achievement gap between white students and minority students and between students from high-income families and students from low-income families. Rather than finding in these tests an aid to good teaching and a means to be held accountable for it, teachers find that their professional judgment and their ability to provide good instruction are inhibited by the need to teach students to succeed on multiple-choice tests.

ASSESSMENT FOR LEARNING AND ACCOUNTABILITY

Making schools genuinely accountable for student learning will require involving teachers in the development of methods and modes of assessment that measure what students know and are able to do. In this way assessment can be tied to instruction, improvement of practice, and the creation of greater knowledge and shared standards across the educational enterprise as a whole. Many schools, districts, and states have already begun to develop alternative forms of assessment based on essays, research projects, exhibitions, and portfolios of students' work.

We need high standards, but those standards must be adapted to local needs. They must be flexible, situational, and multicultural, rather than national, mandated, and standardized. They must be tied to instructional decisions that teachers have helped to make and for which they can be held accountable.

Research evidence from the past two decades documents the fact that testing policies have not had the positive effects intended, while they have had unintended negative consequences for the quality of American schooling and for the equitable allocation of school opportunities. Both direct and indirect indicators show that these negative effects stem from the kind of tests and the ways they are used (and abused) in the quest for higher standards. To date the evidence on the use of alternative assessments is far more consistent with the current research on teaching and learning and provides a clearer connection between what is taught, what is learned, and what is assessed. This developing system of accountability can be a most powerful force for raising standards, reforming schools, and rethinking American education. K

The Implications of Testing Policy for Quality and Equality

In the long run, assessment cannot be a constructive means of reform unless we invest in more educationally useful and valid measures of student learning, Ms. Darling-Hammond warns.

By Linda Darling-Hammond

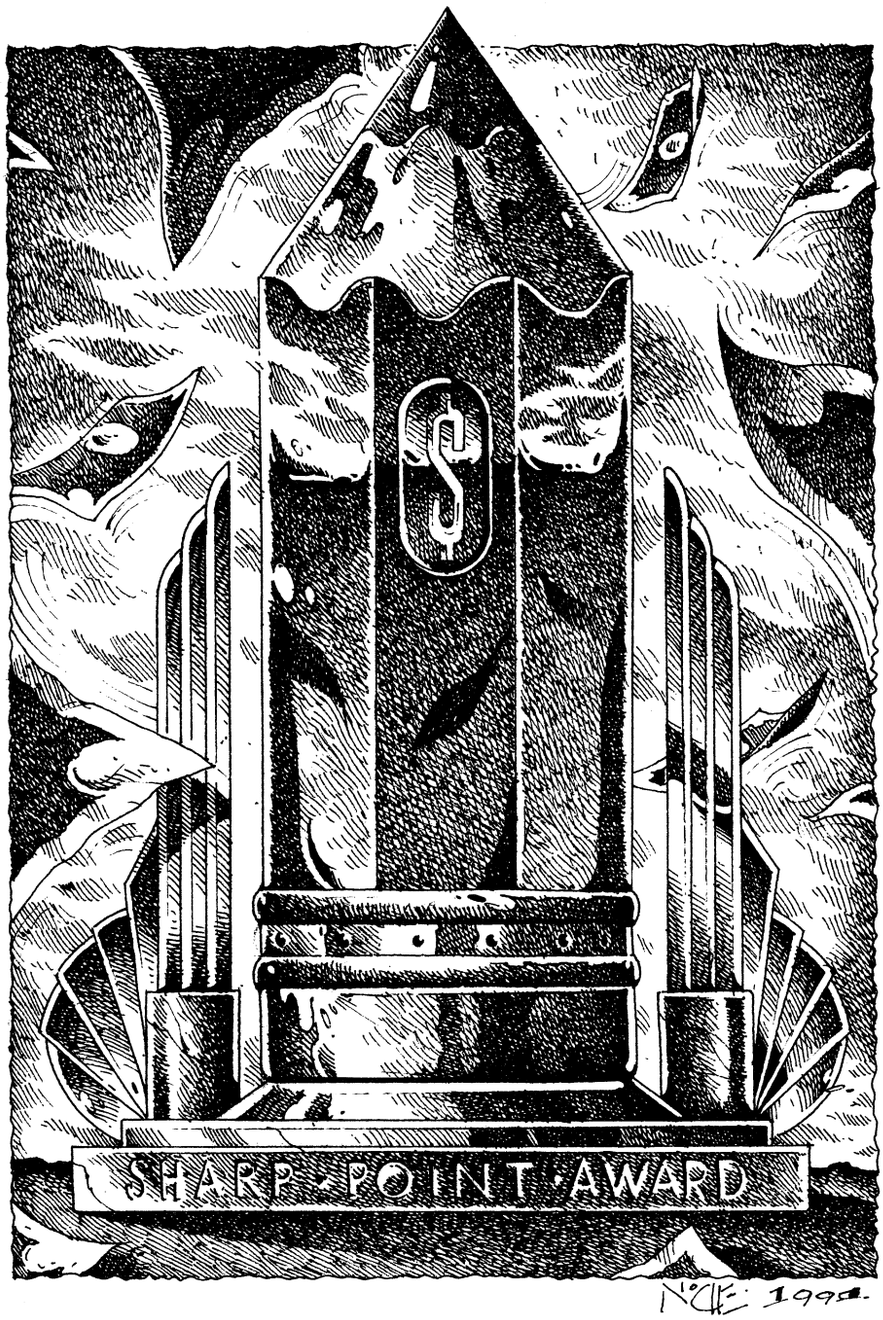
OVER THE PAST decade, efforts to improve American schools have increasingly focused on the use of standardized tests as measures of student achievement and as arbiters of decisions about student placements, teacher competence, and school quality. Some recent policies have sought to "hold schools accountable" by using tests scores to trigger rewards, sanctions, or remedial actions.

However, the evidence now available suggests that, by and large, these testing policies have not had many of the positive effects that were intended for them. Indeed, they have had many negative consequences for the quality of American schooling and for the equitable allocation of school opportunities. These negative effects stem partly from the nature of American tests and partly from the ways in which the tests have been used for educational decision making. As the discussion of national examinations moves forward, we need to ponder the lessons of our previous experiences with testing as a policy tool. Then we can incorporate what we have learned into a new approach to assessment that holds greater promise for improving teaching, learning, and schooling.

LINDA DARLING-HAMMOND is a professor of education and co-director of the Center for School Reform at Teachers College, Columbia University, New York, N.Y.

PROBLEMS WITH AMERICAN TESTS

In contrast to testing in most other countries, testing in the U.S. is primarily controlled by commercial publishers and nonschool agencies that produce norm-referenced, multiple-choice instruments designed to rank students cheaply and efficiently. These instruments were initially created to make tracking and sorting of students more efficient; they are not intended to support or enhance instruction. Because of the way the tests are constructed, they ignore a great many kinds of knowledge and types of performance that we expect from students, and they place test-takers in a passive, reactive role, rather than engage their capacities to structure tasks, generate ideas, and solve problems.[1] Even the criterion-referenced tests developed in some states tend to be poor measures of curriculum attainment and of students' abilities to undertake independent tasks.[2] Current re-



As test scores have come to play an increasingly important role in educational decisions, tests' flaws loom larger.

search on human learning and performance suggests that many tests now being used fail to measure students' higher-order cognitive abilities or to support their capacities to perform real-world tasks.[3]

These shortcomings of American tests were less problematic when the tests were seen as only one source of information about student learning among many others and when test scores were not directly tied to decisions about students and programs. However, as test scores have come to play an increasingly important role in educational decisions, their flaws have become more damaging. As schools have begun to "teach to the tests," the test scores have become ever-poorer measures of students' overall abilities, because classwork oriented toward recognizing the answers to multiple-choice questions does not heighten students' proficiency in those aspects of the subjects that are not tested, such as analysis, complex problem solving, and written and oral expression.[4]

The results can be seen in U.S. achievement trends. Since about 1970, scores on basic skills tests have been increasing, while scores on assessments of higher-order thinking have been steadily declining in virtually all subject areas.[5] Officials of the National Assessment of Educational Progress (NAEP), the National Research Council, the National Council of Teachers of English, and the National Council of Teachers of Mathematics, among others, have all attributed this decline to the schools' emphasis on tests of basic skills. They argue that the uses of the tests have corrupted teaching.

The NAEP found that "only 5% to 10% of students can move beyond initial readings of a test; most seem genuinely puzzled at requests to explain or defend their points of view." The NAEP assessors explained that current methods of teaching and testing reading call for short responses and lower-level cognitive thinking. The outcome is "an emphasis on shallow and superficial opinions at the expense of reasoned and disciplined thought. . . . [Thus] it is not surprising that students fail to develop more comprehensive thinking and analytic skills."[6]

American students' consistently poor showings on international assessments of achievement are also partly related to the orientation of the U.S. curriculum toward "basic" skills taught largely by rote. For example, the most recent comparative studies of mathematics and science performance found that American students were less likely than students in other countries to engage in science experiments, to take part in cooperative learning activities, or to use resources other

Illustration by Mario Noche

than textbooks, such as computers, calculators, or manipulatives. But they are more likely to be given lectures and worksheets. The researchers concluded that these teaching strategies accounted for poorer performance on problem-solving tasks than on tasks involving rote procedures.[7]

National data demonstrate that, during the decade when many test-oriented accountability measures were instituted in American schools, the use of teaching methods appropriate to the teaching of higher-order skills decreased. Between 1972 and 1980, public schools showed a decline in the use of such methods as student-centered discussions, the writing of essays or themes, and research projects or laboratory work.[8] This should not be surprising. As evidence from many studies indicates, when high stakes are attached to scores, tests exert a strong influence on "what is taught, how it is taught, what pupils study, how they study, and what they learn."[9]

Teachers who participated in a RAND study of the classroom effects of education policies described why this occurs. They reported that, in districts and schools that link decision making to scores on multiple-choice tests of basic skills, the tasks of preparing for tests, administering tests, and keeping records took time away from "real teaching." Under the heading of real teaching these teachers included the teaching of nontested subjects – such as science and social studies – and of such nontested modes of thinking and performance as reading books, discussing ideas, writing, engaging in creative activities, and completing projects requiring research, invention, or problem solving.[10]

Both John Goodlad and Ernest Boyer noted these trends in their major studies of schooling in the early 1980s, and both attributed them to the influence of basic skills tests on American schools. In his massive study of more than 1,000 classrooms, Goodlad found that, under the influence of state and district testing programs, students listen, read short sections in textbooks, respond briefly to questions, and take short-answer and multiple-choice quizzes. They rarely plan or initiate anything, create their own products, read or write anything substantial, or engage in analytical discussions.[11]

Misuse of basic skills tests has damaged

Boyer's similiar findings led him to conclude: "The pressure is on to teach the skills that can be counted and reported.

students the tests were intended to help.

As one teacher said, 'We are so hung up on reporting measured gains to the community on nationally normed tests that we ignore teaching those areas where it can't be done.' "[12]

In many ways, the misuse of basic skills tests has been most damaging to the very students the tests were especially intended to help. Many studies have found that students placed in the lowest tracks or in remedial programs are those most apt to experience instruction geared only to multiple-choice tests; such students work at a low cognitive level on test-oriented tasks that are profoundly disconnected from the skills they need to learn. Rarely are they given the opportunity to talk about what they know, to read real books, to write, or to construct and solve problems in mathematics, science, or other subjects.[13] In short, they are denied the opportunity to develop the capacities they will need in the future, in large part because our tests are so firmly pointed at educational goals of the past.

IMPLICATIONS FOR EQUITY

Testing policies affect students' opportunities to learn in other important ways. In addition to determining whether or not students graduate, tests are increasingly used to track students and to determine whether they can be promoted from one grade to the next. Research suggests that both practices have had harmful consequences for individual students and for American achievement generally.

Tracking. Tracking in American schools is much more extensive than in most other countries. Starting in elementary schools with the assignment of students to instructional groups and programs according to test scores, the process becomes highly formalized by junior high school. As a result, challenging curricula are rationed to a very small proportion of students, and far fewer of our students ever encounter the kinds of curriculum students in other countries typically experience.[14]

Students placed in the lower tracks are exposed to a limited, rote-oriented curriculum, and they ultimately achieve less than students of similar aptitude who are placed in academic programs or in untracked classes. Furthermore, these curricular differences explain much of the disparity between the achievement of white and minority students and between the achievement of higher- and lower-income students.[15] In this way, the uses of tests have impeded rather than supported the pursuit of high and rigorous educational goals for all students.

Retention in grade. In addition, some states and local districts have enacted policies requiring that test scores be used as the criteria for decisions regarding the promotion of students from one grade to the next. Since the policies on promotion were enacted, a substantial body of research has demonstrated that the effects of this kind of test-based decision making are much more negative than positive. When students who were retained in grade are compared to students of equal achievement levels who were promoted, the students who were retained are consistently behind on both achievement and social/emotional measures.[16] As Lorrie Shephard and Mary Lee Smith put it: "Contrary to popular beliefs, repeating a grade does *not* help students gain ground academically and has a negative impact on social adjustment and self-esteem."[17]

Furthermore, the practice of retaining students is a major contributor to increased dropout rates. Research suggests that being retained in grade once increases the likelihood of a student's dropping out by 40% to 50%. A second retention increases the likelihood by 90%.[18] Thus the policy of automatically retaining students based on their test scores has

actually produced lower achievement, lower self-esteem, and higher dropout rates.

Graduation. Perhaps the ultimate test-related sanction for students is denying them a diploma because they failed to achieve a specified test score. The rationale for this practice is that, in order to graduate, students should show they have mastered the "minimum skills" needed for employment or future education. It is assumed that tests can adequately capture whatever those skills are. Yet research indicates that neither employability nor earnings are significantly affected by students' scores on tests of basic skills, while their chances of being employed and of staying off welfare are tightly linked to graduation from high school.[19] Thus the use of tests as a sole determinant of graduation imposes heavy personal and societal costs, without obvious social benefits.

Rewards and sanctions. Finally, a few states and school districts have also tried to use student test scores to allocate rewards or sanctions to schools or teachers. The President's proposal for a national test includes such a suggestion as a means of allocating some federal funds. This use of test scores is simplistic and potentially damaging, as it will create perverse incentives for school improvement – even if we invest in better examinations than we currently have.

Because schools' average scores on any measure are sensitive to the population of students taking the test, these kinds of policies create incentives for schools to keep out those students who might lower the average scores, such as children who are handicapped, children with limited proficiency in English, or children from educationally disadvantaged environments. This kind of reward system confuses the quality of education offered by schools with the needs of the students enrolled; it will work against equity and integration and against any possibilities for fair and open school choice, because it discourages good schools from opening their doors to educationally needy students.

Dysfunctional consequences have already been reported from efforts to use average school test scores for making decisions about rewards and sanctions for schools and teachers. These include designating large numbers of low-scoring students for placement in special education so that their scores won't "count" in school reports, retaining students in grade so that their relative standing will look better on "grade-equivalent" scores, excluding low-scoring students from admission to "open enrollment" schools, and encouraging low-scoring students to drop out.[20]

Frank Smith and his colleagues describe the widespread engineering of student populations that they found in their study of a large urban school district that used performance standards as a basis for school-level sanctions:

Applying sanctions to schools with low test scores penalizes disadvantaged students twice.

> Student selection provides the greatest leverage in the short-term accountability game. . . . The easiest way to improve one's chances of winning is 1) to add some highly likely students and 2) to drop some unlikely students, while simply hanging on to those in the middle. School admissions is a central thread in the accountability fabric.[21]

Equally important, these policies will further sweeten existing incentives for talented staff members to opt for placements in schools where students are easy to teach and stability is high. Why should capable teachers and administrators risk losing rewards or incurring sanctions by volunteering to teach in schools in which many students have special needs and performance standards will be difficult to attain? Such incentives only further compromise the educational chances of disadvantaged students, who are already served by a disproportionate share of those teachers who are inexperienced, unprepared, and underqualified.[22]

Applying sanctions to schools with low test scores penalizes already disadvantaged students twice over. Having given them inadequate schools to begin with, society now punishes them further for failing to perform as well as students attending schools with more resources.[23]

The use of test scores as mechanical policy triggers will surely undermine rather than enhance accountability, since this practice shirks responsibility for careful analysis and complex decision making in favor of a simplistic and potentially quite damaging "cure." If policy makers and educators are to be truly accountable for serving students in responsible and responsive ways, then they must combine information about educational conditions and progress with knowledge about sound educational practice to arrive at strategies that are likely to encourage and support student success.

IMPROVING AMERICAN ASSESSMENT

Recognizing these problems, many schools, districts, and states have recently begun to develop different forms of assessment for students. Much like the assessment systems that prevail in many other countries, these approaches include essay examinations, research projects, scientific experiments, exhibitions, and performance in such areas as debating and the arts. They also include the use of portfolios of students' work in various subject areas and projects that require analysis, investigation, experimentation, cooperation, and written, oral, or graphic presentation of findings.

In contrast to standardized multiple-choice tests, these assessment strategies present ill-structured problems that require students to think analytically and to demonstrate their proficiency as they would in real-life situations. They include many different types of tasks – some conducted within the classroom over many months or years, others performed within examination settings. As in other countries, these assessments often require students to respond to questions from classmates or from external examiners, thus helping them learn to think through and defend their views.[24]

A number of schools (such as those in Theodore Sizer's Coalition of Essential

Schools) are engaged in creating authentic assessments of student learning, as are such school districts as Pittsburgh and Albuquerque. In addition, Vermont, California, Connecticut, Maryland, and New York are beginning to experiment with new forms of assessment, including student portfolios, performance-based assessments (for example, requiring students to perform a science experiment or solve a real-world problem using mathematical and scientific concepts), and writing assessments that engage students in complex writing tasks, sometimes requiring several days of work, including revisions.

Initiatives such as these seek to make schools genuinely accountable for helping students to acquire the kinds of higher-order skills and abilities that they will need in the world outside of school. Many of these initiatives also share another important characteristic of other countries' examinations: they involve teachers in developing and scoring the assessments, in supervising the development of student work for portfolios, and in examining their own students and those of teachers in other schools. Thus assessment is tied directly to instruction and to the improvement of practice, creating greater knowledge and shared standards across the educational enterprise as a whole.

Tests that are externally designed and imposed can never play this important role in school improvement, since they deny teachers and students the opportunity to be a part of the process of developing and wrestling with the standards. In order to achieve challenging new educational goals, members of school communities need to participate in assessment practices that are clear and open about the capabilities expected from students. In the long run, this will produce much more learning than the secretive, top-down approaches that have characterized American testing in the past.

The new initiatives will falter or flourish depending on the directions taken by federal and state policy makers seeking to track progress on the national goals. Some proposals for a national assessment *system* (as opposed to a national test) would build on these initiatives, encouraging further local and regional development in the spirit of American creativity and diversity. Such an approach is implied in some of the recommendations developed by the National Education Goals Panel. Research that could lead to such a system is being pursued at the University of Pittsburgh's Learning Research and Development Center and elsewhere.

However, the President's proposal for national achievement tests, as it is currently outlined, would turn back the clock on efforts to reform American testing and American education. As a top-down initiative based on current, primarily multiple-choice testing technologies, these tests would be far behind the innovations already being pursued in many states and localities and could undermine those efforts. (The frameworks for the existing National Assessment of Educational Progress are proposed as the basis for the "new" American achievement tests. Within the proposed time frame of only 18 months, it is certain that no major changes in testing technology could be accomplished.)

No major changes in testing technology are possible in a mere 18 months.

For many reasons, including cost, continuity with past tests, and needed connections to current teaching practice, the NAEP cannot in the near term reflect the state of the art in modern assessment. Equally important, though, is the fact that as soon as the NAEP becomes a high-stakes test that schools will feel obliged to "teach to," it will lose its value as a national indicator of progress toward our educational goals.[25] In addition, planned improvements in the quality of the NAEP will probably be sacrificed to offset the cost of extending its reach to greater numbers of students.

Rather than supporting the American traditions of experimentation and local control, the national test as currently proposed would create a de facto national curriculum – and a fairly limited one at that. This kind of national test would be likely to stifle further curriculum reform, and, by failing to involve teachers or principals in a more pervasive, local, school-based assessment process, it would lose the opportunity to encourage meaningful instructional improvement. Finally, proposals to use the test results as a basis for awarding federal funds, far from stimulating school improvement, would create perverse incentives for schools to exclude low-scoring students and for talented staff to avoid teaching in challenging schools.

In the long run, assessment cannot be a constructive lever for reform unless we invest in more educationally useful and valid measures of student learning. Rushing to create a national test in the image of our current tests will only slow our progress toward better-grounded and more challenging approaches to teaching and learning.

Investing in the creation of authentic assessments of students' actual performance is a strategy with the potential to yield much greater benefits in the long run. However, it must also be said that assessments of any kind will not be sufficient to stimulate all the changes and improvements that America needs and wants from its schools.[26] Investing in the instructional capacities of the schools themselves – and in the welfare of the students they serve – will be a necessary foundation for the success of other reforms aimed at inventing a system of American education for the 21st century.

1. Alexandra K. Wigdor and Wendell R. Garner, eds., *Ability Testing: Uses, Consequences, and Controversies* (Washington, D.C.: National Academy Press, 1982).
2. Ibid.; and George Madaus et al., "The Sensitivity of Measures of School Effectiveness," *Harvard Educational Review*, vol. 49, 1979, pp. 207-29.
3. Lauren B. Resnick, *Education and Learning to Think* (Washington, D.C.: National Academy Press, 1987); and Robert J. Sternberg, *Beyond I.Q.* (New York: Cambridge University Press, 1985).
4. Walter Haney and George Madaus, "Effects of Standardized Testing and the Future of the National Assessment of Educational Progress," working paper prepared for the NAEP Study Group, Center for the Study of Testing, Evaluation, and Educa-

tional Policy, Boston College, Chestnut Hill, Mass.; Daniel Koretz, "Arriving in Lake Wobegon: Are Standardized Tests Exaggerating Achievement and Distorting Instruction?," *American Educator*, Summer 1988, pp. 8-15, 46-52; and Linda Darling-Hammond and Arthur E. Wise, "Beyond Standardization: State Standards and School Improvement," *Elementary School Journal*, vol. 85, 1985, pp. 315-36.

5. *Crossroads in American Education* (Princeton, N.J.: Educational Testing Service, 1989).

6. National Assessment of Educational Progress, *Reading, Thinking, and Writing: Results from the 1979-80 National Assessment of Reading and Literature* (Denver: Education Commission of the States, 1981).

7. *A World of Differences: An International Assessment of Mathematics and Science* (Princeton, N.J.: Educational Testing Service, 1989); and Curtis C. McKnight et al., *The Underachieving Curriculum: Assessing U.S. School Mathematics from an International Perspective* (Champaign, Ill.: Stipes Publishing, 1987).

8. National Center for Education Statistics, *The Condition of Education 1982* (Washington, D.C.: U.S. Department of Education, 1982), p. 83.

9. Madaus et al., p. 226.

10. Darling-Hammond and Wise, op. cit.

11. John I. Goodlad, *A Place Called School: Prospects for the Future* (New York: McGraw-Hill, 1984).

12. Ernest L. Boyer, *High School: A Report on Secondary Education in America* (New York: Harper & Row, 1983).

13. Jeannie Oakes, *Keeping Track: How Schools Structure Inequality* (New Haven, Conn.: Yale University Press, 1985); Eric Cooper and John Sherk, "Addressing Urban School Reform: Issues and Alliances," *Journal of Negro Education*, vol. 58, 1989, pp. 315-31; Donna G. Davis, "A Pilot Study to Assess Equity in Selected Curricular Offerings Across Three Diverse Schools in a Large Urban School District: A Search for Methodology," paper presented at the annual meeting of the American Educational Research Association, San Francisco, 1986; and Kimberly Trimble and Robert L. Sinclair, "Ability Grouping and Differing Conditions for Learning: An Analysis of Content and Instruction in Ability-Grouped Classes," paper presented at the annual meeting of the American Educational Research Association, San Francisco, 1986.

14. McKnight et al., op. cit.

15. Oakes, op. cit.; and Valerie Lee and Anthony Bryk, "Curriculum Tracking as Mediating the Social Distribution of High School Achievement," *Sociology of Education*, vol. 61, 1988, pp. 78-94.

16. C. Thomas Holmes and Kenneth M. Matthews, "The Effects of Nonpromotion on Elementary and Junior High School Pupils: A Meta-Analysis," *Review of Educational Research*, vol. 54, 1984, pp. 225-36; and Lorrie A. Shepard and Mary Lee Smith, "Synthesis of Research on School Readiness and Kindergarten Retention," *Educational Leadership*, November 1986, pp. 78-86.

"I hope this isn't another ploy to raise your grade, Haskell."

17. Shepard and Smith, p. 84.

18. Carnegie Council on Adolescent Development, *Preparing Youth for the 21st Century* (New York: Carnegie Corporation of New York, 1990); Dale Mann, "Can We Help Dropouts? Thinking About the Undoable," in Gary Natriello, ed., *School Dropouts: Patterns and Policies* (New York: Teachers College Press, 1987); and *Before It's Too Late: Dropout Prevention in the Middle Grades* (Boston: Massachusetts Advocacy Center and the Center for Early Adolescence, 1988).

19. Richard M. Jaeger, "Legislative Perspectives on Statewide Testing: Goals, Hopes, and Desires," paper prepared for the American Educational Research Association Forum, Washington, D.C., 5 June 1991; Bruce K. Eckland, "Sociodemographic Implications of Minimum Competency Testing," in Richard M. Jaeger and Carol K. Tittle, eds., *Minimum Competency Achievement Testing: Motives, Models, Measures, and Consequences* (Berkeley, Calif.: McCutchan, 1980), pp. 124-35; and Gordon Berlin and Andrew Sum, "Toward a More Perfect Union: Basic Skills, Poor Families, and Our Economic Future," Ford Foundation Project on Social Welfare and the American Future, Occasional Paper No. 3, Ford Foundation, New York, N.Y., 1988.

20. Linda Darling-Hammond, "Policy Uses of Indicators," paper prepared for the Organization of Economic Cooperation and Development, Paris, in press; Frank Smith et al., *High School Admission and the Improvement of Schooling* (New York: New York City Board of Education, 1986); Lorrie A. Shepard and Mary Lee Smith, "Escalating Academic Demand in Kindergarten: Counterproductive Policies," *Elementary School Journal*, vol. 89, 1988, pp. 135-45; Mary Lee Smith and Lorrie A. Shephard, "Kindergarten Readiness and Retention: A Qualitative Study of Teachers' Beliefs and Practices," *American Educational Research Journal*, vol. 25, 1988, pp. 307-33; Haney and Madaus, op. cit.; and Koretz, op. cit.

21. Smith, pp. 30-31.

22. Jeannie Oakes, *Multiplying Inequalities: The Effects of Race, Social Class, and Tracking on Opportunities to Learn Mathematics and Science* (Santa Monica, Calif.: RAND Corporation, 1990); and Linda Darling-Hammond, "Teacher Quality and Equality," in John Goodlad and Pamela Keating, eds., *Access to Knowledge: An Agenda for Our Nation's Schools* (New York: College Board, 1990).

23. For documentation of the lower levels of resources available to students in high-minority, low-income, and central-city schools, see Oakes, *Multiplying Inequalities*; and Darling-Hammond, "Teacher Quality and Equality."

24. Doug A. Archbald and Fred M. Newmann, *Beyond Standardized Testing: Assessing Authentic Academic Achievement in the Secondary School* (Reston, Va.: National Association of Secondary School Principals, 1988); and Coalition of Essential Schools, "Performances and Exhibitions: The Demonstration of Mastery," *Horace*, vol. 6, March 1990, pp. 1-12.

25. For explanations of how test validity is impaired by efforts to teach to the test, see Haney and Madaus, op. cit.; Koretz, op. cit.; and Robert L. Linn, M. Elizabeth Graue, and Nancy M. Sanders, "Comparing State and District Test Results to National Norms: The Validity of Claims That 'Everyone Is Above Average,' " *Educational Measurement: Issues and Practices*, Fall 1990, pp. 5-14.

26. See Linda Darling-Hammond, "Achieving Our Goals: Superficial or Structural Reforms," *Phi Delta Kappan*, December 1990, pp. 286-95. K