CHAPTER 2

Testing in Transition

Contents

Boxes

Figures

Tables

CHAPTER 2

Testing in Transition

Highlights

- Since the 1960s testing in elementary and secondary schools has been caught in a tug between two powerful forces: increased public attention to test scores because of demands for evidence that the schools are educating children, and increased demands from educators and students for tests that more accurately reflect changing educational goals, new curricula, and reforms in teaching.
- State-level concerns about the quality of education were the dominant force behind the rise of high-stakes testing beginning in the mid-1970s. Minimum competency testing, for example, was embraced by many State policymakers who believed that the imposition of external standards would boost educational quality. Since then, however, studies of the effects of this testing have led most educators to question the utility of tests as an instrument of reform.
- Two decades of research about learning and cognition have produced important findings about how children learn and acquire knowledge. These findings challenge most traditional models of classroom organization, curricula, and teaching methods. Among the most important findings are that teaching thinking skills need not await mastery of so-called "basic" skills, and that all students are capable of learning thinking skills. Many educators now charge that significant changes in classrooms cannot go forward if traditional tests are to remain the primary indicator of achievement and program success. The tests must change, they argue, if schools are to change.
- Many of the recent challenges to traditional tests have been directed at the norm-referenced multiple-choice tests most often used to assess educational achievement. It is not just the tests themselves that create controversy, however. Testing practices-the ways tests are used and the types of inferences drawn from them-also create many of the problems associated with testing. Appropriate testing practices are difficult to enforce and few safeguards exist to prevent misuse and misinterpretation of scores, especially once they reach the public.
- Test-use policy is 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. Concern for the increasing consequences being attached **to** test scores has helped fuel a backlash against standardized testing that had been brewing since the expansion of high-stakes testing in the 1970s, when issues of fairness, test bias, due process, individual privacy, and disclosure were debated in Congress and the courts.
- Although demands for accountability have not abated amid this environment of testing reform, *most* educators now urge the development and implementation of new testing and assessment technologies, and all caution against the use of tests as the sole or principal indicator of achievement.

Overview

Two decades of discussion about school quality have convinced many Americans that their educational system needs substantial reform to meet the demands of the next century. Although the country is far from consensus about **exactly** what types of reform are needed, nearly all the initiatives call for changes in educational testing.

Some school reformers, primarily at the State level, have called for changes in testing to monitor student progress in mastering fundamentally new curricula. Others have pinned their hopes on more high-stakes testing-including yet-to-be-developed national tests—to spur greater student and teacher diligence. This group includes educators and policymakers who believe that new and better ***tests can*** lead to improved learning, as well as those who believe in conventional tests as a catalyst of change. Still others fear that more testing of any type will only exacerbate the problems of test misuse and unfairness, and will be counterproductive to school reform. These debates should not surprise anyone familiar with the U.S. education system: standard-

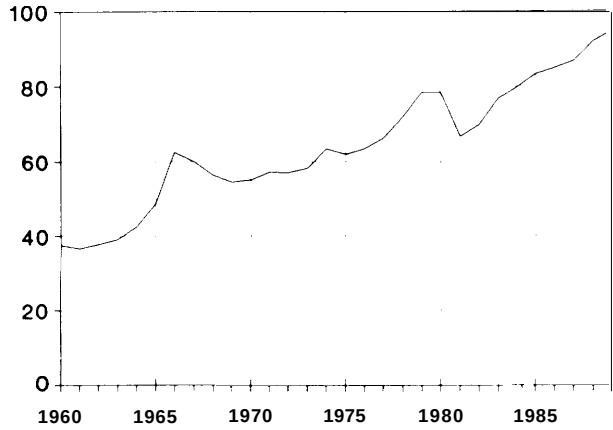
ized tests have always been prominent, and discussion of educational reform inevitably involves an examination of testing.

Since the 1960s Americans have turned increasingly to testing as a tool for measuring student learning, holding schools accountable for results, and reforming curriculum and instruction. Testing in elementary and secondary schools has, therefore, increased in both frequency and significance. As shown in figure 2-1, revenues from sales of commercially published standardized tests for K-12 more than doubled between 1960 and 1989; i.e., from about $40 million in 1960 to about $100 million in 1989 (in constant 1982 dollars). A recent report of the National Commission for Testing and Public Policy estimates that the 44 million American elementary and secondary students take 127 million separate tests annually, as part of standardized test batteries mandated by States and districts.[1]

Much of this growth in testing occurred during a period of economic, social, and demographic turbulence, and is attributable to Federal, State, and local demands for increased accountability.[2] These strategies for change, such as performance reporting, establishing and enforcing procedural standards, and changing school structure or the professional roles of school personnel, rely on test information about schools and students.[3]

At the Federal level, demands for test-based accountability emerged as a consequence of substantial new financial commitments to education on the part of the Federal Government. State-mandated tests, often designed and administered by State authorities (rather than by commercial vendors) have also grown dramatically; State-level concern with the quality of education, and State-level demands for improvement in the outcomes of schooling, have perhaps been the dominant forces behind the rise of standardized testing in the past two decades.

Figure 2-l—Revenues From Sales of Commercially Produced Standardized Tests in the United States, 1960-90



a sales in 1982 dollars.
NOTE: Sales include K-1 2 educational tests.
SOURCE: Office of Technology Assessment, based on data from Filomena Simora (cd.), *The Bowker Annual (New York, NY:* Reed Publishing, 1910-90).

The pattern of increased testing followed by increased controversy dates to the initial uses of tests to stimulate school reform in the 19th century.[4] In different periods the specific causes of controversy over testing have varied. Today the debate stems from three main factors.

First, many of the people and school systems attempting to redesign curricula and reform teaching and learning feel stymied in their efforts by tests that do not reflect new education goals. Moreover, because tests have increasingly high stakes, reformers find that bold new ideas of curricula and instruction cannot surmount the power of tests to reinforce traditional learning. For example, the basic ‘ ‘building-block’ ‘ approach to student learning—

[1]National Commission on Testing and Public Policy, *From Gatekeeper to Gateway: Transforming Testing in America* (Boston, MA: 1990), p. 15. Test publishers claim that the National Commission exaggerates in its estimate of testing. For example, the Vice President for Publishing at one of the largest educational test publishing companies argues that: ‘‘Our data sources indicate that roughly 30 to 40 million standardized tests are administered annually across the country. . . [at an annual] total cost of. ..$100 million to $150 million. . . .“ See Douglas MacRae, ‘‘Topic: Too Much Testing?’’ *CTB* Desk, No. 3, Nov. 15, 1990.

[2]The work of Leon Lessinger, “Accountability for Results, “ (Washington DC: U.S. Office of Education, June-July 1969), is often credited with igniting the most recent wave of accountability in education. For a synthesis and discussion of approaches to accountability in education see Michael Kirst, *State Local Policymakers* (Washington DC: U.S. Department of Education, July 1990).

[3]‘‘An aroused parent group, for example, will follow up on the results of a negative school report card by lobbying the school board for a new principal. ” Kirst, op. cit., footnote 2, p. 7.

[4]See ch. 4.

the idea that children needed to be solidly grounded in the basics before acquiring advanced thinking and problem-solving skills-has been gradually supplanted by new research findings. Curriculum specialists as well as teachers have begun arguing for new approaches to the definition and instruction of "higher order skills," and for changes that could make tests better indicators of learning.

Second, the demand for test-based accountability continues to grow. Advocates of test-based accountability argue it is an efficient and effective way to make students, teachers, and schools work harder. Some go so far as to suggest that raising the stakes of these tests can put America back on the road to global economic hegemony: since teachers will teach and children will study what is tested, the thinking goes, then the tests themselves cart drive educational reform.[5] Opponents of this view charge that high-stakes testing sends the wrong signals to students and teachers, and encourages emphasis on test taking and test preparation rather than genuine learning. They also argue that attaching high stakes to tests threatens the validity of the information provided by the tests and leads to erroneous policy inferences.

Third, as the tension surrounding tests increases, so do concerns about the appropriate use of tests and the effects of tests on individual rights. The history of testing is littered with examples of tests being used in ways not intended by their developers, tempting policy makers and the public to draw inferences not supportable by test data.

The three camps-those who support new approaches to assessment and testing, those who think more high-stakes testing will improve education, and those who are worried about ethical and legal aspects of testing—share a common concern for raising the quality of American schooling. But their strategies are crafted from visions of the educational system and the nature of human learning glimpsed through very different prisms.

Changing Views of Teaching and Learning

A quiet but dramatic transformation is occurring in education as researchers and practitioners rethink basic beliefs about teaching and learning. Two decades of research from developmental and cognitive psychology have produced important findings about how children learn and acquire knowledge.[6] The basic concept in this research is that children are active builders of their own knowledge, not merely passive receptacles for information. These research findings and the instructional theories they have spawned raise serious challenges to traditional classroom organizational models, to conventional curricula, and, in turn, to existing forms of testing. Moreover, they have rekindled an awareness of the close links between instructional goals and assessment.

Evolving Views of Learning

In their teaching methods, curricular materials, and testing methods, many schools today embody a behaviorist model of learning first popularized in the 1920s. In this model:

> . . . learning is seen to be linear and sequential. Complex understanding can only occur by the accretion of elemental, prerequisite learnings. . . . The whole idea was to break desired learnings into constituent elements and teach these one by one. . . . The implications of this model for instruction are conveyed best by . . . [the] metaphor of a brick wall, i.e., it is not possible to lay the bricks in the fifth layer until the first, second, third, and fourth layers are complete.[7]

This model assumes that more complex skills can be broken down into simple skills, each of which can be mastered independently and out of context. When all requisite components are mastered, then more complex thinking skills can accrue. According to this view, the highest levels of knowledge are achieved only at the later grades and, even then, only by some students. In this conventional model, moreover, the teacher is the active partner in the educational

[5] See e.g., Robert Samuelson, "The School Reform Fraud," *Washington Post*, June 19, 1991, p. A19.

[6] The following discussion about constructivist and behaviorist models of learning draws on Lauren B. Resnick and Daniel P. Resnick, "Assessing the Thinking Curriculum: New Tools for Educational Reform," paper prepared for the National Commission on Testing and Public Policy, August 1989; Lorrie A. Shepard, University of Colorado at Boulder, "Psychometricians' Beliefs About Learning," paper presented at the annual meeting of the American Educational Research Association, Boston, MA, Apr. 17, 1990.

[7] Shepard, op. cit., footnote 6, p. 15.

process, imparting knowledge to a passive student as though filling an empty jug.

This hierarchical view of complex thinking is challenged by recent research from the cognitive sciences.

> One of the most important findings of recent research on thinking is that the kinds of mental processes associated with thinking are not restricted to an advanced or 'higher order' stage of mental development. Instead, thinking and reasoning are intimately involved in successfully learning even elementary levels of reading, mathematics, and other school subjects. Cognitive research on children's learning of basic skills reveals that reading, writing, and arithmetic-the three Rs—involve important components of inference, judgment and active mental construction [see box 2-A]. The traditional view that the basics can be taught as routine skills, with thinking and reasoning to follow later, can no longer guide our educational practice.[8]

In fact, the term ' 'higher order" thinking skills seems something of a misnomer in that it implies that there is another set of "lower order" skills that need to come first.

Another implication of the hierarchical "brick wall' model of learning is the notion that slower learners need to master low-level skills before they can move on to more complex skills. This sort of thinking underlies many compensatory education programs, in which educationally disadvantaged children or children who learn more slowly than their peers spend much of their time confined to remedial classes consisting of drill and practice. By a process of remediation through repetition students are expected to master the low-level skills; many, however, spend a good portion (if not all) of their educational careers confined to the mastery of basic skills through remedial methods. The constructivist model of learning indicates that these students are capable of much more than this; this research suggests that all are naturally engaged everyday in problem solving, making inferences and judgments, and forming theories about how the world works.

Several programs designed specifically to focus on increasing the achievement of disadvantaged

Photo credit: Siemens Corp.

Recent research has emphasized that learning is an active process that can best be supported in the classroom by hands-on activities and experimentation. As curricula and teaching practices change, new tests will also be needed.

learners provide evidence to support the notion that these students are capable of learning far more than basic skills. The Accelerated Schools Program is a reform experiment designed to accelerate the learning of at-risk students and close the ''achievement gap' while the students are still in elementary school. The program sets high expectations for student learning and focuses on the teaching of critical thinking and problem solving to all students. Although these programs do not yet have a long track record, teachers report delight and surprise at the gains achieved by participating students.[9] Another program, the Higher Order Thinking Skills (HOTS) project, provides Chapter 1 students in grades four through seven with enhanced thinking skills instead of remediation. The HOTS project has yielded compelling anecdotal evidence of substan-

[8]Resnick and Resnick, op. cit., footnote 6, p. 2.

[9]Gail Meister, Research for Better Schools, "Assessment in Programs for Disadvantaged Students: Lessons From Accelerated Schools," OTA contractor report, April 1991.

Box 2-A—Fourth Grade Scientists Test a Theory[1]

For nine winters, experience had been their teacher. Every hat they had worn, every sweater they had donned, contained heat. "Put on your warm clothes," parents and teachers had told them. So when the children in Ms. O'Brien's fourth grade science class began to study heat one spring day, who could blame them for thinking as they did?

"Sweaters are hot," said Katie.

"If you put a thermometer inside a hat, would it ever get hot! Ninety degrees, maybe," said Neil.

. . . [With O'Brien's help, the students set out to test these theories.] Christian, Neil, Katie, and the others placed thermometers inside sweaters, hats, and a rolled-up rug. When the temperature inside refused to rise after 15 minutes, Christian suggested that they leave the thermometers overnight. After all, he said, when the doctor takes your temperature, you have to leave the thermometer in your mouth for a long time. Folding the sweaters and hats securely, the children predicted three digit temperatures the next day.

When they ran to their experiments first thing the next morning, the children were baffled. They had been wrong. Now they'll change their minds, and we can move on, O'Brien thought.

But . . . the children refused to give up. "We just didn't leave them in there long enough, " Christian said. ''Cold air got in there somehow,' said Katie.

. . . [O' Brien suggested they adjust their experiments and try again.] If, as they insisted, cold air had seeped inside the clothes overnight, what could they do to keep it out? . . . Neil decided to seal the hat, with the thermometer inside, in a plastic bag. Katie chose to plug the ends of the rug with hats. Others placed sweaters in closets or in desks, far away from the great gusts of cold air they seemed to think swept their classroom at night.

. . . On Wednesday morning the children rushed to examine their experiments. They checked their deeply buried thermometers. From across the room, they shared their bewilderment. All the thermometers were at 68 degrees Fahrenheit. Confused, they wrote in their journals. "Hot and cold are sometimes strange," Katie wrote. ''Maybe [the thermometer] didn't work because it was used to room temperature."

Meanwhile, O'Brien wondered in her own journal . . . how long she should let these naive conceptions linger. [She decided to have the students proceed with] . . . one more round of testing. And so the sweaters, hats, and even a down sleeping bag brought from home were sealed, plugged, and left to endure the cold.

. . . For the third day in a row in O'Brien's classroom, the children rushed to their experiments as soon as they arrived. The sweater, the sleeping bag, and the hat were unwrapped. Once again the thermometers uniformly read room temperature. O'Brien led the disappointed children to their journals. But after a few moments of discussion, she realized that her students had reached an impasse. Their old theory was clearly on the ropes, but they had no new theory with which to replace it. She decided to offer them a choice of two possible statements.

''Choose statement A or B," she told them. The frost stated that heat could come from almost anything, hats and sweaters included. In measuring such heat, statement A proclaimed, we are sometimes fooled because we're really measuring cold air that gets inside. This, of course, was what most children had believed at the outset. Statement B, of O'Brien's own devising, posed the alternative that heat comes mostly from the sun and our bodies and is trapped inside winter clothes that keep our body heat in and keep the cold air out.

''Write down what you believe,' O'Brien told the class. [Although some students clung to the "hot hat' theory and some did not know what to think, most choose theory B.]

"How can we test this new theory?" O'Brien asked. Immediately Neil said, ''Put the thermometers in our hats when we're wearing them. ' And so the children went out to recess that day with an experiment under their hats.

As Deb O'Brien relaxed during recess, she asked herself about the past three days. Had the children really changed their minds? Or had they simply been following the leader? Could they really change their ideas in the course of a few class periods? Would any of their activities help them pass the standardized science test coming up in May? O'Brien wasn't sure she could answer any of these questions affirmatively. But she had seen the faces of young scientists as they ran to their experiments, wrote about their findings, spoke out, thought, asked questions—and that was enough for now.

---

[1]Excerpted from Bruce Watson and Richard Konicek, ''Teac~g for Conceptual Change: Confronting Children's Experience, ' *Phi Delta Kappan*, vol. *71, No.* **9, May 1990,** pp. 680685.

tial gains in self-esteem and enthusiasm for learning—as well as achievement test scores—when children participate in the program for 35 minutes a day over 2 school years.[10]

Additional evidence suggests that thinking and reasoning skills can be taught.[11] A number of programs have been designed to teach thinking and problem-solving skills; some focus on developing these skills within particular disciplines (e.g., mathematics and reading) while others are aimed at enhancing general thinking skills that would, presumably, be applicable in many different settings. The effectiveness of these programs is difficult to evaluate in the absence of appropriate outcome measures. Evaluations show students improving on measures tied to the material taught: students appear to learn to do the things the program teaches. The question of whether that learning generalizes is more difficult to assess, in part because there are few good outcome measures for these skills.[12]

The results of these studies suggest some hopeful beginnings for the design of curricula and teaching methods focused on thinking and reasoning skills. Much of this work is new and experimental. Experimentation is needed to discern how much emphasis to place on general thinking skills and how much to emphasize thinking skills for specific knowledge and information. Moreover, knowledge of how to teach those reasoning skills-at what ages, using what methods-is still very rudimentary.

In sum, although educators have always attempted to foster reasoning skills, research about learning and the structure of knowledge suggests two major changes in how those skills should be taught. First, thinking skills need not be learned only after other, more basic skills are mastered. Second, all students are capable of learning thinking skills.

Evolving Views of the Classroom

Recent developments in education have converged to make more and more classrooms into vital laboratories for new teaching and learning methods. First, the growing presence of educational technology in the classroom, especially computers and integrated learning systems, is changing the definitions of what children need to know and how to teach it.

Second, educators are radically rethinking the structure and content of their disciplines. For example, the National Council of Teachers of Mathematics (NCTM) has proposed fundamental changes in the content and delivery of elementary and secondary school mathematics instruction, changes that emphasize the use of manipulative objects and the teaching of analytical reasoning and problem-solving skills. Mathematics educators have recognized that: ". . . the world is changing so rapidly that, unless those involved in mathematics education adopt a proactive view and develop a new assessment model for the twenty-first century, the mathematical understanding of children will continue to be inadequate into the future;"[13] and they have worked to build consensus on a set of curriculum standards for K-12 education. Initiatives to revisit science curricula and teaching methods have also taken hold, with particular efforts to stress ''hands-on' science experiments. In addition, many schools are experimenting with the idea of the ''integrated curriculum," in which central themes or ideas are taught across disciplines and the school day is no longer divided into discrete periods labeled by subject.

Third, attention is being directed toward the development of materials and methods for cultivating higher order thinking skills (see box 2-B). The emphasis on fostering reasoning skills has been bolstered by the widespread recognition that changing economic and technological conditions will

---

[10] S. Pogrow, "Challenging At-Risk Students: Findings From the HOTS Program," *Kappan*, vol. 71, No. 5, January 1990, pp. 389-397.

[11] For descriptions of some of these efforts see R. Glaser, "Education and Thinking: The Role of Knowledge," February 1984, pp. 93-104; Lauren B. Resnick, *to* (Washington, DC: National Academy Press, 1987); Lauren B. Resnick and Leopold E. Klopfer (eds.), *Thinking* Yearbook of the Association for Supervision and Curriculum Development (Alexandria VA: Association for Supervision and Curriculum Development 1989); and Norman Frederiksen, "Implications of Cognitive Theory for Instruction in Problem Solving," vol. 54, No. 3, fall 1984, pp. 363-407.

[12] Resnick, op. cit., footnote 11.

[13] Thomas A. Romberg, E. Anne Zarrinnia, and Kevin F. Collis, "A New Worldview of Assessment in Mathematics," Kulm (cd.) (Washington, DC: American Association for the Advancement of Science, 1990), p. 21.

Box 2-B—Thinking About Thinking Skills

What are "higher order thinking skills"? What do they look like and how do we know when students have them? The first truism seems to be that they are difficult to define; the second is that they are even harder to measure.

Social scientists from many disciplines have studied mental processes such as thinking, problem solving, reasoning, and critical thinking; although they have produced many carefully wrought definitions, consensus about the nature of these processes has eluded them. Educational practitioners, on the other hand, have less interest in understanding the precise nature of all possible thinking processes; instead, practitioners are most concerned about the ". . . complex thought processes required to solve problems and make decisions in everyday life, and those that have a direct relevance to instruction. "[1] One recent attempt to synthesize the perspectives of philosophers, psychologists, and educators has produced the outline of thinking skills shown in table 2-B 1. As this table suggests, at least some consensus exists about the kinds of skills educators would like to include in a thinking curriculum.

[1] J.A. Arter and J.R. Salmon, Northwest Regional Educational Laboratory, "Assessing Higher Order Thinking Skills: A consumer's Guide," unpublished report, April 1987, pp. 1-2.

Table 2-B1—List of Thinking and Reasoning Skills

- I. Problem solving
  - A. Identifying general problem
  - B. Clarifying problem
  - G. Formulating hypothesis
  - D. Formulating appropriate questions
  - E. Generating related ideas
  - F. Formulating alternative solutions
  - G. Choosing best solution
  - H. Applying the solution
  - L Monitoring acceptance of the solution
  - J. Drawing conclusions
- II. Decisionmaking
  - A. Stating desired goal/condition
  - B. Stating obstacles to goal/condition
  - C. identifying alternatives
  - D. Examining alternatives
  - E. Ranking alternatives
  - F. Choosing best alternative
  - G. Evaluating actions
- iii. inferences
  - A. inductive thinking skills
    - 1. Determining cause and effect
    - 2. Analyzing open-ended problems
    - 3. Reasoning by analogy
    - 4. Making inferences
    - 5. Determining relevant information
    - 6. Recognizing relationships
    - 7. Solving insight problems
  - B. Deductive thinking skills
    - 1. Using logic
    - 2. Spotting contradictory statements
    - 3. Analyzing syllogisms
    - 4. Solving spatial problems
- IV. Divergent thinking skills
  - A. Listing attributes of objects/situations
  - B. Generating multiple ideas (fluency)
  - C. Generating different ideas (flexibility)
  - D. Generating unique Ideas (originality)
  - E. Generating detailed ideas (elaboration)
  - F. Synthesizing information
- V. Evaluative thinking skills
  - A. Distinguishing between facts and opinions
  - B. Judging credibility of a source
  - C. Observing and judging observation reports
  - D. Identifying central issues and problems
  - E. Recognizing underlying assumptions
  - F. Detecting bias, stereotypes, cliches
  - G. Recognizing loaded language
  - H. Evacuating hypotheses
  - i. Classifying data
  - J. Predicting consequences
  - K. Demonstrating sequential synthesis of information
  - L Planning alternative strategies
  - M. Recognizing inconsistencies in information
  - N. Identifying stated and unstated reasons
  - O. Comparing similarities and differences
  - P. Evaluating arguments
- VI. Philosophy and reasoning
  - A. Using diaiogicai/dialectical approaches

NOTE: This list is basedon a compilation and distillation of ideas from many educators and psychologists. See original source.

SOURCE: JA. Arter and J.R. Salmon, Northwest Regional Education Laboratory, "Assessing Higher-Order Thinking Skills: A Consumer's Guide," unpublished report, April 1987, p. 3.

require upgrading the cognitive skills of the work force.[14] The combined effects of research on learning and public concern for the state of the education system have led some educators to suggest that reasoning should be considered as ''the fourth R.''[15] In classrooms across the country, teachers are experimenting with ways to teach critical thinking and comprehension along with basic skills and information.

Implications for Standardized Testing

Educators trying to implement these new ideas and classroom practices have found themselves face to face with the dominance of standardized norm-referenced tests as the sine qua non of educational effectiveness. Many have found their new programs being judged by tests that do not cover the skills and goals central to their innovations, Those working on integrated curricula, anew vision of mathematics, or hands-on learning environments have found their new programs measured by tests designed for very different goals. Thus, anew and energetic movement has emerged focused on developing assessments more closely aligned with new curricula, learning methods, and valued skills.

The press for reform of tests to better match instruction and curricula comes from many sources. Educators are recognizing the potential of computers to change testing just as they are changing learning. Curriculum reform groups, such as the NCTM standards committee, are seeking assessments better matched to their curricular and evaluation standards. Educators working to increase the achievement of disadvantaged learners express frustration that many of their critical program goals are not measured by existing standardized tests.[16] A common theme is that transformation of education cannot occur as long as tests embrace obsolete concepts about learning. Without new assessment instruments, it is difficult to ascertain whether reforms in instruction and curriculum are working.

What implications does a focus on thinking skills and active learning have for test design? Reformers trying to implement a thinking curriculum agree on the need for changes that will better focus on reasoning skills and deep understandings. Test designers have always advanced the idea that an achievement test should be designed to reflect the goals of the curriculum. Most current achievement tests were constructed by careful delineation of the subject matter (e.g., reading, language arts, and mathematics); experts in the subject matter areas were largely responsible for specifying the domains of information and the skills to be mastered. However, ". . . a clear definition of the subject-matter content is essential, but insufficient by itself. 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. ' '[17]

Until recently most attempts to incorporate cognitive skills into test design were modeled on Bloom's taxonomy of cognitive behaviors,[18] which attempts to organize and classify the cognitive skills children are supposed to acquire. The taxonomy reflects a behavioral approach to learning; educational objectives are written as clearly delineated, mutually exclusive categories of behavior that can be observed, counted, and classified. Tests based on this taxonomy are organized according to a content-by-

---

[14] Although most analysts agree that some improvement in thinking skills will be beneficial, there is disagreement over how high to raise the threshold. The disagreement stems from conflicting interpretations of data on the productivity of the work force currently and on the effects of technological change on future skill requirements. For an eloquent discussion, see Richard Murnane, "Education and the Productivity of the Work Force: Looking Ahead,' R. Litan, R. Lawrence, and C. Schultze (eds.) (Washington, DC: Brookings Institution 1988), pp. 215-246.

[15] R. Glaser, "The Fourth R: The Ability to Reason,'' paper presented to the Federation of Behavioral, Psychological and Cognitive Sciences Science and Public Policy Seminar, June 1989; and Larry Cuban, "Policy and Research Dilemmas in the Teaching of Reasoning: Unplanned Designs, ' vol. 54, 1984, pp. 655-681.

[16] See Meister, op. Cit., footnote 9.

[17] Robert Linn, ''Barriers Design,'' proceedings of the 1985 ETS Invitational Conference, Eileen E. Freeman (cd.) (Princeton, NJ: Educational Testing Service, 1986), p. 73.

[18] B.S. Bloom (cd.), *Taxonomy of Educational Objectives: The Classification of Educational Goals. Handbook 1—Cognitive Domain* (New York, NY: Academic Press, 1956). This discussion of the applications of Bloom's taxonomy to achievement testing is drawn from Romberg et al., op. cit., footnote 13. See also Edward Haertel and Robert Calfee, "School Achievement: Thinking About What to Test,' " summer 1983, pp. 119-132.

behavior matrix. As the example in figure 2-2 demonstrates, one axis of the matrix lists the content areas and the other axis describes the skills test takers are expected to demonstrate within each content area (in this example, computation, comprehension, application, and analysis). Items are designed for each cell in the matrix. Despite changes over time in the specifics of each axis, the matrix approach to test design has persisted because ''. . . it permits a rapid overview of the entire structure [of a test] and relative emphasis on one part or another."[19]

Some critics of the taxonomic approach feel that the matrix oversimplifies the complexity of knowledge and how students acquire it. Subject matter experts from various content disciplines have criticized the way that such matrices artificially divide both content and skills into mutually exclusive categories, ignoring complex interrelationships. In fact, the matrix form, by its very nature, suggests " . . . relationships which are simple, numerically restricted and linear . . .' '20—an outmoded concept that views thinking skills as hierarchically nested atop one another, with the learner moving from simple thinking skills to more complex ones as achievement advances.

Cognitive Research: Implications for New Test Design

Since the publication of Bloom's taxonomy, considerable research has been conducted about the nature of the cognitive processes involved in learning. The findings from cognitive sciences research provide a basis for different kinds of instruction, curriculum materials, and tests that more closely resemble the processes involved in learning and thinking (see box 2-C). Findings from research on learning and cognition imply at least three broad changes for educational tests:

1. Knowledge is a complex network of information and skills, not a series of isolated skills and facts. Tests designed to assess knowledge must reflect this complexity both in the tasks they require children to complete and the criteria they use to evaluate a child's knowledge.

Figure 2-2—Example of Content-by-Behavior Matrix for a 60-item Mathematics Test

Behavior	Content areas: Number systems	Geometry	Algebra	Total
Computation	15	8	7	30
Comprehension	5	5	5	15
Application	5	3	2	10
Analysis	o	4	1	5
Tot al	25	20	15	60 items

NOTE: The values in the cells represent the number of items on the test. Matrices like this are used in planning and designing tests.

SOURCE: Office of Technology Assessment, 1992. Based on a concept discussed in Thomas A. Romberg, E. Anne Zarinnia, and Kevin J. Collis, "A New Worldview of Assessment in Mathematics," *Assessing Higher Order Thinking in Mathematics, G.* Kulm (cd.) (Washington, DC: American Association for the Advancement of Science, 1990).

2. The research suggests important new possibilities for tests that can diagnose a student's strengths and weaknesses. 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. The shift toward educationally diagnostic tests is an important one; it represents a move away from seeing tests as predictive indicators of a fixed ''ability to learn' to tests that can help shape instruction so ''all can learn. ' '[21]

3. Because research indicates that much learning and thinking is active and occurs within a specific context, assessment of some skills may require testing methods more closely tied to the active learning process. Tasks may need to resemble what students should be able to do, and thus what they spend their time doing in the classroom. It is likely that tests that allow children to manipulate materials, explore naive theories, and demonstrate everyday cognition will more accurately reflect their competence levels across a range of skills. Instruction and assessment can be designed to focus on learning in context; as this happens more, especially in the new forms of assessment commonly referred to as 'performance assess-

[19]Romberg et al., op. cit., footnote 13, p. 9.

[20]Ibid., p. 15.

[21]See, e.g., J.W. Pellegrino, "Anatomy of Analogy, " *Psychology Today*, vol. 19, No. 10, October 1985, pp. 48-54.

Box 2-C—Tests as Avenues to Individualized Learning

Cognitive and developmental psychologists tend to look for patterns and similarities in the way people think and learn. While research has documented some general patterns, it has also found tremendous individual variation in the rates at which children learn and develop. Other research has drawn attention to the importance of individual differences in social, emotional, and motivational characteristics that affect children's learning. Still others have focused on the modality or "style" by which different children learn. Many have reasoned that if tests can diagnose learning styles, then they can aid in the development of improved instructional techniques matched to individual learning styles. There have been many theories, but no consensus on what those different learning styles look like. Attempts to match learning styles to styles of instruction were initially popular in special education, but the research has not held up in part because the measures for diagnosing learning styles are not reliable enough and do not show expected relationships with achievement.[1]

Nevertheless, the research suggests that the "ability to learn" (a commonly used definition of "intelligence") is not a fixed unitary trait: individuals do not have a certain amount of it that predetermines how well and how much they can learn. The model of learning disabilities provides a well-accepted example of how one or two areas of weakness, such as recognition of written words, can interfere with a child's skills across a broad range of academic areas. While in the past these children were often seen as unable to learn, or worse yet as 'dumb,' their capabilities are now recognized. Many such children need alternative learning methods in order to acquire necessary skills. Every child brings to any learning situation a complex profile of strengths and weaknesses as well as past learning and experience. Diagnostic tests can describe in detail the actual skills of a child in areas related to instruction. Strengths can then be used by the teacher to support and guide learning in more difficult areas.

One attempt to describe children's skills more broadly is a recent effort to outline "multiple intelligences." Although theories of the multiple components of intelligence have been around for a long time, Howard Gardner's work suggests that most of our current approach to education, as well as assessment, has relied heavily on developing two types of intelligence, which he calls "logical-mathematical" and "linguistic."[2] Drawing on evidence from multiple sources, including neuropsychology and child development, Gardner has proposed an additional five types of intelligences: musical, spatial, bodily-kinesthetic, interpersonal, and intrapersonal. A student can be represented by a profile of "intelligences,' each of which is relatively independent of the others.[3]

Several educational pilot programs have grown out of this theory. One, the Key School in Indianapolis, attempts to maximize instruction across all seven areas and uses report cards that evaluate children in each. Another, Project Spectrum, has attempted to develop assessment activities that capture the seven competencies in preschool children. The goal of these efforts is to provide a profile of strengths and weaknesses across the seven areas that can be used to direct educational resources to the child; such a profile could help parents and teachers build on strengths or bolster areas of weakness during the early years.[4]

The theory of multiple intelligences provides one model for broadening traditional views about which skills and competencies are important and require nurturing in the school years. As one policy maker has noted:

> Gardner's work has been important in attacking the monolithic notion of intelligence that has undergirded much of our thinking. we are beginning to see that education is not meant merely to sort out a few children and make them leaders, but to develop the latent talents of the entire population in diverse ways?[5]

---

[1] D. Carnine, "Research on the Brain: Implications for Instruction," *Delta Kappan*, vol. 71, No. 5, January 1990, pp. 372-377; and Kenneth A. Kavale and Steven R. Forness, "Substance Over Style: Assessing the Efficacy of Modality Testing and Teaching," *Children*, vol. 54, No. 3, 1987, pp. 228-239.

[2] Howard Gardner, *Frames of Mind* (New York, Basic Books, 1985). For a fuller discussion of the contributions of Spearman, Guilford, Thurstone, and other researchers whose work was based on different theories of the structure of intelligence, see, e.g., Raymond Fancher, *The Intelligence IQ* Controversy (New York NY: W.W. Norton, 1985).

[3] The work of Robert Sternberg, another modern pioneer of multiple intelligence, while focused largely on adults rather than school children, also has important implications for instruction and assessment. See, for example, his book, *Triarchic of* (London, Cambridge University Press, 1985).

[4] For further description of these programs see Marie Winn, "New Views of Human Intelligence," *New York Times Magazine*, part 2, The Good Health Magazine, Apr. 29, 1990; and Howard Gardner and Thomas Hatch, "Multiple Intelligences Go to School," vol. 18, No. 8, November 1989, pp. 4-9.

[5] Rexford Brown, Director of Communications for the Education Commission of the States, quoted in Winn, Op. cit., footnote 4, p. 30.

ment" (see ch. 7), the lines between assessment and instruction blur. Assessment becomes feedback to the learner, which in turn promotes further learning and growth.

There are many more specific ways in which the findings from cognitive psychology could find their way into test design, but few areas of cognitive research are ready for immediate translation into new achievement tests. Thus, any test designed using new cognitive findings is likely to require considerable research and development before the thinking skills that underlie the test can be measured with confidence.

The emergence of new theories of cognition and new instructional strategies raises a fundamental question about the nature of the relationship between curriculum and assessment. Those who advocate reforming tests to more closely parallel new theories of learning tend to believe that tests should ***follow*** curriculum and instruction. In this regard, they echo principles of educational test design well established in the literature of educational measurement.[22] The first step to improving education, according to this view, is to establish what it is students are supposed to learn and how they are most likely to learn it; the next step is to develop instructional approaches; and the last step is to develop assessment instruments that appropriately measure this content and track the learning process.

Tests as Tools of Educational Reform

Everyone would agree that there is bound to be some back-and-forth motion in this process: deciding how children are most likely to learn something can be informed by assessments of their learning in progress. However, another camp of test reformers models the relationship explicitly as one in which tests drive instruction. Since teachers will teach and students will study what is tested, they argue for the development of tests covering content children should learn; curriculum and instruction will then fall into place. This section demonstrates how this view helped spur the rise of high-stakes tests as instruments of policy reform.

Photo *credit:* *American Guidance Service, Inc.*

Many educators urge that tasks on tests resemble the skills students should acquire in school. In mathematics, for example, tests like the one pictured above allow children to manipulate materials or use tools such as calculators.

Educational testing has long been viewed as a means to enforce accountability, inform education policy, evaluate educational progress, and reform the structure and content of teaching and learning.[23] Beginning in the mid-1960s and continuing through the 1970s and 1980s, the reliance on tests toward all these ends began to increase at all levels of government, but especially for accountability purposes and most frequently at the State level.

As accountability became a major force in education policy, the response most often took the form of rising demand for standardized achievement testing. Although many States and the Federal Government continued to collect other school performance data (such as dropout rates and various economic indicators), testing was the vehicle of choice. At the Federal level, policymakers wrote requirements for objective evaluations (usually interpreted as standardized tests) into programs of aid to elementary and secondary schools. At the State level, legislatures in 25 States enacted statewide minimum competency tests that affected critical decisions, such as grade

[22] See e g William Mehrens and Irvin J. Lehmann, *Evaluation and Psychology*, ed. (New York, NY: CBS College Publishing, 1984); and George Madaus and Daniel Stufflebeam, *Educational of Ralph W. Tyler* (Boston, MA: Kluwer Academic Publishers, 1989).

[23] See ch. 4 for a fuller discussion of the history of educational testing in the United States.

promotion or high school graduation. And at the local level, school boards and school administrators began to look at tests as a tool for satisfying public demands for accountability, providing information about how their students compared to others, and gauging their schools' progress toward local goals.

To the chagrin of many school people, Federal, State, and local district demands for test-based accountability data often addressed different issues, with each level of government acting as if data collected for the other levels was off the mark or untrustworthy and making little effort to coordinate the multiple testing requirements. It was hardly an accident that policy makers embraced standardized tests as a means to enforce accountability; this was a tradition with roots in the earliest days of the public school movement (as described in greater detail in ch. 4).

One of the appealing aspects of tests is that they enable outsiders—parents, legislators, and the general public-to leverage the internal workings of schools. One commentator has likened tests to ''remote control' devices, affording policymakers a sense of control over classrooms from a safe distance.[24] Another appealing feature is that testing conforms to a logic that sounds right: if the stakes are high enough, then teachers and students will change their behaviors in ways that improve test scores, leading to increased learning. The facts that tests may not be designed to serve this purpose, and that higher test scores do not necessarily mean increased achievement, are often overlooked. Finally, test scores serve a powerful symbolic function. A steep trend line on a graph can be strong ammunition in political struggles over the quality of schools. Whether the data are reliable and meaningful, though, are issues that are often relegated to the fine print once the headlines have left their marks.

A Climate Ripe for Growth

The reliance on tests as policy tools and the rapid adoption of high-stakes testing programs were not the result of a carefully coordinated national strategy to improve schooling. Rather, they reflected the convergence of several demographic, social, and economic trends that began in the 1960s.

Demographic Trends

The Baby Boomer cohort was a bulge in the demographic python. And as it moved through the K-12 system in the mid-1960s and early 1970s, it created unprecedented demands on school management, particularly in urban and suburban school systems, the centers of growth. As in earlier periods of demographic change, expansion of the school population led to heightened demand for additional sources of information about student achievement, over and above the judgments of teachers and administrators. Moreover, as access to education expanded for minority, immigrant, and low-income children, and in the late-1970s for children with disabilities, schools came under increased pressure to meet the needs of a more diversified student population. Fairness in the allocation of educational opportunities, always a cornerstone of the American public school ethos, rose once again to the top of the education policy agenda.

To confront these demographic changes in an efficient way, schools acted in the 1960s and 1970s in ways that mirrored their reactions to change in decades past: they looked to the world of business, and attempted to adapt techniques such as consolidation, standardization, classification, and, some might argue, bureaucratization. Small districts and rural districts that had lost population to urban and suburban areas consolidated; between school years 1963-64 and 1973-74, the number of public school districts in the United States decreased almost by one-half---horn over 31,000 to less than 16,000. Moreover, school systems of all types began relying more on tests to obtain information on larger student bodies in an efficient and objective manner, as well as to make decisions about sorting and tracking students within these bigger organizational structures.

Social Trends

The civil rights movement had a significant effect on American education in general and on testing policy in particular. In addition to raising issues

[24] See Larry Cuban, ''The Misuse of Tests in Education,'' OTA contractor report, Sept. 9, 1991. As described briefly in ch. 4, the use of standardized tests in schools began around the same time that expansion in the size of business led to the need for standardized data on the performance of business units. See, e.g., George Madaus, ''Testing as a Social Technology,' Inaugural Annual Boisi Lecture in Education and Public Policy, Boston College, Dec. 6, 1990; and Alfred Chandler, *Visible Hand:* (Cambridge, MA: Harvard University Press, 1977).

about student classification and disaggregation of achievement data, the civil rights movement called attention to the vast disparities that existed in the quantity and quality of education available to children from different racial and ethnic backgrounds. It also helped fuel a broader discussion of the educational inequities experienced by poor and disadvantaged children of all backgrounds, including rural white children, migrant children, and limited-English-proficient children.

Passage of the 1964 Civil Rights Act decisively settled the congressional battles over desegregation that had hampered past school aid bills, and paved the way for a significant Federal role in education. On the heels of the Civil Rights Act, Congress passed a host of social legislation—programs for education, welfare, health, labor, housing, and nutrition-all aimed at improving the lot of the economically disadvantaged. With those programs came a renewed interest in survey research and in the development of outcome-based measures to justify the money being spent.[25]

Economic Trends: Concerns About Competitiveness

The Nation's reaction to the Sputnik launch in 1957 foreshadowed the way that school systems would respond in subsequent decades to perceived threats to America's international competitiveness. Looking for ways to explain second-rate technological performance, leaders and the public seized on the apparently uninspired performance of American students in mathematics and science as a key reason why the United States was losing the space race. Consensus began to emerge that schools needed to place more emphasis on these two subjects. Congress passed the National Defense Education Act, the first substantial influx of Federal aid to elementary and secondary education, targeted at mathematics and science, and also containing a notable provision authorizing funds for guidance counseling and testing to identify high-ability students.

Variations on this pattern of concerns about student achievement igniting public debate and propelling a nationwide response were to be repeated in later decades. For example, when *A Nation at Risk* linked falling Scholastic Aptitude Test (SAT) scores with eroding economic competitiveness, it was the States that responded aggressively by adopting more rigorous graduation requirements, initiating a range of other reforms, and, in some cases, providing significant additional funding for schools (developments that led to more standardized testing, as will be noted later.)[26]

Another trend related to economics merits mention. In the 1970s, educational researchers began applying some of the principles and vocabulary of economics to education, assessing the efficiency and cost-effectiveness of education in terms of inputs and outputs. Most of these studies measured outputs in terms of standardized achievement test scores, some in conjunction with other quantitative measures.[27] This trend in the academic research mirrored the shift occurring in the broader policy community. It was during this period that Congress amended several Federal programs-including the Job Training Partnership Act and the Vocational Education Act—to emphasize outcome measures or performance standards in program evaluation.[28]

Changes in School Finance: Growth in Federal and State Support

The debut of the Federal Government as a significant partner in education during the 1960s, and the surge in State reform initiatives during the 1970s and 1980s, transformed the dynamics of school finance. In school year 1959-60, the *lion's* share of revenues supporting public elementary and

25 "Some proponents of social legislation resisted any accountability, believing that such could not be measured when including the social goals of the programs." Donald Senese, former assistant secretary for Educational Research and Improvement, personal communication August 1991.

26 *A Nation at Risk* is among the most cited *government* reports on education in the past 50 years, and arguably one of the most influential in spurring a range of school improvement efforts. It is important to note, however, that the findings in that report did not go entirely unchallenged. See, e.g., L. Stedman and Marshall Smith, "Recent Reform Proposals for American Education," fall 1983, pp.

27 For a recent review of this literature see Eric A. Hanushek, "The Economics of Schooling: Production and Efficiency in Public Schools," *Literature*, 24, 1986, pp. 1141-1 177; Richard Murnane, "Interpreting the Evidence on 'Does Money Matter?' " *Harvard Journal on* vol. summer 1991, pp. 457-464; and Henry M. Levin, "Mapping the Economies of Education: An Introductory Essay," *Educational Researcher*, vol. 18, No. 4, May 1989, pp. 13-16. It is important to note that many of the economists working in this field recognized the limitations of achievement test scores as outcome measures, but the scores did offer a relatively neat quantitative approach to estimating the input-output models of interest.

28 See, e.g., u.S. Congress, Office of Technology Assessment, "PeIfO~ce Standards for Secondary School Vocational Education," background paper of the Science, Education, and Transportation Program, March 1989, for discussion of the shift to outcome-based measures of public programs.

secondary education-almost 57 percent--came from local sources; States provided 39 percent and the Federal Government a mere 4 percent. As shown in table 2-1, by 1969-70, a few years after the Federal Elementary and Secondary Education Act had begun channeling over $1 billion annually to schools, the Federal share had risen to 8 percent, with States holding their own, and local support declining. A decade later, States had become the primary source of educational revenues, with a share approaching 47 percent. In recent years, the State share has continued to move up as the Federal share has declined, so that States now provide about one-half the funding for education.

The increase in Federal and State support brought about some important changes in school finance: it helped reduce revenue disparities between school districts, which formerly had depended on local property tax receipts for over one-half their income; and it targeted additional resources to students, subject areas, or urgent problems deemed to warrant Federal or State attention. But with new money came new overseers and greater demands for measurable results. A principal source of Federal accountability requirements was ''compensatory education, ' a program created in 1965 by Title I of the Elementary and Secondary Education Act. Renamed Chapter 1 in 1981, this program has been the cornerstone of Federal aid to elementary and secondary schools. From the beginning, legal requirements to evaluate the effectiveness of this program in meeting the educational needs of educationally disadvantaged children have resulted in increased reliance on standardized norm-referenced tests. As discussed in depth in chapter 3, the Federal Government has had a powerful impact on U.S. testing practice because of the evaluation and reporting requirements of Chapter 1 legislation.

Developments in the Testing Industry

Economic trends influenced assessment in yet another significant way. Advances in testing technology and psychometric research, accompanied by expansion of the testing industry, made wide-scale testing more affordable for school districts and more profitable for testing companies than ever before. While technological, research, and corporate developments alone did not create the demand for testing-that demand existed well before the advent of specific scoring or testing technologies-they provided powerful efficiency arguments in favor of standardized, machine-scorable tests.

Table 2-I—Sources of Revenues for Public Elementary and Secondary Schools (in percent)

	1959-60	1969-70	1979-80	1987-88
Federal	4.4%	8.0%	9.8%	6.3%
State	39.1	39.9	46.8	49.5
Local	56.5	52.1	43.4	44.1

SOURCE: U.S. Department of Education, National Center for Education Statistics, *Digest of Educational Statistics, 1990* (Washington, DC: 1991), p. 147.

But at the same time as machine-scorable testing was gaining ground as the vehicle of choice to manage the assessment demands of the period, curriculum experts and educational psychologists were busy crafting revised theories of human cognition and learning (as discussed above). Indeed, they, too, were strongly influenced by the apparent decline in American students' performance—compared to students in other nations-and by the fear of America's irreversible loss of international competitiveness. Their response, though, was to rethink thinking, and among the results emerging from this evolving line of research are prescriptions for radical changes in the technologies and uses of educational assessment.

The Net Result

Taken together, these demographic, economic, and social factors created a climate in which the use of tests as policy tools could take root and thrive. As summarized in a seminal National Academy of Sciences report:

> The most significant development in management (and testing) in recent years has been the increasing demand for central oversight of educational results. This comes partly because of the increased reliance of local schools on State funds since the late 1960s, partly because education has come to be viewed explicitly as a weapon with which to combat poverty and increased equality, and partly because of a suspicion that teachers and local administrators are falling down on the job.[29]

[29] Alexander K. Wigdor and Wendell R. Garner (eds.), *Ability Testing: Uses, Consequences, and* part 1, report Of the committee (Washington DC: National Academy Press, 1982), p. 170.

States, Tests, and Minimum Competency

Although the Federal Government has wrought changes in education of indisputable importance, the main arena for the events commonly thought of as the school reform movement has been the States. Education reform can mean many things and can be conducted in quite different ways. In general, the term connotes efforts to improve the quality of educational outcomes through changes in one or more aspects of the school system. Some reforms, such as the decentralization of decisionmaking that took place in the New York City schools in the late 1960s,[30] address the actual organization of schooling. Others focus on curriculum, teacher or administrator salary structures, or student tracking and grouping policies.[31]

Spurred by public demands for more accountability in education, States have taken on new and increasingly activist roles in education—and in education reform-over the past 15 years. In general, State-initiated reforms of the 1970s were ''top down' in nature: States identified their priorities, often in the forums of the legislature and State Board of Education, and set standards for all local school systems.

Tests have been essential components of most State-mandated reforms and have been asked to fulfill many new functions, such as determining the allocation of resources or persuading individuals and organizations to change behavior. In fact, States have been the main practitioners of high-stakes educational testing. For these reasons, the State experience with mandated reforms is a good illustration of some of the effects of externally developed standards on educational practices.

Minimum Competency Testing: Definition

Perhaps the most significant manifestation of the vigor with which States approached reform was the ***growth*** of minimum competency testing (MCT) that occurred during the late 1970s and continued into the 1980s. MCT refers to programs mandated by State or local agencies that have the following characteristics:

- All or almost all students in designated grades take paper-and-pencil tests designed to measure a set of skills deemed essential for future life and work.
- The State or locality has established a passing score or acceptable standard of performance on these tests.
- The State or locality may use test results to: a) make decisions about grade-level promotion, high school graduation, or the awarding of diplomas; b) classify students for remedial or other special services; c) allocate certain funds to school districts; or d) evaluate or certify school districts, schools, or teachers.[32]

Within this general framework, minimum competency tests can vary greatly in their design, format, uses, and applications to high-stakes decisions.

Impetus for MCT

MCT is a genuine example of a grassroots phenomenon, with the impetus coming mostly from outside the educational system.[33] Fueled first by popular writers, employers, and the media, and later by a proliferation of education reform panels, a movement began to catch fire among parents and other citizens who were already somewhat disillusioned with the schools. In the minds of this group, the symptoms of educational distress were all around, apparent to anyone who dared open his eyes: standards had been relaxed to the point that a high school diploma no longer meant anything; students were leaving school without the basic reading and mathematics skills they needed to succeed in work or higher education; pupils were being promoted to higher grades automatically, regardless of achievement; too little time was being spent on instruction and too much on "hills"; and too many teachers

[30] See, e.g., Diane Ravitch, *School Wars:* York, NY: Basic Books, 1974), especially pp. 251-404.

[31] For a review of recent school reform efforts, see, e.g., Educational Testing Service, *The Education* policy information report (Princeton, NJ: 1990). For analysis of the role of testing in the reform movements of the 1970s and 1980s, see Douglas A. Archbald, University of Delaware, and Andrew C. Porter, University of Wisconsin, Madison, ''A Retrospective and an Analysis of the Roles of Mandated Testing in Education Reform, " OTA contractor report, January 1990.

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

[33] Archbald and Porter, op. cit., footnote 31. See also Barbara Lerner, "Good News About American Education,'' vol. **91, No. 3, March 1991, pp. 19-25.**

were incompetent.[34] A symbol that became inextricably linked with deteriorating educational quality and perhaps more responsible than any other for erosion in public confidence was the steady drop in SAT scores that began in 1963 and persisted through the 1970s.[35]

This movement, which led to the adoption of MCT by many States, was an outgrowth of the "back to basics" movement of the 1970s—itself a backlash against the educational experimentation and general social permissiveness that had characterized the previous decade. A public grown suspicious of such innovations as schools without walls and student-centered learning, or the elimination of dress codes and the expansion of electives, came to believe that major changes-more rigorous standards, a curriculum rooted in the ''three Rs"—were needed. But many people believed that since local teachers and administrators were part of the problem, they could not be relied on to make the needed reforms without outside pressure. Seeking support from the Federal Government was an unappealing alternative to those who feared an infringement on State and local control of education or the enactment of Federal mandates.

Eventually public pressure focused on the States as the level of government best positioned to direct education reform. State Government was close enough to grassroots to understand community standards and needs, but possessed enough authority to put pressure on recalcitrant school districts. It was largely elected State officials-State legislators and State Board of Education members—who found themselves at the center of the debate over education reform. It is significant that elected officials, more than professional educators, took the lead on MCT Many State legislators were already sympathetic with the back to basics movement and were willing, even anxious, to show their support through sponsoring legislation. In addition, the fact that State legislators were not part of the educational establishment may explain their faith in the power of tests to bring about major change in education. Finally, as some researchers have observed: ''As non-educators, enthusiasts of competency testing [were] free to focus on the results and to pay little heed to the processes by which they might be achieved. "[36] State legislators may have viewed this freedom as a plus; by enacting MCI' they could appear to be doing something significant about education reform without seeming to encroach too much on local control or venture into instructional areas they knew little about.

The basic idea behind MCT was an appealing one to many State policymakers. In developing the tests, States could create some uniform, external standards that emphasized those skills deemed especially important to literacy and life success. By further tying these standards to promotion, graduation, or other educational way stations, it would focus instruction and learning on critical areas.[37]

The Rise of MCT

By the mid-1970s, the climate was ripe for action in many States. States had already begun to pick up a greater share of the costs of education, and the principle that he who pays the piper calls the tune is a time-honored one in the educational arena. And in many States, the use of tests as accountability tools was a well-established principle (witness the existence of State licensing examinations in a range of professional fields, or the State Regents' examinations in New York). In addition, early MCT programs in Denver, Florida, and Georgia had **set a** precedent and piqued the interest of policymakers from other States.

The major expansion in MCT that occurred during the 1970s and 1980s **was a** watershed event in testing policy. Prior to 1975, only **a** few States mandated MCT. The peak growth period for statewide compe-

---

[34] **Berk, op. cit.,** footnote 32.

[35] George **Madaus,** ''Testing and Policy-True Love, Shot Gun Wedding or Marriage of Convenience?' paper presented at the annual meeting of the National Council on Measurement in **Education,** New Orleans, LA, April 1984. The sudden (and short-lived) upturn in Scholastic Aptitude **Test** scores beginning in 1979 is evidence for some analysts of the effectiveness of the minimum competency testing movement. See Lemer, op. cit., footnote 33, for the most ardent formulation of this causal argument.

[36] **Walt Haney** and George **Madaus,** "Making Sense of the Competency **Testing** Movement, ' November 1978.

[37] **Critics took a much dimmer view** of what they saw as the **real function** Of minimum competency testing: "When penalties associated with failing a certification test are severe **enough,** instruction and study will adjust to prepare pupils to pass it. The test becomes a coercive device to influence both the curriculum and instruction. Unleashing the fear of diploma denial or retention in grade bullies the instructional delivery system into line.' P. **Airasian** and G. **Madaus,** ''Linking Testing and Instruction: Policy Issues,' summer 1983.

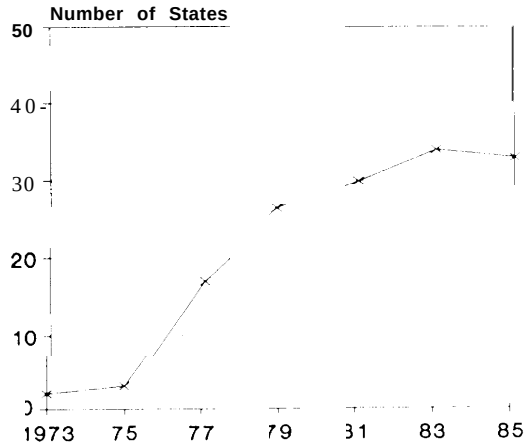
tency testing was between 1975 and 1979 (see figure 2-3). In fact, MCT accounted for most of the overall growth in educational testing in the post-1975 era. By 1980, 29 States had implemented legislation that required students to pass criterion-referenced examinations and 8 more had such legislation pending.[38] 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 minimum competency testing; 11 of these States required the test as a prerequisite for graduation.[39]

Minimum competency tests were altogether different creatures from the ''off-the-shelf' norm-referenced achievement tests that had dominated standardized testing up to that point. Most MCT instruments were custom-made in State education offices or by vendors working from State specifications, and unlike commercial tests, were designed from the start as high-stakes instruments. Most States required students to achieve a predetermined passing score for grade promotion or diploma receipt; usually students were allowed to take the test over if they did not obtain a passing score the first time. Some States mandated remediation for students who did not pass, while in other States it was optional.

Minimum competency tests are criterion-referenced; they measure performance in relation to specified skills objectives in such areas as vocabulary, reading comprehension, mathematical computation, and, in some cases, fictional skills (filling out a job application, for instance, or conducting simple financial transactions). The multiple-choice format is by far the most common, although some competency tests use other approaches, such as essay writing, oral examinations, and problem solving.

Two other features distinguish MCTs from other types of tests. First, because they use specific passing scores, they require some type of standard-setting process to determine and justify the ''cutoff score. ''[40] Since there is no freed, scientific approach to determining what knowledge a person needs to "function" in society, this can be **a** murky process. Second, MCT instruments are always administered on a census basis: each student takes the test. This does not mean, however, that the tests are not also used as instruments of *school-level* accountability. Many States and districts aggregate individual student scores to derive passing rates or average scores for entire schools. The demand for this type of comparative information has actually increased, with business leaders and policy makers often linking support for expensive reform packages to the willingness of State Education Agencies (SEAS) and school districts to accept public disclosure of test results. (Nineteen States now produce public reports comparing districts or schools on State test results.[41])

Figure 2-3--Number of States Conducting Minimum Competency Tests



SOURCE: U.S. Congress, Office of Technology Assessment, "State Educational Testing Practices," background paper of the Science, Education, andTransportation Program, December 1987; supplemented by data from Ronald A. Berk, "Minimum Competency Testing: Status and Potential," *The Future* **Barbara S. Plake and Joseph C. Witt (eds.)(Hillsdale, NJ: L. Erlbaum Associates, 1986), p. 96.**

The Second Wave of State-Mandated Reform

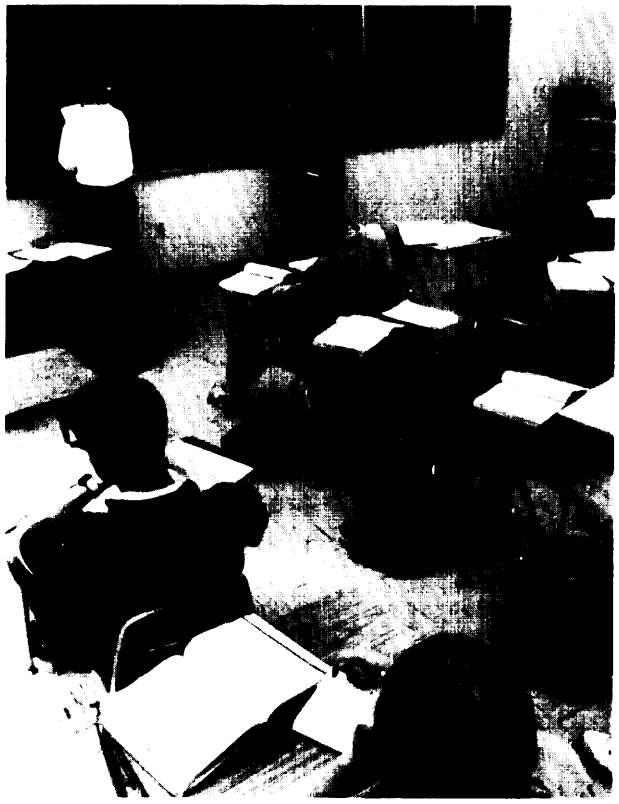
A Nation at Risk and other reform reports of the 1980s set in motion a second wave of State-

[38] Berk, op. cit., footnote 32.

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

[40] See ch. 6; Robert Linn, George Madaus, and Joseph Pedulla, ''hlinimum Competency Testing: Cautions on the State of the Art,' *Journal* November 1982, pp. 1-35; and Richard Jaeger, "An Iterative Structured Judgment Process for Establishing Standards on Competency Tests: Theory and Application," *Educational Evaluation and Policy Analysis,* vol. 4, No. 4, winter 1982, pp. 461-475.

[41] Archbald and porter, **op. cit.**, footnote 31.



credit ran So

Th 9 w tn sed cr d m
acc ed ca S m
ded ma da d w g
q me St g ade g br
q d ak co Lo S w

mandated reform. Reacting to criticisms that not enough students were taking advanced courses in science, mathematics, foreign languages, and other areas deemed critical to American international competitiveness, States assumed greater control of graduation requirements, making them more rigorous. [42] In addition, States pushed for and obtained more authority over curriculum, usually making them more prescriptive and enforcing a greater degree of consistency across the State.[43] Many States with statewide (rather than locally determined) textbook adoption policies also began scrutinizing more closely the match between their textbooks and their curriculum guidelines.[44]

Under public pressure to demonstrate gains in test scores, some States also undertook major ‘ ‘curriculum alignment” efforts, which linked curricular objectives, textbooks, lessons, instructional methods, and assessment. Curriculum alignment is a common strategy at the classroom and school level, but it is only recently that entire districts and States have experimented with it. The idea behind curriculum alignment is straightforward: if the goal is to improve test scores, then instruction should focus on what is tested. At the State level, however, alignment is not always easy to achieve. SEAS must contend with traditions of local curriculum autonomy and wide differences among school districts according to a whole range of characteristics. Moreover, the local variables that affect course content and classroom instructional practice are not easily influenced by State policies.

Nonetheless, many States have gradually tightened control over those curriculum variables that they *can* influence. Districts under pressure to raise test scores on State tests have done the same.[45] In practice, curriculum alignment can range from State officials selecting a norm-referenced test based on how well it matches with loosely defined State education goals, to States conducting exhaustive content analyses to ensure detailed matches among tests, curriculum, and textbook objectives. Off-the-shelf standardized tests—the staple of State testing for decades-increasingly were augmented or re-

[42] William Clune, Paula White, and Janice Patterson, The *Steps* Brunswick, NJ: Rutgers, The State University of New Jersey, Center for Policy Research in Education 1989).

[43] In a survey of 27 State social studies specialists, 26 said course requirements and guidelines had become more specific in the last 4 to 5 years. The investigators concluded: “Despite great differences among the states, a very strong generalization emerges from the study, namely, that the current ‘flavor’ of social studies throughout most of the country is highly prescriptive. Many prescripts have been applied in recent years to students, teachers, and curricula. ” Council of State Social Studies Specialists, *Social Kindergarten-Grade 12* (Washington DC: National Council for the Social Studies, 1986).

[44] Harriet Tyson-Bernstein, *A Conspiracy of Intentions* (Washington, DC: Council for Basic Education, 1988); and Harriet Tyson-Bernstein, “Three Portraits: Textbook Adoption Policy Changes in North Carolina, Texas and California,” occasional paper for the Institute for Educational Leadership, 1989.

[45] Ken Komoski, director of the Educational Products Information Exchange, as cited by Lynn Olson, ‘‘Districts Turn to Nonprofit Group for Help in ‘Realigning’ Curricula to Parallel Tests, ” Week, vol. 7, No. 8, Oct. 28, 1987, pp. 17, 19. Textbook manufacturers market their books in “big-market’ States and districts by demonstrating (in documentation and in sections of the books themselves) the alignment of their textbook content with State curriculum frameworks through ‘‘correlational analyses. ’

Table 2-2—Improvements in Student Achievement Associated With Curriculum Alignment

Locale	Subject	Grade	Period	Gain[a] (in percent)
Alabama	3Rs	3, 6,9	1981-86	1-13%
	3Rs	11	1983-85	4-8
Connecticut	3Rs	9	1980-84	6-16
Detroit	3Rs	12	1981-86	19
Maryland	3Rs	9	1980-86	13-25
	Social studies	9	1983-86	23
New Jersey	Reading/mathematics	9	1977-85	16-19
	Reading/mathematics	10	1982-85	8-11
South Carolina	Readiness	1	1979-85	14
	Reading/mathematics	1-3,6, 8	1981-86	12-20

[a]Figures represent the increased percentage of students who have mastered standards of quality during the Period in question.

SOURCE: W. James Popham, "The Merits of Measurement-Driven Instruction," *Phi Delta Kappan, vol. 68, No. 9,* May 1987, pp. 679-682. Note, numbers in right-most column denote the range of percentage increases across the different grade levels and tests in columns on left.

placed by custom-developed tests designed to assess State curriculum guidelines and goals.

MCT: Lessons for High-Stakes Testing

One problem with drawing conclusions about the effects or influences of State-mandated tests on school improvement is that testing is but one of many forces that shape the learning experiences of young people. Indeed, mandated testing is as much a *result* of widely held beliefs about curriculum, teaching, and learning as it is a *cause* of educational outcomes.

Even so, researchers have made some thorough analyses of State experiences with MCT and other State-mandated reforms and drawn some conclusions about their effects. In general, these researchers have concluded that the movement, which began amid such optimism, has produced results that are on the whole disappointing. A summary and analysis of key findings from studies of MCT are summarized below.

Test Score Gains

A number of States and districts can point to gains over time on minimum competency and other State tests. Gains tend to be more apparent in districts and States that have systematically pursued test and curriculum alignment. For example, on the Texas Assessment of Basic Skills in mathematics, 70 percent of ninth graders achieved mastery in 1980; by 1985, the figure had risen to 84 percent. On the reading portion of the same assessment, passing rates increased from 70 percent to 78 percent during the same period.[46] Similarly, in South Carolina, the percentage of frost graders passing the basic skills reading test rose from 70 percent in 1981 to 80 percent in 1984, and for mathematics the passing rate went from 68 to 81 percent during the same period[47] (see table 2-2).

Impressive as these gains might be, their credibility was severely undermined by analysts who looked more closely at the timing and generality of the trends in test scores.[48] Among the findings in this body of research, the most damning to the MCT movement were: 1) that scores on some tests in some places rose more rapidly and more significantly than in other places, 2) scores rose on tests even in States without MCT,[49] 3) scores began to rise before MCT could have had much impact, and 4) all States were reporting performance of their students on nationally normed achievement tests above the national average, a statistical impossibility (see box 2-D).

---

[46]Office of Technology Assessment op. cit., footnote 39, p. 272.

[47]See w. James Popham, Keith L. Cruse, Stuart Rankin, Paul Sandifer, and Paul L. Williams, ''Measurement Driven Instruction: It's on the Road, *Kappan*, vol. pp. 628-634; cited in Lorrie Shepard and Katharine Dougherty, ''Effects of High-Stakes Testing on Instruction' paper presented at the annual meeting of the American Educational Research Association, Chicago, IL, April 1991.

[48]See especially Daniel Koretz, *Trends in Educational Achievement* (Washington, DC: Congressional Budget Office, April 1986); and Congressional Budget Office, *and Implications* DC: August 1987).

@see also Gerald Bracey, **rejoinder** to Barbara Lerner, *Commentary*, vol. August 1991, p. 10.

Box 2-D—The Lake Wobegon Effect: All the children Above Average?

In radio personality Garrison Keillor's fictional town of Lake Wobegon, ''all the women are strong, all the men are good-looking, and all the children are above average. " To statisticians, of course, average is simply a representation of central tendency, and is a point drawn from an array of numbers. In many norm-referenced tests (NRTs), average represents the "median" and shows that one-half the test takers scored above this point and one-half below. It is statistically impossible for everyone to be above average-but "above average" is in some sense an American ideal.

The word average connotes a certain hum-drum, undistinguished level of achievement, especially when applied to people. Just as the citizens of mythical Lake Wobegon want all their children to be above average, teachers, principals, and parents want to show that their children are doing well.

Thus, the desire for higher test scores may overwhelm the desire to improve actual learning. similarly, in reporting scorns, calculations and methods may be used that do not give a full or accurate picture. Such excessive emphasis on test scores can compromise the value of information, as well as give misleading views of how children and schools ''rank' with regard to one another. For example, students and teachers may focus their efforts on improving performance on samples of what is to be learned, rather than on the body of knowledge from which the samples are drawn, and rising test scores may then be erroneously interpreted as reflecting genuine gains in achievement. Schools or districts seeing the scores of their students rise may be lulled into a false sense of complacency.

Or consider another possible example of how test scores used alone can lead to inaccurate inferences about achievement gains. A school system adds a number of academic high school course requirements in order to increase achievement levels. After several years, test scores go up considerably and administrators conclude that increased course requirements have raised achievement levels throughout the district. However, this gain has been attained at the expense of a number of low-achieving students dropping out. True achievement has not risen; but the lowest scoring students are no longer represented in the data. In this case, achievement test scores examined in combination with another achievement indicator (drop-out statistics) might have demonstrated that the gains were artificial.

The so-called Lake Wobegon phenomenon is by now a familiar example of how excessive focus on test scores can provide misleading information. Issued in 1987 by a group called the Friends for Education, the Lake Wobegon report asserted that all States reporting statewide test scores ranked above the national average; however, many of these same States were doing very poorly on other indicators such as graduation and literacy rates.[1]

The Lake Wobegon report sparked controversy and debate; critics charged that the report contained many inaccuracies and misunderstandings of the technical nature of test scores. Although subsequent analyses by testing experts have acknowledged that such errors do exist in the report, they have largely confirmed the basic conclusions of the Lake Wobegon report-achievement test scores can give a highly exaggerated picture of achievement.[2] Although the causes of the problem are complex and are difficult to collect data about, some of the most well-understood contributions to the Lake Wobegon phenomenon are shown below.

Dated norms. Before a standardized NRT is released, it is administered to a national sample of students to obtain "norms' '-that is, the distribution of scores for children across the Nation. That set of norms, which acts as a national standard, will then be used for about 7 years before a new form of the test is developed and "re-normed" on a new sample of children. When there are upward trends in genuine achievement, old norms become easier to master because children know more than those in prior years.[3] When old norms are used, the average performance of students today is being compared with students who took the test up to 7 years ago. Thus, today's children will appear above average.

---

[1] J. Cannell, *Nationally Normed Elementary Achievement Testing* *All 50* *the National* (Daniels, Friends for Education, 1987).

[2] See Daniel Koretz, ''Arriving in Lake Wobegon: &e Standardized I'&@ Exaggerating Achievement and Distorting Instruction?'' vol. 12, No. 2, summer 1988, pp. 8-15, 46-52; Robert L. Linn, Elizabeth Graue, and Nancy M. Sanders, "Comparing State and District Test Results to National Norma: Interpretations of Scoring 'Above the National Average,''' paper presented at the annual meeting of the American Educational Research Association San Francisco, CA, March 1989.

[3] See, e.g., Linn et al., op. cit., footnote 2.

Repeated use of nonsecure tests. Because the same tests are present in the district and given over a period of years, teachers and students become increasingly familiar with the test questions. This is one of the factors that can contribute to a very focused 'teaching to the test" and leads to the difficulty in defining the gray area between Legitimate test preparation activities and outright cheating (e.g., by having students practice actual test questions). The demarcation between legitimate test preparation activities (e.g., giving practice, coaching, and explanation of instructions to students) and dubious or even unethical practices may vary from school system to school system.[4] Even if all test preparation activities are legitimate and teaching to the testis minimized, however, some gains can probably be attributed to the increased familiarity with a particular form of a test that comes with use of a single test over a number of years.

Selection **of closely aligned tests. Standardized** achievement tests vary in content, emphasis, and form. Administrators typically select tests that most closely match the curricular objectives of their State or district. Students will tend to score higher on a test that is closely aligned with their own curricula than will students who have been taught a different, less closely aligned curricula. Because the norming group of any test is composed of schools which vary in their degree of alignment, a district with a highly aligned curriculum will score higher than the norming group. Thus, administrators who select a highly aligned test, or have a customized test made for them, will often find their students scoring better than the national norming group". . . even if their level of achievement is in some broader sense equivalent, simply because their curricula match the test more closely and thus prepare them better for it."[5]

Selection of students to be tested. Testing manuals usually explain that certain students, such as non-English speakers or special education students have been excluded from the norming sample. However, when the tests are being administered in schools, specific decisions about which children to exclude-who has mastered English well enough to take the test, for example--have to be made at the district and school level. Because many of the students who will be excluded (including truant or chronically absent children) will score well below average, these decisions can have a major impact on a school or district's average score. Schools that decide to exclude all such students are likely to have a higher average than schools with policies that attempt to include all students for whom the test can be considered valid. If the exclusionary policies for a district are more liberal than those used to obtain the norming sample, that district is likely to appear 'above average. "

Although embarrassing to some State policymakers, the Lake Wobegon report illustrated the potential mischief caused by high-stakes testing: higher test scores without more 1earning. And since the publication of the original study, other researchers have replicated the basic result. For example, one recent longitudinal study of a large urban district that uses a high-stakes commercial achievement test found that the improved performance seen over a 4-year period on that test was not confirmed when a different test was also administered in the fourth year. preliminary data indicate that the ". . . results of this district's high-stakes test overstate achievement [in mathematics] by as much as 8 academic months by the spring of grade 3."[6] Policymakers (and the public) are interested in mathematics achievement broadly defined, not just as defined by one particular test. These results suggest that". . . information provided to the public by accountability-oriented tests can be seriously misleading."[7]

The Lake Wobegon episode taught policymakers and the testing community a number of important lessons about norms, test selection, teaching to the test, and the distorting effects of high-stakes testing. Perhaps the greatest significance of the phenomenon was to demonstrate the validity of a warning that has been provided by educational testing experts for many years: no single test should ever be the basis for important policy decisions about schools or individuals.[8]

---

[4]For views on the difference between ethical and unethical test preparation activities see William A. Mehrens and John Kaminski, "Methods for Improving Standardized Test Scores: Fruitful, Fruitless, or Fraudulent?' vol. 8, spring 1989, pp. 14-22; and Thomas M. Haladyna, Susan B. Nolen, and Nancy S. Haas, "Raising Standardized Achievement Test Scores and the Origins of Test Score Pollution," *Educational Researcher, vol. 20, No. 5,* June-July 1991, pp. 2-7.

[5]Koretz, op. cit., footnote 2, p" 14

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

[7]Ibid.

[8]See, e.g., Anne Anastasi, *Psychological* York, NY: Macmillan Publishing Co., 1988).

Proponents of high-stakes testing, however, counter these arguments with data from the National Assessment of Educational Progress (NAEP). Unlike the high-stakes tests, for which score increases can be attributed to test-taking skills rather than genuine achievement, NAEP trends are considered by most experts as abetter gauge of trends in achievement.[50] Thus, the fact that NAEP scores have gone up in the 1970s and 1980s has become a linchpin in the pro-MCT argument.[51]

But, once again, closer inspection of the timing and significance of NAEP trends suggests a more complex picture, one that defies simple attribution to MCT or any other single policy. First, NAEP scores did rise in the 1970s and 1980s, but the rise actually began to be noticed as early as the 1974 assessment, well before MCT was in operation in all but one or two States.

Second, the magnitude of the rise was considerably less impressive than the magnitude recorded on other standardized tests. Although some might argue that NAEP underestimates true achievement because NAEP test takers perceive no particular incentive to do their best, even correcting for this possibility would not erase the large gap between increases on other tests and the increases on NAEP.

Third, the most impressive aspect of longitudinal analysis of NAEP scores is the narrowing of the achievement gap between minority and white students: ". . . the average achievement of Blacks and of Hispanic students is substantially higher now than a decade ago. ' '[52] This is hailed by some as the most convincing proof of the value of MCT,[53] while others note that: 1) the narrowing of the gap is explained largely by improvements at the low end of the range of achievement, 2) the overall gap between achievement of minority and white students remains quite large, and 3) gains among minority students in basic literacy and numeracy skills may have come at the expense of gains in higher order skills, which, according to NAEP data have been stagnant at best.

Undue Emphasis on Basic Skills

Prompted by these trends in NAEP, a number of researchers have investigated the hypothesis that basic skills improvements may have been made possible by a shift of instructional resources away from higher order academic skills. NAEP reports, for example, have emphasized the lack of progress in so-called higher order skills during the period of progress in basic skills. But other studies have been more optimistic. Researchers working with the Iowa Tests of Basic Skills, for example, produced evidence contradicting NAEP's: performance of comparable samples of 9-, 13-, and 17-year-olds increased between 1979 and 1985 on higher order questions even more than on basic skills items, continuing a trend observed from 1971 on.[54]

Contradictory evidence about test score trends notwithstanding, there is widespread agreement that State-mandated testing, and MCT in particular, had damaging effects on classroom behavior of teachers and students. One study combined analysis of survey data and intensive interviews with teachers and school administrators, and concluded that the testing reinforced the already excessive emphasis on basic skills and stymied local efforts to upgrade the content of education being delivered to all students. The authors of this study write:

> Although [the] ability of a Statewide testing program to control local activity maybe praiseworthy in the minds of some educational critics, the activity the program stimulated was not reform. Responding to testing did not encourage educators to reconsider the purposes of schooling; their purpose quickly became to raise scores and lower the pressure directed toward them. Responding to testing did not encourage educators to restructure their districts; they redirected time, money, and effort so that some parts of their systems could more expeditiously address the test score crisis while leaving the parts unaffected by testing or producing 'good' scores unscathed. Responding to testing did not encourage educators to rethink how they should teach or how they should administer schools; once

---

[50] For a fuller discussion of the origins and technical characteristics of the National Assessment of Educational Progress, see ch. 3,

[51] See Lerner, op. cit., footnote 33.

[52] 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, pp. 127-133.

[53] See Lerner, op. cit., footnote 33.

[54] See Elizabeth Witt, Myunghee Han, and H.D. Hoover, "Recent Trends in Achievement Test 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, Boston, MA, 1990.

again they addressed process only in the parts of their system that felt the direct impacts of testing.[55]

Narrowing Effect

While there is agreement among many studies of MCT that local districts have changed curriculum, instructional methods, and textbooks to align them more with the content of MCT instruments, there are differences of opinion about whether this is a good or bad trend. Some studies have bemoaned the narrowing effect that MCT seems to have had on instructional strategies, content coverage, and course offerings. The values embodied by MCT—that there is a fixed body of knowledge that students must absorb by a certain age, that mastery of this content is reflected in student responses to paper-and-pencil tests, and that student failure on the testis the school's responsibility to correct-tend to reinforce educational practices that are mechanical, superficial, and fragmented, such as passive learning, drill and practice, and adherence to age-grade distinctions and subject-matter boundaries.[56] Moreover, alignment to a State standard does not reflect the meaningful differences between localities.

Effects on Achievement and on Teacher Behavior

Recent research suggests that improvements on high-stakes tests do not generalize well to other measures of achievement in the same domain. For example, in one study mathematics performance on a conventional high-stakes test was found to not generalize to other tests for which students have not been specifically prepared. The authors of this study caution, therefore, that: ''. . . information provided to the public by accountability oriented tests can be seriously misleading. "5[7] The evidence is somewhat contradictory about the extent to which teachers modify their instructional practices in ways that are likely to produce higher test scores. One-half of the respondents to one nationally representative survey of eighth grade mathematics teachers (n=552) said they did not prepare students at all for mandated tests; of those who said they did, almost one-half reported spending no more than several periods a year on these efforts (and mathematics is one of the most tested areas) .58 It is also important to note, however, that of the group who said that testing influenced their instruction, 30 percent said they increased basic skills emphasis; 24 percent said they added emphasis on topics covered on the test; and 19 percent said they decreased their emphasis on project work, since it was not directly assessed by the test.[59]

Research studies that focus in particular on teachers in districts with high-stakes testing conditions—such as MCT, school evaluation tests, or externally developed course-end tests-demonstrate a greater influence of testing on curriculum and instruction. A study of four elementary classrooms with both mandated State and district objectives-based testing found that students spent up to 18 hours annually taking tests and about 54 hours receiving instruction that appeared to be directly oriented toward the tests.[60] Teachers of New York Regents courses, which have high-stakes testing at the end of the course, report spending anywhere from a few class periods to about 10 class periods (out of 175) reviewing and preparing for the examinations. Even the upper number reflects a rather modest direct effect of testing.[61]

One recent study, which sought to disentangle the effects of high-stakes testing on teaching and learning, showed fairly convincing evidence of

---

[55]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, pp. 10-11.

[56]Archbald and Porter, op. cit., footnote 31. Also see ibid.

[57]Daniel Koretz, Robert Linn, Stephen Dunbar, and Lorrie Shepard, "The Effects of High Stakes Testing on Achievement: Preliminary Findings About Generalizations Across Tests,'' paper presented at the annual meeting of the American Educational Research Association, Chicago, IL, April 1991, p. 20.

[58]Thomas Romberg, Anne Zarrinia, and Steven Williams, *The Influence of Mandated Testing On Mathematics* Instruction: *Grade* 8 (Madison, WI: National Center for Research in Mathematical Science Education, University of Wisconsin-Madison, 1989), pp. 33-39. Nevertheless, the authors concluded that changes in instruction brought about by the tests were incompatible with the kinds of changes sought by the mathematics community. See discussion below.

[59]See also Shepard, op. cit., footnote 6.

[60]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.

[61]Douglas Archbald, "Curriculum Control and Teacher Autonomy,' paper presented at the annual meeting of the American Educational Research Association, Boston, MA, April 1990.

testing influencing teacher practices. This study found that:

- teachers felt pressured to improve test scores; 79 percent reported "great" or "substantial" pressure by district administration and the media;
- teachers reported giving greater emphasis to basic skills instruction than they would have in the absence of the mandatory tests;
- one-half the teachers reported giving less emphasis to subjects not on the tests;
- one-half the teachers reported spending 4 or more weeks per year giving students worksheets and practice exercises to review content they expected to be on the test and to prepare students for the tests: 68 percent of the teachers reported conducting these preparation activities ''regularly, i.e., throughout the school year and not just in the days or weeks prior to testing; and
- the majority of teachers could identify numerous beneficial uses of the tests, such as-' . . setting instructional goals, providing feedback about student strengths and weaknesses, and identifying gaps in instruction . . . [but] these benefits . . . were offset or greatly outweighed by negative effects such as the amount of instructional time given to test preparation, the amount of stress experienced, unfair or invalid comparisons, and the demoralizing effects on teachers and students. ''[62]

These findings on the effects of high-stakes testing on teacher behavior, which the authors of the study described above caution are not necessarily generalizable, raise fundamental questions about the use of tests for instructional reform.

Misuse of MCT Data for School Comparisons

Another lesson from the MCT experience is that if test data are available they will be used to make comparisons and judgments about districts, schools, and students regardless of the data's original purpose, the ways in which it was collected, or how many caveats are issued as warnings about potential misuse. These types of comparisons, furthermore, ignore differences between school districts with large variations in student populations, resources, and other factors affecting instruction; not only are the comparisons damaging to the self esteem of students and schools, they are also potentially misleading to policymakers seeking information on how to improve the schools.

Conclusions

Viewing the MCT glass as at least half-fill, proponents have argued for more high-stakes testing and, in particular, for more high-stakes testing that covers advanced skills. Their argument is simply that if it worked for the basic skills it can work for the higher order skills.[63] These supporters of high-stakes testing argue that MCT worked because it:

- defined a single performance standard tied to powerful incentives (promotion or graduation);
- allowed teachers latitude in choosing whatever instructional methods they thought would be most appropriate to bring their students closer to the defined standards of performance;
- signaled to students the importance of acquiring basic skills in order to become productive citizens in a democracy; and
- conveyed to all students that they could acquire *the* necessary skills.

Critics contend that MCT is not a genuine tool of reform because it:

- does not provide school systems with information onto how to improve instruction, but rather serves to reinforce the instructional methods already in place;
- ignores differences between school districts with large variations in student populations, resources, and other factors affecting instruction; and
- creates conditions under which true reform is not possible, by emphasizing test scores rather than improved learning.

In the current debate over testing, it is common to hear both sides invoke the lessons of the minimum competency movement. Proponents focus on the powerful effects of high-stakes testing on clarifying and reinforcing curricula, and argue that once the right curricula are established tests will make them work. Critics fear that more high-stakes testing will reinforce outmoded curricula, provide misleading

[62]For a detailed discussion of methods, sample, and results, see Lorrie Shepard and Katherine Dougherty, ''Effects Of High Stakes Testing on Instruction,'' paper presented at the annual meeting of the American Educational Research Association, Chicago, IL, April 1991.

[63]Lerner, op. cit., footnote 33.

information to policy makers, and create artificial obstacles to educational and economic opportunity.

The positive and negative lessons of MCT, and of 100 years of prior experience with standardized tests, should inform policy for the future of testing in America. Although some of the evidence is contradictory, even confusing, one thing is clear: test-based accountability is fraught with uncertainties—it is no panacea. Specific proposals for tests intended to catalyze school improvement must be scrutinized on their individual merits, with certain cautions in mind. First, the evidence seems clear that as the stakes attached to test results heat up, so do teacher and student efforts to do better on the tests, which can lead to instructional activities that do not necessarily promote real learning. Second, there is a compelling rationale to design high-stakes tests that: a) sharpen incentives for students and teachers to practice for them, but b) contain material worth practicing for. Experience to date suggests that designing such tests is harder than originally imagined and that none has yet been implemented successfully.[64] Third, it is dubious that mandated testing alone has the potential to effect the sorts of restructuring needed to substantially reform education.

Increased Concern About the Appropriate Use of Tests

Testing policy in the United States has been influenced by the tugs of two countervailing tides: pressure for more testing with higher stakes on one hand, and cries for a slower pace and more careful examination of consequences on the other. As the influence of educational tests expanded in the 1970s and 1980s, a counterbalancing trend emerged. Individuals with different interests-parents, students, scholars, lawyers, writers, civil libertarians-began questioning the role of tests in their own and others' lives and sounding alarms about the effects of tests on individual privacy, equal opportunity, and fairness in the allocation of future opportunities. This antitesting movement encompassed a variety of sentiments, from skepticism about the validity of tests to apprehension about the damaging effects of their misuse. In addition, the trend gained momentum from the growth of consumerism and *some* key victories in Congress and the courts. The themes of this backlash against standardized testing, in the past and today, have tended to cluster around certain passion-inspiring issues: fairness, bias, due process, individual privacy, and disclosure.

In the late 1960s, for example, the idea of a "self-fulfilling prophecy' gained a foothold in the American consciousness, supported in part by a controversial study of teacher expectations. In this study, teachers were told that a test had identified a subset of children as ''bloomers' whose achievement could be expected to flourish during the school year.[65] Despite the fact that these bloomers were actually chosen at random, many showed impressive gains, outpacing their "nonbloomer" classmates. This study, which has since been found to contain many weaknesses, caught the public fancy and helped to support the arguments of many that disadvantaged children were failing in school due to teachers' low expectations about their abilities. It also alerted the public to the potential dangers of labeling children on the basis of test scores, and thus limiting their educational futures.[66]

As this example illustrates, it is not only the tests themselves that create controversy. Testing practices and policies-the ways tests are used and the types of inferences drawn from them-also create many of the problems associated with testing. There is widespread agreement among educators, analysts, measurement experts, and test publishers that tests are often used for functions for which they were not designed or validated, and that test results are often misinterpreted.

What Constitutes Fair Testing Practice?

Attempts to develop ethical and technical standards for tests and testing practices have a long

[64] The possibility that certain types of performance assessments might solve the dilemma has generated enthusiastic research and experimentation. See ch. 6.

[65] Robert Rosenthal and Lenore Jacobson, *Pygmalion in the Classroom: Pupils' Development* (New York, NY: Holt, Rinehart and Winston, 1968).

[66] For other sources on the self-fulfilling prophecy and rejoinders to the original study see Ray C. Rist, "Student Social Class and Teacher Expectations: The Self-Fulfilling Prophecy in Ghetto Education, ' *Harvard* vol. 40, No. August 1970, pp. 41 1-451; J.D. Elashoff and Richard E. Snow (eds.), (Worthington, OH: Jones, 1971); and Samuel S. Wineburg, "The Self-Fulfillment of the Self-Fulfilling Prophecy: A Critical Appraisal, ' and replys by Robert Rosenthal and Ray C. Rist, vol. 16, December 1987, pp. 28-44.

history. These efforts have been made primarily by professional groups involved in the design and administration of tests, such as psychologists and educational measurement specialists. Although discussions of such standards began at the turn of the century, the first organized efforts, at mid-century, resulted in the adoption of a formal code of ethics for psychologists in 1952 and a set of technical recommendations regarding test use developed by three professional groups in 1954.[67] This latter document, known in its most recent version as the *Standards for Educational and Psychological Testing* (hereafter referred to as the *Standards),* has been revised three times in the intervening years.[68]

Some of these technical standards pertain to tests themselves: the methods by which they should be developed, the data required to support their use, and evidence of their fairness. Although aimed primarily at the developers and publishers of tests, the standards have relevance for test users, who must evaluate the adequacy of the tests they buy or commission.

Many of the technical standards contain guidelines for test use: appropriate procedures for the selection, administration, and interpretation of tests, and guidelines affecting the rights of test takers. The two incidents quoted below, for example, represent violations of principles of appropriate testing practice.

> A high school newspaper carried a page one headline: "Meet the geniuses of the incoming class" and listed all pupils of IQ 120 and up with numerical scores. Then under a heading: ''These are not geniuses, but good enough" were listed all the rest, with IQ scores down to the 60's.
>
> ********
>
> A new battery of tests for reading readiness was introduced in a school. Instead of the customary two or three, 12 beginners were this year described by the test as not ready for reading. They were placed in a special group and given no reading instruction. The principal insisted that if the parents or anyone else tried to teach them to read 'Their little minds would crack under the strain. ' In at least two cases parents did teach them to read with normal progress in the first semester, and later mental tests showed IQ's above 120.[69]

As these examples suggest, one of the major problems with the professional *Standards* is that most of the principal interpreters of educational test results (such as policymakers, school administrators, teachers, and journalists) are unaware of them and are untrained in appropriate test use and interpretation.

A set of testing standards should consider the needs of three main participants in the testing process: 1) the test developer who constructs and markets tests, 2) the test user (usually the institution that selects tests and uses them to make some decision), and 3) the test taker who takes the test "... by choice, direction, or necessity."[70] Some form of consumer protection or assurance is needed for both the test user and the test taker, but particularly for the latter: ". . . who is still the least powerful of the three."[71] As depicted in figure 2-4, the test-taker's fate rests on the assumption that good testing practice has been upheld by both the test developer when it constructed the test and the test user (such as the school) when it selected, interpreted, and made a decision on the basis of the test. With few exceptions, the test taker has no direct contact with or access to the test developer; the test user serves as the primary falter through which testing information reaches the test taker.[72] Just as the patient undergoing an electrocardiogram must assume that the machine is soundly built and correctly calibrated, that the technician is admini-

---

[67] The American Psychological Association, the American Educational Research Association and the National Council on Measurement in Education; and Walter Haney and George Madaus, "The Evolution of Ethical and Technical Standards for Testing,' " R. Hambleton (cd.) (Amsterdam, The Netherlands: North-Holland Publishing Co., in press).

[68] In 1966, 1974, and 1985.

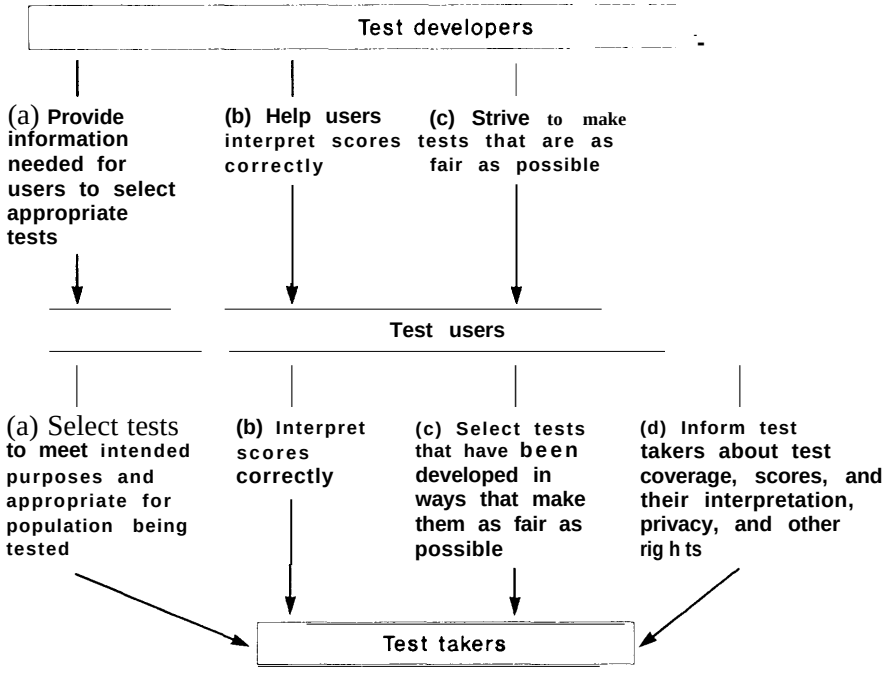
[69] American psychological Association quoted in Haney and Madaus, op. cit., footnote 67.

[70] Melvin R. Novick, ''Federal Guidelines and Professional Standards,'' vol. 10, October 1981, p. 1035.

[71] James V. Mitchell, Jr., "Testing and the Oscar Buros Lament: From Knowledge to Implementation to Use," Barbara S. Plake (cd.) (Hillsdale, NJ: L. Erlbaum Associates, 1984).

[72] For college and graduate admissions tests such as the SAT, ACT, and GRE, test takers do have direct contact with test developers. On these tests, students register directly with the test developers and receive explanations of the test, scoring methods, test-taking strategies, as well as score reports from them. Records of test scores, in these cases, remain in the hands of test developers, so privacy protection must also be assured by the developer. In contrast, the responsibility for and control of the test-takers' scores remains with the school system for most educational achievement tests administered during elementary and secondary years.

Figure 2-4-Appropriate Testing Practice in Education: Four Major Obligations of Test Developers and Test Users to Test Takers[a]



[a]This chart is based on *The Code of Fair Testing Practices in Education* which outlines four areas of major obligation to test takers: 1) developing/selecting tests, 2) interpreting scores, 3) striving for fairness, and 4) informing test takers. See the *Code* for the specific principles in each area.

NOTE: For some kinds of tests, such as college admissions tests, test developers have direct contact with test takers; in these cases, they are also obligated to the set of principles (d) regarding appropriately informing test takers.

SOURCE: Joint Committee on Testing Practices, *Code of Fair Testing Practices in Education* (Washington, DC: National Council on Measurement in Education, 1988).

stering the test properly. and that the physician is interpreting the information appropriately, so must the test taker assume that the choice of test, its method of administration, and its interpretation are correct. Currently, few mechanisms exist to assure such protection for educational tests.

The assurance of good testing practice for the test taker is further complicated by the absence of information about tests. Testing manuals, which document development and validation processes, are highly technical, and considerable training is required to evaluate the statistical properties of much of this test data. In addition, most tests are closely supervised by developers and users, in order to maintain the secrecy of test items, which is important to assuring that the test remains fair for all current and future test takers .73 The compulsory nature of most schoolwide testing programs presents yet another complication: students and their parents can exercise little choice about whether a child should be tested. In sum, a social and ethical tension exists between the need for close professional supervision of tests and the need for open public discussion and knowledge about tests by test takers—especially those whose educational opportunities may be affected by their use.

Since the 1977 version of the *Standards,* more attention has been given to the rights of the persons being tested. This attention to consumers' rights, however, appears to conflict somewhat with the need for test security. For example,

> Concerning testing, the 1977 *Standards* states that "Persons examined have the right to know results, the interpretations made, and where appropriate the original data on which final judgments were made. In light of the very next sentence, the

[73]In fact the ethical principles of psychologists prohibit them from releasing tests to unqualified persons; dissemination of any standardized test risks invalidating the test and giving some test takers an unfair advantage over others.

> modifier ''where appropriate' looms large and uncertain: "Test users avoid imparting unnecessary information which would comprise test security. . . ." An obvious question remains: When do the rights of test takers leave off and the need for test security begin?[74]

Agreement about what constitutes good testing practice is far from unanimous even among professionals; as the above example suggests, considerable latitude of interpretation is allowed for any one of the standards. For the most part each standard is a general principle, a goal to strive for and uphold; the specific criteria by which it is met are not explicitly stated. The principles governing the appropriate administration of standardized achievement tests in schools are a good example. What one school district may call legitimate test preparation activities (practice, coaching, and explanation of instructions to students), another may deem dubious or even unethical. These different interpretations are one of the principal causes of test score "inflation."[75]

Recently some professional groups have been working to translate the more technical *Standards* into principles for untrained users of tests, such as administrators, policymakers, and teachers. The *Code of Fair Testing Practices in Education*[76] (for basic provisions, see figure 2-4) attempts to outline the major obligations that professionals who use or develop educational tests have to individual test takers. These principles are widely agreed on and endorsed by professional groups as central to the fair and effective use of tests.[77]

What agreement is there about the rights of test takers? Is there a consistent set of ethical principles that should be followed? Most professional groups seem to agree that test takers should be provided with certain basic information about:

- content covered by the test and type of question formats;
- the kind of preparation the test taker should have and appropriate test-taking strategies to use (e.g., should they guess or not?);
- the uses to which test data will be put;
- the persons who will have access to test scores and the circumstances under which test scores will be released to anyone beyond those who have such access;
- the length of time test scores will be kept on record;
- available options for retesting, rescoring or canceling scores; and
- the procedures test takers and their parents or guardians may use to register complaints and have problems resolved.[78]

An important question arises regarding the principle of "informed consent, " defined by the *Standards as:*

> *The* granting of consent by the test taker to be tested on the basis of full information concerning the purpose of the testing, the persons who may receive the test scores, the use to which the test score maybe put, and such other information as may be material to the consent process.[79]

Since most children cannot give truly informed consent, an adult serving as a proxy must give consent. Although in most cases such a proxy will be the parent, there appears to be certain circumstances under which school officials are allowed to grant permission for collecting and using pupil information. Currently, the *Standards* suggest that test data collected on a schoolwide basis or by a legislated requirement are exempt from parental informed

---

[74]Haney and Madaus, op. cit., footnote

[75]See, e.g., Thomas M. Haladyna, Susan Nolen, and Nancy Haas, Standardized Achievement Test Scores and the Origins of Test Score Pollution,'' vol. June-July 1991, pp. 2-7.

[76]Authored b, the Joint Committee on Testing Practices initiated by the American Educational Research Association, *the American* Psychological Association, and the National Council on Measurement in Education in 1988. Joint Committee on Testing Practices, *in* (Washington, DC: National Council on Measurement in Education, 1988).

[77]Similar efforts are under way in other countries. For example, a number of professional groups in Canada, drawing on the experience Of the Joint Committee who developed the working on a set of principles for Canadian testing programs.

[78]See, e.g., American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, *Standards for Educational* (Washington DC: 1985); Joint Committee on Testing Practices, op. cit., footnote 76; Russell Sage Foundation, *Guidelines for* York, NY: 1969); U.S. Department of Education, Office of Educational Research and Improvement, (Washington DC: 1980).

[79]American Educational Research Association et al., op. cit., footnote 78, pp. 91-92.

consent-consent is given in this case by school officials.[80]

Informed consent also implies that the test takers are aware that they are being tested. As high-stakes tests are now conducted, children are certainly well aware that they are being tested: instructions, setting, and testing booklets all serve to clearly mark the testing session as something different from the everyday business of the classroom. Parents and children are usually notified in advance when tests will be given, in part so that parents can assure that their children are well rested and fed on testing day. Conditions and circumstances of testing are made clear so that all children have the chance to do their best.

How can parents be assured that tests are being used appropriately by schools to make decisions, particularly about individual students? One of the persistent problems with tests is that they are used for purposes not originally intended. Those being tested are not always directly informed about the uses and purposes of testing. Although it has long been considered to be the ethical responsibility of test administrators and developers to assure that tests are used only for purposes intended, there are few, if any, safeguards to assure this. Furthermore there are even fewer protections for the test score information once it is obtained-scores that sit in a child's record can be used by anyone who has access to that record whether or not that person knows anything about the particular test that was administered. It is difficult to prevent the misuse of test-based information once that information has been collected.

How is Fair Testing Practice Encouraged or Enforced?

It follows from this analysis that the first step toward fair testing practice is agreement on a set of principles or guidelines about appropriate and inappropriate test practices. Achieving such a consensus is not always a simple or clear-cut process. But given that some agreement already exists about what constitutes appropriate and inappropriate test use, how can these practices be encouraged or enforced and unfair practices be discouraged?

Right now there are four mechanisms for encouraging fair and appropriate testing practices: professional self-regulation, education, litigation, and legislation.

Professional Self-Regulation

Professional self-regulation is the primary mechanism for promoting good testing practices in education. Standards and codes for testing developed by professional associations, critical reviews of tests by experts, and individual professional codes of ethics all contribute to better testing practices among testing professionals; nevertheless, many professionals agree that these codes lack sufficiently strong enforcement mechanisms.[81] The Buros Institute of Mental Measurement has long been concerned with the education of test users and the assurance of quality tests. As part of these efforts the Institute publishes the *Mental Measurement Yearbook (MMY)*, first published in 1938, which contains critical reviews by experts of nearly ail commercially available psychological and educational tests. Recently, Institute personnel concluded that 41 percent of the tests reviewed in *The Eighth Mental Measurements Yearbook* were lacking in reliability and/or validity data.[82] In the years before his death, Oscar Buros often lamented the lack of effect that either the *Standards* or the Buros Institute had on test quality or use. In a speech in 1968, for example, Buros reported the following:

> At present, no matter how poor a test may be, if it is nicely packaged and if it promises to do all sorts of things which no test can do, the test will find many gullible buyers. When we initiated critical test reviewing in had no idea how difficult it would be to discourage the use of poorly constructed tests of unknown validity. Even the better informed test users who finally become convinced that a widely used test had no validity after all are likely to rush to use a new instrument

[80] The *Standards* read: ". . . informed consent should be obtained from test takers or their legal representatives before testing is done except (a) when testing without consent is mandated by law or governmental regulation (e.g., statewide testing programs); (b) when testing is conducted as a regular part of school activities (e.g., schoolwide testing programs and participation by schools in norming and research studies); or (c) when consent is clearly implied (e.g., application for employment or educational admissions)." Ibid., p. 85.

[81] See, e.g., George Madaus, "Public Policy and the Testing Profession—You've Never Had it so Good?" and reactions by former National Council on Measurement in Education presidents William E. Coffman, Thomas J. Fitzgibbon, Jason Millman, and Lorrie A. Shepard, in *Educational Measurement:* winter 1985, pp. 5-16.

[82] Mitchell, op. cit., footnote 71.

which promises far more than any good test can possibly deliver.[83]

In addition, the efforts by professionals to self-regulate are often aimed at developing technically sound tests and thus at the transactions between test developers and test users. Less attention has been directed toward the even more intractable problem of how to assure that tests are used appropriately *once developed and chosen by a school. How can good* testing policies be assured once a testing program, over which test takers have no choice about participation, is put in place?

Education and Public Discussion

Education and public discussion about tests, their limitations (as well as their value), and the principles of appropriate test use is the second way better testing practices could be encouraged. If the general public, parents, and test takers understood what questions to ask about tests and what protections to expect, then those who administer and choose tests would be more accountable for their testing practices. A number of testing experts believe that more open examination of test use and its social consequences could help encourage better practices on the part of those responsible for administering and interpreting tests.[84]

Teachers, principals, school boards, superintendents, and others who set testing policies for schools are another audience for educational efforts. Some proposals have recommended mandatory training for teachers to help them better understand tests and good testing practices.[85] Recently several professional associations jointly drew up a set of "Standards for Teacher Competence in Educational Assessment of Students, ' which established guidelines for what teachers should know in order to use various assessment techniques appropriately.[86] Others have called for better training of administrators and have encouraged rewarding of administrators for good assessment practices in their schools.[87]

Litigation

Litigation is the third route toward better testing practice. "Before the 1960's, the courts were rarely concerned with testing or evaluation of students. Most likely, their concern was limited because, under the standard of 'reasonableness, ' standardized testing was a subject left principally to the professional discretion of school teachers and administrators.' '[88] And since the courts showed little interest in test-related issues, as characterized in this quotation, lawyers had no incentive to bring legal actions about testing practices.

As the use of tests increased, so did their potential for causing legally significant harm to test takers.[89] The court's "hands off" approach changed in the 1970s and 1980s, with the filing of several lawsuits challenging the uses of standardized tests in education. The activism of parents, civil rights advocates, and civil liberties groups was an important spur to the development of case law in this area. Overall, however, educational tests have received far fewer legal challenges than have employment-related tests.[90]

Most litigation involving standardized educational tests involves individuals who, alone or as a class, claim violations of fundamental rights. These include the constitutional rights of due process and equal protection, and the rights guaranteed by Federal laws, such as civil rights, equal opportunity, and education of individuals with disabilities. The issues tend to center on the use of tests for classification, exclusion, and tracking, or the privacy of individual test takers. In these cases, the defendants are usually State and local school administra-

---

[83] Oscar K. Buros, "The Story Behind the Mental Measurements Yearbooks," *Guidance*, vol. 1, 1968, p. 94.

[84] Mitchell op. cit., footnote 71; and Walter Haney, "Testing Reasoning and Reasoning About Testing,'' winter pp.

[85] John R. Hills, ''Apathy Concerning Grading and Testing,'' *Kappan*, vol. 72, No. 7, March 1991, pp. 540-545; and Richard J. Stiggins, ' 'Assessment Literacy, " *Delta Kappan*, March 1991, pp. 534-539; and Robert Lynn Canady and Phyllis Riley Hotchkiss, "It's a Good Score! Just a Bad Grade," *Delta Kappan*, 1, September 1989, pp. 68-73.

[86] American Federation of Teachers, National Council on Measurement in Education, and National Education Association, ''Standards for Teacher Competence in Educational Assessment of Studenta," unpublished document, 1990.

[87] Hills, Op. cit., footnote 85.

[88] James E. Bruno and John C. Hogan, ''What Public Interest Lawyers and Educational Policymakers Need to Know About Testing: A Review of Recent Cases, Laws and Areas of Future Litigation' vol. p. 917.

[89] Donald N. Bersoff, ''Social and Legal Influences on Test Development and Usage, " in Plake (cd,), op. cit., footnote 71.

[90] See Wigdor and Garner (eds.), op. cit., footnote 29, for an overview of legal issues in employment and educational testing.

tors. Some of the earliest challenges to testing practices focused on racial discrimination. Under attack were certain classification and tracking policies—not uncommon in Southern schools resisting desegregation —that used I.Q. and other tests in ways that resulted in resegregation. Federal courts quickly barred these types of programs.[91]

Often it is the testing policy or the way a test is being used, rather than the test itself, that is challenged in court. In addition, most legal challenges have dealt with tests used for the so-called ''gatekeeping' functions: college admissions, minimum competency, or special education placement. *Thus, tests are most likely to receive legal scrutiny and challenge when they are used to make significant decisions about individual students.* In general, the courts have most often sought guidance from and upheld the *Standards.*

Some of the most significant cases involving due process and testing were spawned by the minimum competency movement. The first such case, the landmark *Debra P. v. Burlington,* claimed that the Florida law requiring students to pass a functional literacy test before obtaining a high school diploma violated the student plaintiffs' rights to due process and equal protection, as well as the Equal Educational Opportunities Act. After examining such issues as whether the test assessed skills that were actually taught, whether there was adequate notice of the requirement, whether students had access to adequate remediation, and whether they had opportunities to take the test over, the court enjoined Florida from implementing the law until 1982-83, after the vestiges of the State's formerly segregated school system were presumed to have dissipated.

As in other cases, the court referred to the *Standards in* reaching its decision. However, this case also demonstrated quite clearly the considerable latitude for interpretation and professional judgment required to translate the *Standards* into specific recommendations for practice. During the trial, two testing experts, both of whom were members of the committee who drew up the *Standards in 1974,* offered divergent and conflicting expert views about the kind of validity evidence the State of Florida should have provided.[92]

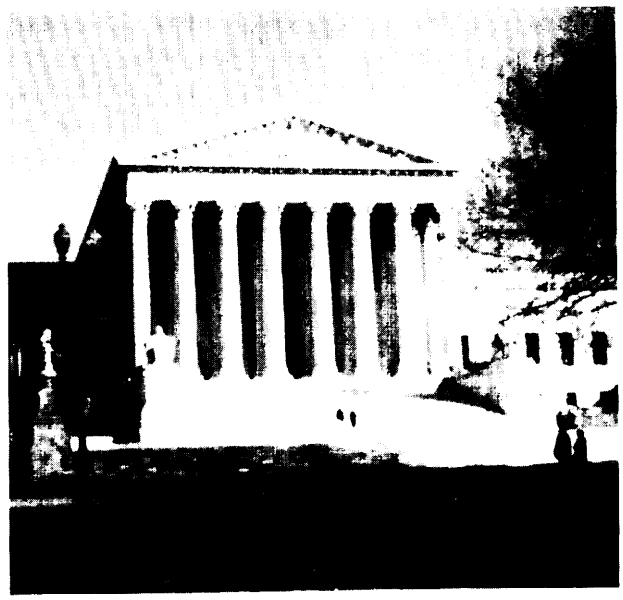


Photo credit: **Broun,**

Resorting to the courts to settle issues of good testing practice is often a last recourse. Most legal challenges to educational tests have occurred when these tests have been used for selection, certification, or placement of students.

The body of case law reveals some broad themes about how courts view tests, and some general principles about acceptable and unacceptable uses of tests. In general, courts have a great respect for well-constructed, standardized tests that are clearly tied to the curriculum. They do not find them arbitrary or irrelevant to the legitimate State interest in improving education. A minimum competency test, for example, is a reasonable method of assessing students' basic skills. In addition, Federal courts have hesitated to interfere in the education process or second guess local school district personnel.

Courts tend to look at how the results of the tests are used. If there are allegations that tests were used to deny graduation diplomas, place students in lower education tracks, or misclassify students as mentally disabled—any situations in which a test taker can claim serious injury-then the cases will be given more careful scrutiny. Cases involving historically vulnerable groups of students, such as minorities and children with disabilities, also raise flags.

91 Norman J. Chachkin, ''Testing in Elementary and Secondary Schools: Can Misuse Be Avoided?" *Law,* Bernard R. Gifford (cd.) (Boston, MA: Kluwer, 1989).

92 See Haney, **op. cit., footnote 84.**

Usually Federal lawsuits involving the use of tests have been successful only where there was a claim that the test violated some other, independently established Federal right, such as the right of due process or protection from racial discrimination.[93] State courts have shown similar deference to local judgment.

Court decisions have established some other basic guidelines about tests and their applications. Tests should accurately reflect their intended content. Students should have opportunities to learn the material on the tests in school. Students should receive adequate notice to prepare for the tests. The examinations should not be used as the sole factor in determining placement or status. The scoring procedures should accurately assess mastery of the content.[94]

Courts have protected the privacy of the parent-child relationship when testing of a very personal nature, such as certain psychological and diagnostic tests, has interfered with family relationships or the parents' rights to rear their children. On the flip side, courts have also tended to protect the security of tests by reaffirming the applicability of copyright laws to test materials.

Resorting to the courts to settle issues of good testing practice is often a last recourse. However, many testing experts as well as educators feel that courts are not the optimal arena in which to set policies regarding tests and their use. "If educators have a difficult time matching students with appropriate educational placements, judges have no experience at all. ' '[95]

One clear alternative to courts as watchdogs is to encourage school systems and policymakers to be more careful about the testing policies they implement. Many school testing policies are not set clearly and explicitly nor are they publicly available. As one litigator, involved for many years in testing and tracking litigation in schools, has written: " . . . the most difficult part of such litigation is the process of factual investigation to determine ***exactly what use is being made of what tests in a particular district."***[96]

A recent case in New York State suggests that educational administrators may have an important role to play in providing guidance and supervision regarding the fairness of school testing policies. The mother of an eighth grade student who had been excluded from enrichment programs because of her test scores on the Iowa Tests of Basic Skills (ITBS) appealed that decision. The district superintendent denied her appeal, supporting the school board's policy of using this test as the screening criteria for the enrichment program. This mother then appealed her case to the New York State Commissioner of Education who, after reviewing the evidence about the ITBS, issued an order prohibiting the district's use of test scores as the sole determinant for eligibility for educational enrichment programs. In part the order reads:

> Given the proviso in the ITBS testing manual, respondents' use of its test scores as a screening device that automatically excludes a student from further consideration for placement in an enrichment program is inconsistent with the specific guidelines provided by the developers of the ITBS test. Furthermore, because the results of a single test may be adversely affected by factors such as anxiety, illness, test-taking ability, ability to process directions or general distractibility (which have little to do with ability or achievement), use of standardized test scores as a screening device may serve to exclude pupils prematurely who are otherwise eligible. Based on the foregoing, I conclude that respondents' (the district) policy which denies a student the possibility of further consideration for placement in an enrichment program solely on the student failure to achieve above a certain score on a subpart of the ITBS is not a legitimate measure for screening a student's capacity for success in an enriched program and is, therefore, *arbitrary, capricious and contrary to sound educational policy.*[97]

As the attorney cited above notes:

As we (litigators) accumulate more knowledge about both test construction and test misuse in

[93] Chachkin, op. cit., footnote 91.

[94] Bruno and Hogan, *Op.* cit., footnote 88.

[95] William H. Clune, "Courts as Cautious Watchdogs: Constitutional and Policy Issues of Standardized Testing in Education," report prepared for the National Commission on Testing and Public Policy, 1988, p. 1.

[96] Chachkin, Op. cit., footnote 91, p. 186, emphasis added.

[97] Order #12433 of the State Education Department of New York, issued Dec. 7, 1990 by Thomas Sobol, Commissioner Of Education, p. 3, emphasis added.

educational settings, it will become easier for attorneys to gather these facts and litigation will continue and expand. For this reason, policymakers, legislators, and educational administrators are well advised to conduct their own reviews for the purpose of restricting test use to appropriate functions within their institutions and systems.[98]

Federal Legislation

Federal legislation is the fourth avenue to improved test practice, Some of the practices commonplace today in educational testing are the result of legislative efforts. In the mid-1970s, Congress passed a series of laws with significant provisions regarding testing and assessment, 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 compliance with these legislated mandates regarding appropriate testing practices.

The Family Education Rights and Privacy Act of 1974 (FERPA)—FERPA, 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. This legislation drew heavily on a set of voluntary guidelines regarding pupil records, called the Russell Sage Foundation Conference Guidelines, drawn up in 1969 by a panel of education professors, school administrators, sociologists, psychologists, professors of law, and a juvenile court judge.[99] The basic provisions of this legislation are twofold. First it establishes the right of parents to inspect school records. Second, it protects 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 written parental consent for the release of identifiable data (see table 2-3).

Table 2-3—Federally Legislated Rights Regarding Testing and School Records

1. The Family Education Rights and Privacy Act of 1974
- **A. Right to inspect records:**
  1. **Right to see all of a child’s test results that are part of the child’s official school record.**
  2. **Right to have test results explained.**
  3. **Written requests to see test results must be honored in 45 days.**
  4. **If child is over 18, only the child has the right to the record.**
- **B. Right to privacy: Rights here limit access to the official school records (including test scores) to those who have legitimate educational needs.**

ii. The Education of All Handicapped Children Act of 1975 and The Handicapped Rehabilitation Act of 1973
- **A. Right to parent involvement:**
  1. **The first time a child is considered for special education placement, the parents must be given written notice in their native language, and their permission must be obtained to test the child.**
  2. **Right to challenge the accuracy of test scores used to plan the child’s program.**
  3. **Right to file a written request to have the child tested by other than the school staff.**
  4. **Right to request a hearing if not satisfied with the school’s decision as to what are the best services for the child.**
- **B. Right to fairness in testing:**
  1. **Right of the child to be tested in the language spoken at home.**
  2. **Tests given for placement cannot discriminate on the basis of race, sex, or socioeconomic status. The tests cannot be culturally biased.**
  3. **Right of child to be tested with a test that meets special needs (e.g., Braille or orally).**
  4. **No single test score can be used to make special education placement decisions. Right to be tested in several different ways.**

SOURCE: E.B. Herndon, *Your Child and Testing* **(Washington, DC: U.S. Department of Education, National Institute of Education, October 1980), pp. 26-27.**

FERPA was an early victory for the proponents of public disclosure of test results and to date their only significant success in the Federal arena. During the 1980s, several “truth in testing” bills were introduced in Congress, intended to make tests more accessible to individuals who took them. Amid press reports about serious scoring mistakes and the publication of books accusing major testing companies of greed and arrogance, these bills gained momentum for a while, but none were enacted. The

[98] Chachkin, **op. cit.,** footnote 91, p. 186.

[99] With respect to ‘‘informed consent,’ the Russell Sage Foundation Conference op. cit., footnote 78, state that: “. . . no information should be collected from students without the prior informed consent of the child and his parents, p. 16. However, these guidelines also specify the types of data for which the notion of *representational* consent can be accepted. Representational consent means that permission to collect data is given by appropriately elected officials, such as the State Legislature or local school board. The *Guidelines go* on to clarify that: ‘‘no statement of consent, whether individual or representational, should be binding unless it is freely given after: The parents (and students where appropriate. . .) have been fully informed, preferably in writing, as to the methods by which the information will be collected; the uses to which it would be put; the methods by which it will be recorded and maintained; the time period for which it will be retained; and the persons to whom it will be available, and underwhat conditions, ’ p. 17.

drive for Federal action to ensure better testing practices has since stalled.

These bills were patterned, to some extent, on legislation passed by New York and California requiring testing companies to disclose to State commissions information about tests and testing procedures, as well as the answers to test questions. In general these laws have contained three main provisions: 1) that test developers file information about the reliability and validity of the test with a government agency, 2) that they inform students what their scores mean, how scores will be used and how access to the scores will be controlled, and 3) that individual test takers have access to corrected questions (after the test), not just the score they receive. It is largely this third provision that has made this type of legislation so controversial; the first two provisions (assuring access to information about the test's development and assuring that the test taker is appropriately informed and privacy protected) are basic tenets of good testing practice.[100] The premise behind these laws is that by increasing public scrutiny of tests, their development and their uses, potential harm to individuals can be headed off in the early stages—as when a testing company makes a scoring error-and the tests themselves will become more accurate and fair.

Legislation Affecting Individuals With Disabilities-The Rehabilitation Act of 1973 bars recipients of Federal funds from discriminating against individuals with disabilities. In the educational arena, the act has been interpreted to protect against misclassification of people as retarded, learning disabled, or mentally disabled in other ways.

One of the most consistent recommendations of testing experts is that a test score should never be used as the single criterion on which to base decisions about individuals. Significant legal challenges to the overreliance on I.Q. test scores in special education placements led to an exemplary Federal policy on test use in special education decisions. 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. Congress included eight provisions designed to protect students and ensure fair, equitable, and nondiscriminatory use of tests in implementing 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.[101] This legislation provides, then, a number of significant safeguards against the simplistic or capricious use of test scores in making educational decisions.

Conclusion: Toward Fair Testing Practice

Legal challenges have affected testing practices in some important ways. First, they have . . made the [psychological and testing] profession, as well as society in general, more sensitive to racial and cultural differences and to how apparently innocent and benign practices may perpetuate discrimination, [Second, they have] . . . alerted psychologists to the fact that they will be held responsible for their conduct. "[102] Third, by drawing some attention to the rights of test takers and responsibilities of test administrators, they have accelerated the search for better means of assessing human competencies in all spheres.[103]

Even after the enactment of FERPA and 25 years of court challenges, the current level of protection against test misuse remains rather low when compared with some other areas of consumer interest. Protections consist primarily of warnings in test publishers' manuals and a handful of State laws. Few public school districts, except for the very largest, have staffs with adequate backgrounds in psychometrics, fully trained in professional ethics and responsibilities governing test use and misuse. For most school systems, there is an abundance of public and government pressure to test students extensively, but a minimum of support to help them

[100] The truth in testing legislation has focused primarily on college and graduate admissions tests, ''. . . probably in part because such tests seem to have more visible consequences for the fate of individual test-takers than did testing of students below the college age, but surely also because college age test-takers had considerably more political clout than test-takers too young to vote. ' Mehrens and Lehmann, op. cit., footnote 22, p. 629; and Haney, op. cit., footnote 84.

[101] John Salvia and James E. Ysseldyke, Assessment 3rd ed. (Boston, MA: Houghton Mifflin Co., 1985).

[102] Donald N. Bersoff, "Testing and the Law," vol. 36, No. 10, October 1981, p. 1055.

[103] Ibid.

make ‘‘. . . proper, cautious interpretations of the data which are produced."[104]

As educational test use expands, examination of the social consequences of test use on children and schools must also be a priority. More social dialog and openness about what constitutes acceptable and unacceptable testing practices should be encouraged. Furthermore, tests used for the gatekeeping functions of selection, placement, and certification should be very carefully examined and their social consequences considered. If high-stakes testing spreads into new realms, such as a national test, we can expect to see the number of court challenges and the demand for legislative and regulatory safeguards multiply. Options for Congress to consider to foster better testing practice are discussed in chapter 1.

[104]Chachkin, op. cit., footnote 91.