inspiration, "Aha! I see how it all comes together now!" In addition, these tests can signal to the students what skills and content they should learn, help teachers adjust instruction, and give students clear feedback.

Much of the research about learning and cognitive processes suggest important new possibilities for tests than can diagnose a student's strengths and weaknesses. Although traditional achievement tests have focused largely on subject matter, researchers are now recognizing that ''. . . an understanding of the learner's cognitive processes-the ways in which knowledge is represented, reorganized, and used to process new information-is also needed. '[28]

[28]Robert L. Linn, ''Barriers t. New Test Design,'' *The* Redesign *of Testing for the 21st Century,* proceedings of the 1985 ETS Invitational Conference, Eileen E. Freeman (cd.) (Princeton, NJ: Educational Testing Service, 1986), p. 73

Box 1-E—Mr. Griffith's Class and New Technologies of Testing: Before and After

To understand how teaching and testing are traditionally used in the classroom, consider this fictional account of a fourth grade teacher's efforts to understand his students' progress, and the role standardized tests play in that understanding. We start with mathematics, or, as it is known in most fourth grade classrooms, arithmetic.

Mr. Griffith is working on fractions. Among the 28 children in his class, 3 raise their hand to every one of the teacher's prompts, and usually have the answer right. Some of the other children seem to be on safe ground when it comes to adding and subtracting fractions, but appear puzzled over the rules of multiplying. The majority appear lost when it comes to division. Griffith has a sense of these differences based on his constant interaction with his class, but he needs more systematic information to know how to adjust his lessons.

Before

For starters, Griffith turns to his own tests, which are tightly linked to his instructional objectives and to the material he has covered in class. He also assesses the children in other ways: he checks their workbooks, calls on them to do problems at the blackboard, poses questions and invites answers, and eavesdrops while his students work in small groups. As an experienced teacher, Griffith can synthesize his observations of children at work into fluid judgments of their strengths and weaknesses and go that next vital step of adjusting his pedagogy accordingly.

An additional source of information is the summary of statistics from last spring's administration of a nationally normed standardized mathematics test. From these data, Griffith could get a sense of how well the students in his class stack up against others in the school and even in the Nation as a whole, as measured by their performance on that test several months earlier. For example, he might find that Sarah and Jonathan, two of the three students who seem to know all the answers, scored high on the test. But he might also find that Richard, the third one, did less well than his current classroom performance would indicate. (Did he have a bad day in the spring, or did he work on his fractions over the summer?) He might also find that Noreen, another bright child in the class, did very well on the test but still gets stuck when she has to perform at the blackboard

On the whole, this test data provides information, but probably not enough for Griffith to get a complete picture of his students' learning needs or to structure his lesson plans. One problem is that a handful of his students were not even present for the spring testing, and he has no test data for them. Another problem is that the standardized test scores do not distinguish between fractions and other applications of addition and subtraction. When Griffith moves beyond fractions, there is no guarantee that the next topic on the curriculum will have been covered on the standardized test.

It is not much better with reading and writing. The children read a lot of books on their own, but the reading tests supplied by the district still give passages out of context that have no mcaning for many of the students. And, even though Griffith feels it is important to have his students do as much writing as possible, the tests are mainly questions on spelling and vocabulary. If he wants to make the children's scores look go@ and the principal happy, he has to drill his students a lot on the mechanics. Important as they are, they do not inspire much enthusiasm in either the students or, truth be known, in Griffith. But scores are important for merit pay in his district, so Griffith knows where his priorities should be.

After

Consider again the situation of Mr. Griffith, our fourth grade teacher. In the last few years, his school has gradually invested in technology. Each class now has several computers linked together in an integrated learning system (IL-S) that corresponds to the mathematics and language arts curriculum taught in his school. Money from the PTA made it possible for Griffith to purchase two additional stand alone computers and a VCR, which connect to a television that had been locked in the storage room until a few years back. Occasionally he borrows the school's video camera from the library. While he is far from considering himself a''tekkie,' Griffith took a few courses on teaching with computers and has grown pretty comfortable with their use, especially since he knows that his colleague, Mrs. Juster, a computer whiz, is just across the hall and willing to help him when he gets stuck.

Mr. Griffith finds that, as he uses these technologies for teaching, common sense requires that he use them for testing as well. Like the teaching, the testing varies. Some of the testing he does is the same as before, but made simpler by the technology. With the help of a testmaker software package, he can design his own short-answer, essay, or multiple-choice quizzes geared to the material he has been teaching. He appreciates the fact that the

Box 1-E—Mr. Griffith's Class and New Technologies of Testing: Before and After-Continued

software can automatically translate questions into Spanish, so Maria and Esteban, who recently arrived from El Salvador, can take tests with the rest of the class. The children say these tests are much easier to read than the handwritten ones he had to crank out on the school's ancient mimeograph machine. He keeps better track of their records with "gradebook" software that automatically computes and updates student averages and lets him know who is slipping in time for him to set up his little "fireside chats" with students.

But the real change has been in being able to link his testing closer to the point for instruction. Griffith has been having his students do a lot of writing on the word processor. Now he has the students pass their writing around on the computer, make comments on each other's works, and save their first drafts. They seem more comfortable making revisions, and he can grade final products that are indeed more finished. He has each student collecting their written work in electronic portfolios on disk; at the end of each semester they chose their best works and print them out for inclusion in the portfolio they take with them to the fifth grade. Some, like Regine, have a hard time deciding what is best and why. She'd like to print it all!

The mathematics they have been working on is included in the software in the ILS: same old fractions and long division--the material that Griffith has watched, over the years, turn some students off mathematics forever while others just breeze through it. But at least now he can get a better handle on where the potholes are for which children. Dana is no problem--he has already moved on to two- and three-digit long division. At the end of his work, the system prints out a report that shows he got all 10 problems in the mini-test right, and completed it in 20 minutes. Griffith makes a note to himself-"Move Dana ahead to the next unit on the program and see how he does. It's far better than having him staring out the window while I'm going over the basics with the other kids." Michelle, who did fine with multiplication, continues to have difficulty in division problems. A quick printout of the problems she missed-with the step-by-step procedure she followed-reveals that her problem lies in subtraction-she keeps forgetting principles of carrying. ''Maybe I can get Brad to work with heron some of those problems,' he thinks. "Oops, Brad is too much of a tease. Better ask Kevin instead."

Before it is time for the first grading period, Griffith prints a summary report on all the children's work. There is still a huge range in their skills, especially in mathematics. Even with the bells and whistles added in the computer programs, the curriculum can still be pretty deadly, Griffith knows. He decides to try using some of the new videos Mrs. Juster told him about as ways to get his students more interested in using mathematics to solve problems. "The one about the abandoned bell tower at the edge of town, in which the bell starts mysteriously ringing, might get their interest," he thinks. They like working in groups and digging out the clues in the video; looking for patterns and doing the mathematics to solve the problem might put some of these dry mathematics facts into context. Maybe.

While they are watching the video, Griffith plans to get Elise, a student who just came into his class yesterday from a neighboring school district, started on the computer-adaptive test she will need for placement. It looks like she is quite far behind the other students; this will give a quick picture of her abilities and can be used in determining whether she might benefit from the Chapter 1 program in the school. "Shoot, I hate to have her miss that video, though. I suppose I can see if she can stay after school and take the test. She'll miss her bus home, though, and I'll be late picking up the baby at the day care center. And then there's the video report I promised to help Lindsey, Scott, and Sherri with. They are working on a report on 'Why we need new playground equipment' and interviewing students playing in the schoolyard after school. I can see they'll need a lot of help with that! Whoever said technology makes teaching easier?"

SOURCE Fictional scenario prepared by Office of Technology Assessment 1992.

New diagnostic tests, informed by cognitive science research, may help teachers recognize more quickly the individual learner's difficulties and intervene to get the learner back on track. Similarly, computer-administered tests open up new possibilities for keeping records of a student's errors or ineffective problem-solving strategies, and for providing immediate feedback so that children can recognize their errors while still involved in thinkin g about the questions.[29]

[29] See, for example, Isaac Bejar, "Educational Diagnostic Assessment," vol. 21, No. 2, summer 1984, pp. 175-189.

Using New Testing Technologies Beyond Classrooms

Teaching has always been an art more than a science, and what works in one classroom with one teacher does not easily transfer to other classrooms with other teachers.[30] Consequently, many of 'he methods used by teachers to gauge the progress of their students and adjust their lessons are not standardized. As long as teachers can correct their judgments on a continuing and fluid basis, day by day and hour by hour, teacher experimentation with a wide range of inferential assessment methods presents no particular harm and can offer many benefits.

When judgments about student performance are moved outside the classroom, however, they must be comparable: ''. . . whatever contextual understanding of their fallibility may have existed in the classroom is gone. '[31] Using tests fairly and appropriately for management decisions about schools or students, therefore, imposes special constraints. As explained in detail in chapter 6, standardization in test administration and scoring is the first necessary condition to make test results comparable. It is precisely the recognition that individual teachers' judgments may be insufficient as the basis for crucial decisions affecting children's futures that historically has fueled public interest in standardized tests originating from outside the classroom or school.[32]

It is important to recall that the basic concept of direct assessments of student performance is not new. American schools traditionally used oral and written examinations to monitor performance. It was the pressure to standardize those efforts, coupled with the perceived need to test large numbers of children, that led eventually to the invention of the multiple-choice format as a proxy for genuine performance. Evidence that these proxies were more efficient in informing administrative decisions rapidly boosted their popularity, despite their less obvious relevance to classroom learning. *The modern performance assessment movement is based on the proposition that new testing technologies can be more direct, open ended, and educationally relevant than conventional tests, and also reliable, valid, and efficient.*

How can performance assessments and computer-based tests contribute to system monitoring and selection, placement, and credentialing decisions? A growing number of States are experimenting with answers to this question. Thirty-six States currently use writing assessments and nine others are planning to introduce writing assessment in the near future. Twenty-one States currently use other performance assessment methods including portfolios, constructed response, and hands-on demonstrations; 19 States plan to adopt some or all of these methods. Figure 1-3 shows the current geographic distribution of States using writing and other performance assessments. Some States are using sampling technologies to reduce the direct costs of performance assessments and are seeking to resolve various technical problems. Most States are using these tests in combination with the more familiar multiple-choice test.

To the extent that decisions about school resources could be based on these statewide assessments, they are potentially high stakes. Advocates maintain that performance assessments have a clear advantage over standardized multiple-choice tests, because they assess a wider range of tasks. Although these assessments do not necessarily provide different estimates of individual student progress than some conventional tests, many educators believe their advantage lies in their more obvious relevance to learning goals. The involvement of teachers in developing and scoring performance assessments is crucial to keeping them closely linked to curricula and instruction.

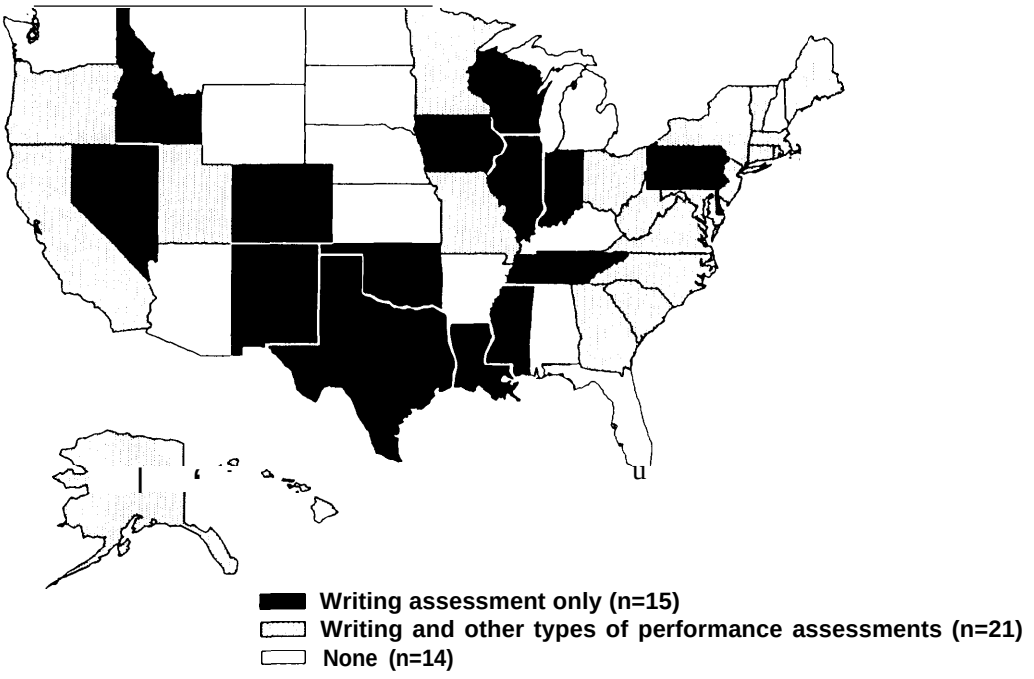
Using performance assessments beyond the confines of classrooms raises a set of important research and policy issues:

3? See Richard Murnane and Richard Nelson, ''Production and Innovation When Techniques are Tacit: The Case of Education,'' *Behavior and* vol. pp. 353-373; also Pauly, op. cit., footnote 12.

31 Stephen Dunbar, Daniel Koretz, and H.D. Hoover, ''Quality Control in the Development and Use of Performance Assessments, '' paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL, April 1991, p. 1.

32 If decisions about children's future opportunities are at stake, then the tests must also demonstrate sufficient "predictive validity, " i.e., they must provide reasonably accurate information about individual potential for future behavior in school, work, or elsewhere. For discussion of issues pertaining to the use of test scores in predicting future performance, see, e.g., Henry Levin, ''Ability Tests for Job Selection: Are the Economic Claims Justified?" *and the* B. Gifford (cd.) (Boston, MA: Kluwer, 1990); and James Crouse and Dale Trusheim, *the SAT* (Chicago, IL: University of Chicago Press,

Figure 1-3--Statewide Performance Assessments, 1991



NOTE: Chart includes optional programs.
SOURCE: Office of Technology Assessment, 1992.

- The most common form of performance assessment is the evaluation of written work: essays, compositions, and creative writing have been widely used in large-scale testing programs. Other forms of performance assessment are still in earlier stages of development and, though promising, require considerable experimentation before they can be used for high-stakes decisions.
- If performance assessment is to be successfully adopted, continuing professional development for teachers will be critical. Most teachers receive little formal education in assessment. Performance assessment may provide a great opportunity for teacher development that links instruction with assessment.
- Some parents and educators are worried that a move to greater use of performance assessment could have a negative impact on minority groups. It is critical that the issues of cultural influence and bias be scrutinized in all aspects of performance assessment: selection of tasks, administration, and scoring.
- Administration and scoring of performance assessment are both time consuming and labor intensive. If the time spent on testing is viewed as integral to instruction, however, new methods could be cost-effective.

Computer technologies, too, may play a powerful role in system monitoring and high-stakes testing of individual students. In particular:

- Adaptive testing, in which the computer selects questions based on individual students' responses to prior questions, can provide more accurate data than conventional tests, and in less time.[33]
- Advances in software could make possible automated scoring that closely resembles human scoring.
- Large item banks made possible by advanced storage technologies could lower the costs of test development by allowing State or district

[33]For discussion of the state-of-the-art in computer-adaptive testing, see Bert F. Green, The Johns Hopkins University, "Computer-Based Adaptive Testing in 1991," monograph May 9, 1991.

testing authorities to tap into common pools of questions or tasks.

. With the combination of large item banks, computer-adaptive software, and computerized test administration, tests would no longer need to be composed in advance and printed on paper; rather, each student sitting at a terminal could theoretically face a completely individualized test. This could reduce the need for tight test security, given that most students cannot memorize the many thousands of items stored in item banks.

An important policy question regarding computers in testing is whether to invest in new technologies for scarming hand-written responses to open-ended test items. Since more tests may one day be administered by computer, investing in new scanning technologies could be wasteful.

Special Considerations for System Monitoring

Performance assessments and computer-based tests could be designed to provide information on the effectiveness of schools and school systems. As with all tests, though, the outcomes of these new tests need to be interpreted judiciously: the relative performance of schools or school systems must be viewed in the context of many factors that can influence achievement.

Because individual student scores are not necessary for system monitoring, innovative sampling methods can be used that offer many important advantages for implementing performance assessments. When sampling is used, inferences can be made about a school system based on testing either a representative subsample of students or by giving each student only a sample of all the testing tasks. These methods can lessen considerably the direct costs of using long and labor-intensive performance tasks, allow broader coverage of the content areas that appear on the test, and still keep testing time limited. Furthermore, sampling methods provide important protection against misuse of a test for other functions (such as selection, placement, or certification), since students do not receive individual scores.

However, the use of sampling methods raises specific concerns: one issue is whether students' less obvious incentives to do well on such tests—given that no individual consequences are attached to performance---could lead to erroneously low estimates of aggregate achievement. A related issue is whether tests administered to samples of students will effectively signal to *all* students what they are expected to learn. A third question is whether it would be fair to administer new testing methods, intended as tools for enriched instruction, to samples of students rather than to all students.

Photo credit: IBM Corp.

Computers can change testing just as they change learning. Recent advances in computers, video, and related technologies could one day revolutionize testing.

These issues warrant further research as a prerequisite to using new testing methods for system monitoring functions.

Special Considerations for selection, Placement, and Credentialing

New testing technologies have considerable potential to enrich selection and certification decisions. For example, portfolios of student work can provide richly detailed information about progress and achievement over time that seems particularly relevant and useful for certification decisions. One example is the Advanced Placement (AP) studio art

examination, administered by the Educational Testing Service (ETS), which is based on a portfolio of student artwork. This examination is used to award college credit, and, as such, certifies that a student has mastered the skills expected of a first-year college student in studio art.

Tests based on complex computer simulations of ''on-the-job' settings are being developed for architecture, medicine, and other professions, as a basis for professional licensing and certification; the integration of graphics, video, and simulation techniques can create tests more closely resembling the actual tasks demanded by those professions. Although promising, these initial efforts have uncovered some technical issues that will require considerably more research before the tests can accurately and fairly assess the skills of interest, and be used to make high-stakes decisions about individuals.[34]

OTA has identified the following central policy issues concerning the design of new tests for selection, placement, and certification.

Technical requirements-These tests must meet very high technical standards. Inferences drawn from them must be based on rigorous standards of empirical evidence not necessarily required of tests used for other functions. Because tests used to select, place, or certify individuals can have potentially long term and significant consequences, their uses need to be limited to the specific functions for which they are designed and validated. Similarly, because test scores are only estimates, very high levels of reliability, or consistency, must be demonstrated for the test as a whole. Finally, because of the amount of day-to-day variability in individuals, no one test score should be used alone to make important decisions about individuals.[35]

Generalizability—Another issue pertains to the content coverage of new assessment formats, such as exhibitions, portfolios, science experiments, or computer simulations. The advantage of these formats is in their coverage of relevant factors of performance and achievement; however, this usually means that only a few such long and complex tasks can be completed by a single child in the allotted time.[36] Are inferences about achievement made on the basis of just a few tasks generalizable across the whole domain of achievement? When each child can complete only a few tasks, there is a much higher risk that a child's score will be specific to that particular task. Selection and certification decisions cannot be made on the basis of these tasks unless results are stable and generalizable.

Security--Currently most high-stakes selection, placement, or certification tests are multiple-choice, and precautions are taken to keep items secret. Scores would be suspect if some (or all) test takers knew the items in advance.[37] Given the relatively low number of performance-based tasks that might appear on some new tests, sharing of information from one cohort of test takers to another could become a problem undermining the test's validity. Computers with enough memory to accommodate very large item banks may provide some technological relief, although the question remains open as to whether a sufficient number of different items could be written at reasonable cost.

Fairness-Most previous legal challenges have targeted tests used to make significant decisions about individuals. Any test designed for selection, placement, or certification will be carefully scrutinized by those concerned with equity and bias. Designing a performance-based selection or certification test will require considerable research to ensure elimination of bias.

---

[34] See, for example, David B. Swanson, John J. Norcini, and Louis J. Grosso, "Assessment of Clinical Competence: Written and Computer-Based Simulations," pp. 220-246.

35& additional reason for insisting on high standards is that high-stakes tests can lead inadvertently to the labeling of individual-by themselves or by others-with uncertain and potentially harmful consequences. For discussion of these issues see, e.g., U.S. Congress, Office of Technology Assessment, "The Use of Integrity Tests for Pre-Employment Screening," background paper of the Science, Education, and Transportation Program, September 1990.

[36] Increasing the time allotted to assessment does not necessarily imply reduced time for instruction, as long as the two activities are well integrated. But completely ''seamless' * integration of testing and instruction could raise problems of its own, such as potential infringement of students' rights to know whether they are being tested and for what purposes.

[37] The concept of ''test openness'' is controversial. Most traditional measurement experts argue that allowing students access to test items in advance would irreparably compromise the test's validity. For opposing viewpoints, however, see, e.g., Judah Schwartz and Katherine A. Viator (eds.), A Report to the Ford Foundation (Cambridge, MA: Harvard Graduate School of Education, September 1990); and John Fredrickson and Alan Collins, "A Systems Approach to Educational Testing,'' December 1989, pp. 27-32.

Cost Considerations: A Framework for Analysis

A common challenge posed to advocates of alternative assessment methods is an economic one: can they be administered and scored as efficiently as conventional standardized tests?[38] Indeed, one of the attractive features of commercially published standardized tests is their apparently low cost. As shown in box l-F, OTA estimated outlays for standardized testing in a large urban school district were approximately $1.6 million for 1990-91 ($0.8 million per test administration), or only about $6 per student per test administration

But these outlays on contracted materials and services and district testing personnel do not tell the whole story. First, they neglect the dollar value of teacher time devoted to test administration. Because a teacher's many activities are not typically itemized on a school district budget, the costs associated with teacher time spent administering tests are less obvious than other testing expenses. But they can be significant: in the district studied by OTA, the portion of total teacher salaries attributable to time spent administering tests was roughly $1.8 million per test, or $13 per pupil.

Another important component of cost is the time spent by teachers in test preparation. This factor is more variable than administration time and is more difficult to estimate. It depends largely on the degree to which teachers can distinguish their regular instruction from classroom work that is driven by the need to prepare students for specific tests. The question is whether the test preparation activities would take place even in the absence of testing: this issue hinges partly on test content—how closely does the test reflect curricular and instructional objectives?—and partly on how individual teachers allocate their classroom time across various activities, including test-related instruction. (Tests that are intended to be linked to instruction might not be perceived as such by some teachers, and tests that are apparently separate from regular instruction could be useful tools in the hands of other teachers.) In the district OTA studied, teachers reported spending anywhere from O to 3 weeks in preparing their students for each test administration-at a cost as high as $13.5 million per test, or close to $100 per pupil.[39]

Just as counting material and testing personnel outlays alone can lead to deceptively low estimates of the total resources devoted to testing, accounting fully for teacher administration and preparation time can lead to deceptively high cost estimates. To correctly account for teacher time requires attention to the indirect or *opportunity costs* of that time. An opportunity cost is defined generally as ". . . the value of foregone alternative action. "[40] With respect to testing, analysis of opportunity costs focuses attention on the following question: to what extent does the time spent by teachers on preparation and administration of tests contribute to the core classroom activities of teaching and learning?

If testing is considered integral to instruction, then teacher time spent on preparing students and on administering the tests has lower opportunity costs than if the testing has little or no instructional value. To *estimate the opportunity costs, then, requires information or assumption about the degree to which any particular test is intended as an instructional tool, and information or assumptions about the extent to which individual teachers use testing as part of their instructional program.*

As shown in box l-F, some teachers in the district OTA studied spent as much as 3 weeks preparing students for each of the two standardized tests, plus 4 days administering each test. The worst case would be one in which this time was completely irrelevant to coursework: the district would have incurred steep opportunity costs—about $15 million per test, or close to $110 per pupil. The best case, in which all preparation time was relevant to coursework, would have cost under $2 million per test, or $13 per pupil.

Thus, the total costs of a testing program consist of both direct and opportunity components: direct expenditures on materials, services, and salaries, and

[38]The efficiency advantages of standardized multiple-choice tests are discussed in several places in this report. See especially ch. 4 fOr a historical synopsis, ch. for general discussion of item formats, and ch. 8 for review of technological change in test scoring and administration.

[39]A full accounting of direct costs would also include overhead on the school building and grounds, i.e., depreciation attributable to time spent on test preparation and administration. To simplify the analysis, OTA omitted this element.

[40]David W. Pearce (cd.), 3rd ed. (Cambridge, MA: MIT Press, 1986), p. 310.

Box 1-F-Costs of Standardized Testing in a Large Urban School District

Because testing policy decisions are still primarily made at the local and State levels, OTA has analyzed the kind of data on standardized testing costs that school authorities would likely include in their deliberations over testing reform. Data for this illustrative example were provided by the director of Testing and Evaluation in a large urban school district with 191,000 enrolled students, among whom 32 percent are in Chapter 1 programs. The district employs 12,000 teachers, including regular classroom and special teachers. The total 1990-91 district budget was $1.2 billion.

Approximately 140,000 students in grades kindergarten through 12 take tests, once a year in kindergarten and twice a year (fall and spring) in all other grades (absenteeism and student mobility account for the large number of untested students). During each test administration, students take separate tests in English, mathematics, social studies, and science. The tests typically consist of norm-referenced questions supplemented with locally developed criterion-referenced items. (In kindergarten, first, second, and third grades, criterion-referenced checklists filled out by teachers supplement the Paper-and-pencil tests.) The tests are machine scored by the test publisher, who provides computer-generated score reports to district personnel. Tests are administered by 4,500 regular classroom teachers; them are no other special personnel involved, except for a small group of district staff who design the criterion-referenced items, manage the overall testing program, and conduct research based on test results.

Although the district purchases tests from a large commercial publishing company that has many school districts as customers, the cost figures discussed below are not necessarily representative of other school districts in the United States.

Materials and Services

Inmost years, the district purchases only a limited supply of test booklets, replacing the complete set only once every few years when they become damaged or when test items are revised OTA computed average annual expenditures on test booklets based on test publishers' estimates that booklets are recycled typically once every 7 years. As shown in table 1-F1, total annual outlays for the standardized testing program in 1990-91--including materials, contracted scoring and reporting services, and nonteaching personnel-were approximately $1.6 million, or $5.70 per student per test administration.[1]

Teacher Time

Based on the specified time allotments for the various tests in the various grades, and on conversations with district staff, OTA found that Ml-time teachers in the district spend roughly 2 percent of their annual work time in the *administration* of tests to students. The total salary cost to the district for teacher time spent administering tests was roughly $3.6 million for two testing administrations ($1.8 million per testing cycle).

Table 1-F1--Outlays on Materials, Services, and Personnel

	Cost
Materials	
Contracted:	
Test booklets: new purchases plus annualized costs based on assumed 7-year cycle	$369,000
Practice books	49,400
Examiner manuals	26,200
Checklists and worksheets	100,600
Kindergarten program	33,300
other :	
Kindergarten Chapter 1 tests	$3,000
Labels	1,200
Pencils	17,900
Answer sheets	23,000
Headers	2,700
Language battery	1,300
Special tests	14,100
Materials subtotal	$641,700
services	
Contracted:	
scoring	$175,600
Report generation	141,800
Collection	14,800
Scanning	146,500
Distribution	9,000
Services subtotal	$487,700
Nonteaching personnel:	
Assistant director	$58,200
Research manager	56,500
Research associates (2)	108,700
Research assistants (3)	127,800
Secretaries	56,500
Clerks	45,600
Nonteaching personnel subtotal	$453,300
Total	$1,582,700

SOURCE: Office of Technology Assessment, based on data supplied by a large urban school district, 1990-91 academic year.

[1] To understand how this district's cost of standardized testing compares with others, OTA looked at cost data from the November 1988, "Survey of Testing Practices and Issues," conducted by the National Association of Test Directors (NATD). The survey was sent to testing directors in approximately 125 school districts. For 38 districts providing their cost information, the average direct cost per student was $4.80 per year, slightly lower than the $5.70 per student in this example. Most of the districts responding to the NATD survey administer achievement tests only once a year, compared to OTA's example district, which tests twice a year in grades 1 to 12.

In conversations with district teachers, OTA found that the time they spend in classroom *preparation* of students for the standardized tests varies from O to 3 weeks per testing administration. Some teachers *claim they* spend no time doing test preparation that is distinguishable from their regular classroom instruction; others use the standardize test as a final examination and offer students the benefit of lengthy in-class review time. OTA therefore estimated the salary costs for preparation time under three scenarios: O, 1.5, and 3 weeks (per test). These estimates are summarized in table 1-F2.

Table 1-F2--Salary Costs of Teacher Time Spent on Testing, per Test Administration[a]

Test administration[b]	Test preparation	Total[c]
$1.8 million	0 weeks: 0	$1.8 million
	1.5 Weeks: [d]$7.2 million	9.0 million
	3 weeks: $13.5 million	15.3 million

[a]Based on average salary of $40,500 per year.
[b]Based on an estimated 2 percent of total time spent on test administration for two testing periods.
[c]Based on teachers.
[d]8 days.

SOURCE: Office of Technology Assessment, based on data supplied by a large urban school district, 1990-91 academic year.

Total **Direct Costs**

The **total direct costs** of testing can be computed by adding the expenditures on materials and services to the costs of teacher time for test preparation and administration. It is important to note, however, that this analysis does not account for the degree to which teacher time spent on testing is considered to be a necessary and well-integrated part of regular instruction. The importance of indirect or opportunity costs as it pertains to the analysis of testing costs is illustrated in box l-G.

indirect costs of time spent on testing activities.[41] For a graphical exposition of this concept, see box 1 G.

Federal Policy Concerns

Several proposals now pending before Congress could fundamentally alter testing in the United States. Three issues already on Congress' agenda are proposals for national testing, changes to the National Assessment of Educational Progress (NAEP), and revisions to the program that assists educationally disadvantaged children (Chapter 1). Federal action could also focus on ensuring the appropriate use of tests, and speeding research and development on testing.

These policy opportunities combined with the current national desire to improve schooling provide Congress with an opportunity to form comprehensive, coordinated, and far-reaching test policy. Rather than allowing test activity to occur haphazardly in response to other objectives, decisionmakers can bring these several concerns together in support of better learning.

National Testing

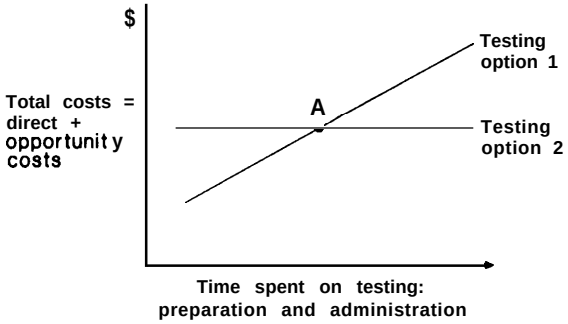
As discussed in chapter 3, the past year has witnessed a flurry of proposals to establish a system of national tests in elementary and secondary schools. Momentum for these efforts has built rapidly, fueled by numerous governmental and commission reports on the state of the economy and the educational system; by the National Goals initiative of the President and Governors; by casual references to the superiority of examination systems in other countries (see box l-H); and most recently by the President's "America 2000" plan.

The use of tests as a tool of education policy is fraught with uncertainties. *The first responsibility of Congress is to clarify exactly what objectives are attached to the various proposals for national testing, and how instruments will be designed, piloted, and implemented to meet these objectives. The* following questions warrant careful attention:

- If tests are to be somehow associated with national standards of achievement, who will participate in setting these standards? Will the content and grading standards be visible or invisible? Will the examination questions be

[41]In addition to teacher time, there are opportunity costs associated with student time: assuming that instruction time is an investment with economic returns, student time spent on testing can be valued in terms of foregone future income. This follows a ''human capital' investment model of education. See, e.g., Gary Becker, *Human* 2nd ed. (New York, NY: National Bureau of Economic Research, 1975). For application of the concept of indirect costs to educational testing see also Walter Haney, George Madaus, and Robert Lyons, Boston College, "The Fractured Marketplace for Standardized Testing,'' unpublished manuscript, September 1989.

Box 1-G—Direct and Opportunity Costs of Testing



This figure illustrates the relationship between time spent on testing activity and the total costs of testing. Hypothetical test 1 is assumed to contribute little to classroom learning. It costs little in direct dollar outlays, but is dear in costs. Total costs begin relatively low but rise rapidly with time devoted by teachers and students to activities that take them away from instruction.

Hypothetical test 2, which is a useful instruction and learning tool, requires relatively high direct expenditures. But the opportunity costs of time devoted to testing are relatively low.

At point A, a school district would be indifferent between the two testing programs, if cost was the main consideration.

SOURCE: Office of Technology Assessment, 1992.

kept secret or *will they be* disclosed after the test?

- If the objective of the test is motivational, i.e., to induce students and teachers to work harder, then the test is likely to be high stakes. What will happen to students who score low? What resources will be provided for students who do not test well? What inferences will be made about students, teachers, and schools on the basis of test results? What additional factors will be considered in explaining test score differences? Finally, will the tests focus the attention of students and teachers on broad domains of knowledge, as desired, or on narrower subsets of knowledge covered by the tests, as often happens?
- If the Nation is interested in using tests to improve the qualifications of the American work force, how will valuable nonacademic skills be assessed? What should be the balance of emphasis between basic skill mastery and higher order thinking skills?
- If there is impatience to produce a test quickly, it is likely to result in a paper-and-pencil machine-scorable test. What signal will this give to schools concerning the need to teach all students broader communication and problem-Solving skills?
- What effects will national tests have on current State and local efforts to develop alternative assessment methods and to align their tests more closely with local educational goals?
- Would the national examinations be administered at a single setting or whenever students feel they are ready?
- Would students have a chance to retake an examination to do better?
- Would the tests be administered to samples of students or all students?
- At what ages would students be tested?
- What legal challenges might be raised?

If a test or examination system is placed into service at the national level before these important questions are answered, it could easily become a barrier to many of the educational reforms that have been set into motion, and could become the next object of concern and frustration within the American school system.

Given that a national testing program could be undertaken through State and/or private sector initiatives, the role of Congress is not yet entirely clear. However, to the extent that congressional action regarding NAEP, Chapter 1, and appropriate test use will affect the need for and impact of any national examinations, Congress has a strong interest in clarifying the purposes and anticipated consequences of such examinations. Also, Congress must carefully analyze the pressures the national test movement is exerting on these programs, such as the idea of converting NAEP into a national test for all students.

Future of the National Assessment of Educational Progress

NAEP has proven to be a valuable tool to track and understand educational progress in the United States. It was created in 1969 and is the only regularly conducted national survey of educational achievement at the elementary, middle, and high

Box 1 -H—National Testing: Lessons From Overseas'

The American educational system has a traditional commitment to pluralism in the definition and control of curricula as well as the fair provision of educational opportunities to all children. Lessons from European and Asian examination systems, which have historically been geared principally toward selection, placement, and credentialing, need to be considered judiciously. OTA finds that the following factors should be considered when comparing examination systems overseas with those in the United States:

- Examination systems in almost every industrialized country are in flux. Changes over the past three decades have been quite radical in several countries. Nevertheless, there is still a relatively greater emphasis on tests used for selection, placement, and certification than in the United States.
- None of the countries studied by OTA has a single, centrally prescribed examination that is used for all purposes--classroom diagnosis, selection, and school accountability. Most examinations overseas are used today for certifying and sorting individual students, not for school or system accountability. Accountability in European countries is typically handled by a system of inspectors charged with overseeing school and examination quality. Some countries occasionally test samples of students to gauge nationwide achievement.
- External examinations before age 16 have all but disappeared from the countries in the European community. Primary certificates used to select students for secondary schools have been dropped as comprehensive education past the primary level has become available to all students.
- The United States is unique in the extensive use of standardized tests for young children. Current proposals for testing all American elementary school children with a commonly administered and graded examination would make the United States the only industrialized country to adopt this practice.
- There is great variation in the degree of central control over curriculum and testing in foreign countries. In some countries centrally prescribed curricula are used as a basis for required examinations (e.g., France, Italy, the Netherlands, Portugal, Sweden, Israel, Japan, China and, most recently, the United Kingdom). Other countries are more like the United States in the autonomy of States, provinces, or districts in setting curriculum and testing requirements (Australia, Canada, Germany, India, and Switzerland).
- Whether centrally developed or not, the examinations taken during and at the end of secondary school in other countries are not the same for all students. Syllabi in European countries determine subject-matter

[1]This draws on information from George Madaus, Boston College, and Thomas Kellaghan, St. Patricks College, Dublin, "Student Examination Systems in the European Community: Lessons for the United States," OTA contractor report, June 1991.

next

school levels. It was designed to be an educational indicator, a barometer of the Nation's elementary and secondary educational condition. NAEP reports group data only, not individual scores.

NAEP has also been an exemplary model of careful and innovative test design. As discussed in chapter 3, NAEP has made pioneering contributions to test development and practice: "matrix" sampling methods, broad-based processes for building consensus about educational goals, an emphasis on content-referenced testing, and the use of various types of open-ended items in large-scale testing.

If Congress wishes to develop a new national test—to be administered to each child and used as a basis for important decisions about children and schools—OTA concludes that NAEP is not appropriate. This objective would require fundamental redesign and validation of NAEP, and would alter the character and value of NAEP as the Nation's independent gauge of educational progress. It would also greatly increase both the cost and time devoted to NAEP at every level.

A better course for Congress is to retain and strengthen NAEP's role as a national indicator of educational progress. To do this, Congress could:

- require NAEP to include more innovative items and tasks that go beyond multiple choice;
- fund the development of a clearinghouse for the sharing of NAEP data, results of field trials, statistical results, and testing techniques, giving States and local districts involved in the design of new tests better access to the lessons from NAEP;
- restore funding for NAEP testing in more subject areas, such as the fine arts;

Box 1-H—National Testing: Lessons From Overseas--Continued

content and examinations are based on them, organized in terms of traditional subject areas (language, mathematics, sciences, history, and geography) and, in some cases, levels at which the subject is studied (general or specialized). Even in European Community (EC) countries with a national system, the examinations are differentiated: all students do not take the same examination at the same time. The examinations may also be differentiated by locale (depending on the part of the country) or by track (there are high-level, low-level, and various curricular options).

- With differentiated examinations, multiple options give students on lower tracks the chance to choose lower level examinations. It appears, though, that these school-leaving examinations can discourage students who do not expect to do well from staying in school.
- In no other system do commercial test publishers play as central a role as they do in the United States. In EC and other industrialized nations, tests are typically established, tinted, and scored by ministries of education, with some local delegation of authority. In Europe, Japan, and the U.S.S.R. the examinations have traditionally been dominated by and oriented toward the universities. In Europe, most examination systems are organized around a system of school inspectors, with quasi-governmental control through the establishment of local boards, or multiple boards in larger countries.
- Psychometrics does not play a significant role in the design or validation of tests in most European and Asian countries. Although issues of fairness and comparability are important, they are treated differently than in the United States.
- Teachers in other countries have considerable responsibility for administering and scoring examinations. In some countries (Germany, the U.S. S.R., and Sweden) they even grade their own students. Teacher contracts often include the expectation that they will develop or score examinations; they are sometimes offered extra summer pay to read examinations.
- Syllabi, topics, and even sample questions are widely publicized in advance of examinations, and it is not considered wrong to prepare explicitly for examinations. Annual publication of past examinations strongly influences instruction and learning.
- In European countries, the dominant form of examination is "essay on demand." These examinations require students to write essays of varying lengths in responses to short-answer or open-ended questions. Use of multiple-choice examinations is limited, except in Japan, where they areas prevalent as in the United States. Oral examinations are still common in some of the German *lander* and in foreign language testing in many countries. Performance assessments of other kinds (demonstrations and portfolios) are used for internal classroom assessment.

- support the continued development of methods to communicate NAEP results to school officials and the general public in accurate and innovative ways (particular emphasis could be placed on informing the public about appropriate ways to interpret and understand such test data and on minimizing misinterpretation by the press and general public);
- add testing of nonacademic skills and knowledge relevant to the world of work;
- restore funding for the assessment of out-of-school youth at ages 13 and 17, to provide a better picture of the knowledge and skills of an entire age cohort;
- request data on the issues surrounding test-takers' motivation to do well on NAEP in various grades;[42]
- expand NAEP to assess knowledge in the adult nonschool population; and
- ensure that matrix sampling is retained, to minimize both costs and time requirements of NAEP.

An experiment in extending the uses of NAEP to provide data on educational progress at the State level and to measure this progress against national standards is now under way.

OTA has identified three potential problems of using NAEP for State-by-State comparisons that

[42]In particular, questions have been raised about the accuracy of information derived from tests of 12th graders who are about to graduate. Further trial efforts and research could shed light on this issue. Ed Roeber, Michigan Educational Assessment Program, personal communication, October 1991.

Photo credit: National *Assessment of Educational Progress*

The **National Assessment of Educational Progress (NAEP) has pioneered the use of performance assessments in large-scale testing programs. In this science task, 7th and 11th grade students figure out which of the three materials would make the box weigh the most.**

Congress should review before making a final decision on a permanent use of NAEP for this purpose. First, States could be pressured to introduce curriculum changes to improve their NAEP performance on certain subjects, regardless of whether such changes have educational merit. For example, following the release in 1991 of the State-by-State results from the first such trial, some States (e.g., the District of Columbia) announced plans to revamp their mathematics curricula. It could be argued that the use of NAEP as a prod to State education authorities to rethink their curricula is a good thing; however, *it is clear that the pressure to perform on the test can outweigh the stimulus for careful deliberation about academic policy, and that many States could make changes for the sake of higher scores rather than improved learning opportunities for children.* This signifies putting the cart of testing before the horse of curriculum, exactly the kind of outcome feared by the original designers of NAEP who insisted that scores not be reported below broad regional levels of aggregation.

Second, the presentation of comparative scores could lead to intensified school-bashing--even when differences in average State performance are statistically insignificant or when those differences reflect variables far beyond the control of school authorities. Critics of comparative NAEP reporting point out that low-scoring States need real help--financial, organizational, and educational-not just more testing and public humiliation.

Finally, extending NAEP to State-level analysis and reporting is a costly undertaking. NAEP funding jumped from $9 million in 1989 to $19 million in 1991. It is not clear that this extra money provides a proportional amount of useful information: one researcher interested in this question showed that roughly 90 percent of the variance in average State performance on NAEP could be explained by socioeconomic and demographic variables already available from other data.[43] In a time of scarce educational resources, NAEP extensions need to be weighed carefully on the scale of anticipated benefits per dollar. State-by-State comparisons of NAEP performance may not pass this cost-benefit test.[44]

These issues notwithstanding, many education policymakers at the State and national levels have insisted that State-level NAEP could provide new and useful information to support curricular and instructional reform. Their arguments should be taken as potentially fruitful research hypotheses and treated as such: just as new medical treatments undergo careful experimentation and evaluation before gaining approval for general public use, extensions and revisions to NAEP should be postponed pending analysis of research data.

In education, the line between research and implementation is often blurred; few newspapers noted that the 1990 State mathematics results were the first in a "trial" program-the results were

[43] See Richard Wolf, Teachers College, Columbia University, ''What Can We Learn From State NAEP?' unpublished document, n.d.

[44] See also Daniel Koretz, ''Shte Comparison Using NAEP: Large Costs, Disappointing Benefits, ' vol. 20, No. 3, April 1991, pp. 19-21.

treated as factual evidence of relative effectiveness of State education systems.

The NAEP standard-setting process also raises questions of feasibility and desirability. As discussed in chapter 6, the translation of broad educational goals-such as emphasizing problem-solving skills in the mathematics curriculum--into specific test scores is a complex and time-consuming task. The particular performance standards selected must be validated empirically: how closely educators in different parts of the country will concur on standards of proficiency for children at different stages of schooling is not known. Standard setting has always been a slippery process—in employment, psychological, or educational testing-in large part because of difficulties surrounding the designation of acceptable ''cutoff scores. Not surprisingly, controversy surrounded the initial attempts to reach consensus on standards for NAEP, with experts disagreeing among themselves on key definitions and interpretations of items.

Educators and policymakers continue to debate whether nationwide standards are desirable, especially if children who do not reach the defined standards are somehow penalized. In addition to the potential effects on children, turning NAEP into a higher stakes test—with implicit and explicit rewards pegged to achievement of the given proficiency standards--could irreparably undermine NAEP's capacity as a neutral barometer of educational progress.

While continued research on State-by-State NAEP and on standard setting will be useful, Congress needs to find ways to ensure that data from this research are reported as such and that the results are not prematurely construed as conclusive.

Chapter 1 Accountability

Because of its scope and influence, Chapter 1 represents a powerful lever by which the Federal Government affects testing practices in the United States. OTA's analysis of Chapter 1 testing and evaluation requirements (see ch. 3) suggests several congressional policy options that could improve Chapter 1 accountability while reducing the overall testing burden in the United States.

Chapter 1, the largest Federal program of aid to elementary and secondary education, provides supplementary education services for disadvantaged children. Over its 25-year history, Chapter 1 evaluation and assessment requirements have been revised many times. The result is an elaborate web of legal and regulatory requirements with standardized norm-referenced achievement tests as the basic thread. The tests fulfill several functions: Federal policymakers and program a administrators use nationally aggregated scores to judge the program's overall effectiveness; and local school districts and States use scores to determine which schools are not making sufficient progress in their Chapter 1 programs, to place children in the program, to assess children's educational needs, and for other purposes.

As a result of the 1988 amendments to Chapter 1, which introduced the "program improvement" concept, Chapter 1 testing became even more critical. At the national level, there has been growing concern that the aggregated test data-collected by school districts with widely divergent expertise in evaluation-do not provide an accurate and well-rounded portrait of the program's overall effectiveness. At the school district level, educators argue that the test data often target the wrong schools for program improvement or miss the schools with the weakest programs in the district or the subject areas and grade levels most in need of help. At the classroom level, teachers tend to feel that their own tests and assessments, as well as some externally designed criterion-referenced tests, afford a much better picture of individual students' progress than do the norm-referenced tests.

Congress' principal challenge vis-á-vis Chapter 1 is to find ways to separate Federal evaluation needs from *State and local needs.* It is a tough dilemma: to balance the national desire for meaningful and comparable program accountability data against State and local needs for useful information on which to base instructional and programmatic decisions. Congress will consider reauthorization of Chapter 1 in 1993. Hearings and analysis on these complex questions in 1992 would provide an excellent basis for a major revision of the evaluation and testing requirements.

One way to improve Chapter 1 accountability is to create a system that separates national evaluation needs from State and local information needs. It is the perceived need for nationally aggregated data that drives the use of norm-referenced tests. If Congress separated national

evaluation purposes from State and local purposes and articulated different requirements for each, State Education Agencies (SEAS) and local education authorities would be free to use a variety of assessment methods that better reflect their own localized Chapter 1 goals. The national data would be used to give Federal policymakers, taxpayers, and other interested groups a national picture of Chapter 1 effectiveness, while the State and local information would be used in modifying programs, placing students, targeting schools for program improvement, deciding on continuation of schoolwide projects, and other purposes.

Congress could obtain national data on Chapter 1 through a well-constructed, periodic testing of Chapter 1 children, similar to the way NAEP is used to assess the progress of all students. This assessment would rely on sampling (rather than testing of every student) and could be administered less frequently than the current tests. In addition to relieving the testing burden on individual students and reducing the time devoted to testing by teachers, principals, and other school personnel, this procedure could also result in higher quality data. As the principal client of the data, the Federal Government could identify the areas to be assessed, instill greater standardization and rigor in test administration and data analysis, and avoid the aggregation problems that arise from thousands of school districts administering different instruments under divergent conditions. This type of Federal assessment could be designed and administered by either an independent body or the Department of Education, with the help of the Chapter 1 Technical Assistance Centers.

The system might be designed to provide a menu of assessment options--criterion-referenced tests, reading inventories, directed writing, portfolios, and other performance assessments—from which States could establish statewide evaluation criteria for Chapter 1 programs. If Congress preferred maximum local flexibility, the discretion to choose among the assessment options could be left to school districts, as long as they administered the instruments uniformly and consistently across schools. The Chapter 1 Technical Assistance Centers could help the States and school districts select and implement appropriate measures.

Either a State or local option would increase the latitude for linking assessments to specific program goals. However, if States or districts were to select instruments that put their Chapter 1 programs in the best light, the information could be misleading. Congress should take steps to see that this does not happen. For example, a strict approach would require programs to show growth in student achievement using multiple indicators, perhaps including one indicator based on a standardized test, A looser version of this option would allow States or districts to develop their own evaluation methods, and set their own standards of acceptable progress, subject to Department of Education approval.

An advantage of separating evaluation requirements would likely be local development of new testing methods, which have not been widely used in Chapter 1 because of the need for national aggregation and comparability. Congress could encourage this choice by reserving some of the Federal Chapter 1 evaluation and research funding to advance the state of the art.

For example, competitive grants could be authorized for local education agencies, SEAS, institutions of higher education, Technical Assistance Centers, and other public and private nonprofit agencies to work on issues such as calibrating alternative assessments, training people to use them, bringing down the cost, and making them more objective. Congress could also consider allowing funds from the 5-percent local innovation set-aside to be used for local development and experimentation.

Since Chapter 1 is a major national influence on the amount, frequency, and types of standardized testing, a broad research and development effort for Chapter 1 alternative assessment would have an impact far beyond Chapter 1. The instruments, procedures, and standards developed by this type of effort would spill over into other areas of education, such as early childhood assessment, and would increase local districts' experimentation in other components of their educational programs.

An important issue for congressional consideration is the appropriate grade levels for Chapter 1 evaluations. There is considerable agreement that testing of children in the early grades is inappropriate, especially if standardized norm-referenced paper-and-pencil tests are used; the 1988 reauthorization eliminated testing requirements for children in kindergarten and first grade. On the other hand, there are compelling arguments that from a program evaluation point of view it is important to have "pre" and "post" data, which means collecting

some baseline information. Lack of a reliable method to demonstrate progress during the early years could discourage principals from channeling Chapter 1 funds to very young children, despite evidence that early intervention is very effective. If testing is required to show progress, these tests should be developmentally appropriate.[45]

A related congressional issue concerns the assessment of school children who have only been in a given school's Chapter 1 program for a short period of time; school districts throughout the country cite the high mobility of Chapter 1 children as a logistical obstacle to meaningful evaluation. Despite regulatory guidance, confusion continues to reign in State and local Chapter 1 offices about how to deal with a mobile student population. Clear and consistent policies regarding testing of these children would alleviate some of that confusion.

Appropriate Test Use

The ways tests should be used and the types of inferences that can appropriately be drawn from them are often not well understood by policymakers, school administrators, teachers, or other consumers of test information. Perhaps most important, many parents and test takers themselves are often at a loss to understand the reasons for testing, the importance of the consequences, or the meaning of the results. School policies about how test scores will be used are important not only to students and parents but also to teachers and other school personnel whose own careers may be influenced by the test performance of their pupils. Many of these problems result from using tests for purposes for which they are not designed or adequately validated. Fairness, due process, privacy, and disclosure issues will continue to fuel public passions around testing.

As reviewed in chapter 2, attempts to develop ethical and technical standards for tests and testing practices have a long history. The most recent attempt to codify standards for fair testing practice (in the *Code of Fair Testing Practices in Education)*[46] led to a set of principles with which most professional testing groups concur.

Educational testing practices in some areas have been defined by Federal legislation. In the rnid-1970s, Congress passed laws with significant provisions regarding testing, one affecting all students and parents and the others affecting individuals with disabilities and their parents. In both cases this Federal legislation has had far-reaching implications for school policy, because Federal financial assistance to schools has been tied to mandated testing practices. The Family Education Rights and Privacy Act of 1974-commonly called the 'Buckley Amendment" after former New York Senator James Buckley-was enacted in part to attempt to safeguard parents' rights and to correct some of the improprieties in the collection and maintenance of pupil records. The basic provisions of this legislation established the right of parents to inspect school records and protected the confidentiality of information by limiting access to school records (including test scores) to those who have legitimate educational needs for the information and by requiring parental written consent for the release of identifiable data.

Given the growing importance of testing and the precedent for Federal action, several avenues are open if Congress wishes to foster better educational testing practices and appropriate test use throughout the Nation.

One option for congressional action would aim at improved disclosure of information. Individual rights could be better safeguarded by encouraging test users (policymakers and schools) to do a careful job of informing test takers. Many critical decisions about test use, such as the selection and interpretation of tests, are made in a professional arena that is well-protected from open, public scrutiny. This occurs in part because of the highly technical nature of testing design. Although the professional testing community is not unanimous about what constitutes good testing practice, there is considerable consensus on the importance of carefully informing individual test takers (and their parents or guardians in the case of minors) about the purpose of the test, the uses to which it will be put, the persons who will

[45] See, e.g., Robert E, Slavin and Nancy A. Madden, Center for Research on Effective Schooling for Disadvantaged Students, The Johns Hopkins University, "Chapter 1 Program Improvement Guidelines: Do They Reward Appropriate Practices?" paper prepared for the Office of Educational Research and Improvement, U.S. Department of Education, December 1990. See also Nancy Kober, "The Role and Impact of Chapter 1 ESEA, Evaluation and Assessment Practices," OTA contractor report, June 1991.

[46] Joint Committee on Testing Practices, *Code of Fair Testing Practices in Education* (Washington, DC: National Council on M easurement in Education, 1988).

have access to the scores, and the rights of the test taker to retake or challenge test results.[47]

Congress could require, or encourage, school districts to:

- develop and publish a testing policy that spells out the types of tests given, how they are chosen, and how the tests and test scores will be used; and
- notify parents of test requirements and consequences, with special emphasis on tests used for selection, placement, or credentialing decisions.

A second approach for Congress is to encourage good testing practice by modeling and demonstrating such practice at the Federal level. The Federal Government writes much legislation that incorporates standardized testing as one component of a larger program. For example, the Individuals With Disabilities Education Act (Public Law 101476), formerly the Education for all Handicapped Children Act of 1975 (Public Law 94-142), was designed to assure the rights of individuals with disabilities to the best possible education; this legislation included a number of explicit provisions regarding how tests should be used to implement this program.

Among the provisions were: 1) decisions about students are to be based on more than performance on a single test, 2) tests must be validated for the purpose for which they are used, 3) children must be assessed in all areas related to a specific or suspected disability, and 4) evaluations should be made by a multidisciplinary team.

Through these assessment provisions, Public Laws 101-476 and 94-142 have provided a number of significant safeguards against the simplistic or capricious use of test scores in making educational decisions. Congress could adopt similar provisions in other legislation that has implications for testing. A recent example of Federal legislation that could lead to questionable uses of tests is a provision in the 1990 Omnibus Budget Reconciliation Act. The objective of this provision is to reduce the high loan default rate of students attending postsecondary training programs (largely but not exclusively in proprietary technical schools). The policy lever is testing: the act requires students without a high school diploma to pass an ''ability-to-benefit' test, on the assumption that students who are able to benefit from postsecondary training will be more likely to get jobs and pay back their loans than students who are not able to benefit. Basic questions arise about the appropriateness of using existing tests to sort individuals on this broad ''ability criterion. Even the most prevalent college admissions tests do not make claims of being able to predict which students will "benefit" in the long run, but rather which students will do well in their freshman year.

A third course of action would focus on various proposals to certify, regulate, oversee, or audit tests. If Congress wants to play a more forceful role in preventing misuse of tests—in particular, preventing tests designed for classroom use or system monitoring from being applied to individual selection or certification decisions—this option is the clear choice. If testing continues to increase and takes on even more consequences, pressure for congressional intervention will grow. Proposals include Federal guidelines for educational test use, labeling of all mandated tests and test requirements, labeling of all commercially available tests, and creating a governmental or quasi-governmental entity to regulate, certify, and disseminate information about tests. This last option, which echoes a concept endorsed by the National Commission on Testing and Public Policy, has been discussed in testing policy circles for some years now.[48]

Finally, Congress could pursue more indirect ways to inform and educate consumers and users of tests. This might include supporting continuing professional education for teachers and administrators, or funding the development of better ways to analyze test data and convey the results more effectively to the public.

Federal Research and Development Options

Test development is a costly process. Even for a test or test battery that has already been in use for many years, it can take from 6 to 8 years to write new

[47] See, for example, American Psychological Association, *Standards for Educational and Psychological Testing* (Washington, DC: 1985); Joint Committee on Testing Practices, op. cit., footnote 46; and Russell Sage Foundation (New York, NY: 1%9), especially Guideline 1.3.

[48] See, e.g., D. Goslin, "ThePresent and Future of Assessment: Towards an Agenda for Research and Public Policy,' draft report of a planning meeting sponsored by the U.S. Department of Education, Mar. 23-25, 1990, draft dated July 19, 1990.

items, pilot test, and validate a major revision.[49] Most investigators working on new testing designs are wading into uncharted statistical and methodological waters. For a new test, consisting of open-ended performance tasks or other innovative items, development and validation are substantially more expensive, even if test content and objectives are clearly defined. For example, the development of a set of new performance measures assessing specific job-related skills for the armed services cost $30 million over 10 years. The results of this sustained research effort, coordinated by the Department of Defense and carried out by the individual service organizations, were a set of hands-on measures, new supervisory ratings, job-knowledge tests, and computer-based simulations representing the skills required in some 30 well-defined jobs. The main purpose of the research was to improve the outcome or criterion measures used to validate the Armed Services Vocational Aptitude Battery, the standardized test used to qualify new recruits for various job assignments.[50]

In elementary and secondary school testing, however, the first step--defining the content that tests should cover—is much more complex than defining specific job performance outcomes for a number of jobs. The omnipresent issue of achieving consensus on content poses formidable barriers to test design. Even in a subject like mathematics, for which there is some agreement on outcomes and standards (as exemplified by the National Council on Teachers of Mathematics' recent work on standards for mathematics education), the definition of those standards took 6 years to develop. In most other subjects consensus on goals and curricula is more difficult to reach, adding substantially to research and development (R&D) costs. Moreover, separate standards, content, and tests would need to be developed for each grade level and subject to be tested.

Another factor making testing R&D expensive is the question of how new assessment methods will affect students and teachers. Much of the interest in developing new assessments (see ch. 6) stems from the desire to see those assessments eventually become the basis for system monitoring and other high-stakes decisions. Validation studies are therefore critical. Random assignment experiments, which are costly, could encounter legal barriers because students' lives and educational experiences could be affected. Validation studies, therefore, may need to be conducted with quasi-experimental designs, which suffer from various statistical and methodological problems.[51]

Congress has an important role to play in supporting R&D in educational testing, because adequate funding cannot be expected from other sources. Commercial vendors are not likely to make the requisite investments without some assurance of a reasonable return; they face strong market incentives to sell generic products that match the curricula of many school systems. But if these products are so general in their coverage that they reflect only a limited subset of skills common to virtually all curricula, schools may not see the advantage of adding them to an already strapped instructional materials budget. States might be willing to foot the R&D bill, although their education budgets are generally quite constrained. Moreover, in addition to costs associated with consensus-building on test content and evaluation of the anticipated effects of testing, new performance assessment and/or computer-based methods require basic research on learning and cognition. Basic education research has traditionally been a Federal responsibility.

The question becomes how much: how much should the Federal Government spend on educational testing R&D? The answer depends on the choice Congress makes regarding the value of dramatically enlarging the currently available range of testing methods. For example, Federal spending on educational assessment research is roughly $7 million for fiscal year 1992, out of a total education research budget of close to $100 million.[52] This money is divided almost evenly among NAEP (for validation studies, evaluation of trial State assessment, and secondary data analysis); development of new mathematics and science assessments ($6

[49] Rudman, op. cit., footnote 8, p. 8.

[50] See Alexandra Wigdor and Bert Green (eds.), *Performance Assessment for the* vol. 1 (Washington, DC: National Academy Press, 1991).

[51] See, e.g., Anand Desai, ''m~cd Issues in Measuring Scholastic Improvement Due to Compensatory Education Programs, " *Socio-Economic* vol. 24, No. 2, 1990, pp. 143-153.

[52] Education research and statistics spending in fiscal year 1990 was $94 million. See U.S. Department of Education, *Statistics, 1990*, op. cit., footnote 1, p. 344.

million over 3 years, administered through the National Science Foundation); and general assessment research (through the Center for Research on Evaluation, Standards, and Student Testing).

Substantially more funding would be needed if Congress chooses to support:

- cognitive science research on learning and testing,
- development of new approaches to consensus building for test content and objectives,
- research on the generalizability of new testing methods across subjects and grades, and
- validation studies of new testing methods.

An intermediate funding approach would be to target Federal dollars toward:

- the creation of a clearinghouse to facilitate continuing and more widespread dissemination of testing research results and innovations,
- continuing professional education for teachers in the applications of new testing and assessment methods and in the appropriate interpretations and uses of test results, and
- the creation of a nationwide computer-based clearinghouse of test items from which States and local districts could draw to develop their own customized tests.