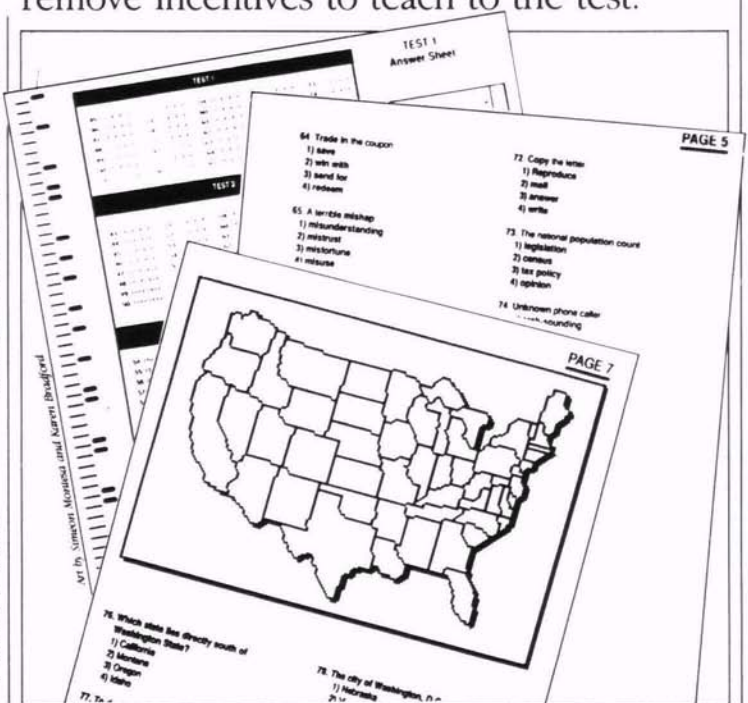
LORRIE A. SHEPARD

Why We Need Better Assessments

Educators should use a variety of assessment measures, make substantive improvements to standardized tests, and remove incentives to teach to the test.

In the United States today, standardized testing is running amok. Newspapers rank schools and districts by their test scores. Real estate agents use test scores to identify the "best" schools as selling points for expensive housing. Superintendents can be fired for low scores, and teachers can receive merit pay for high scores. Superintendents exhort principals and principals admonish teachers to raise test scores—rather than to increase learning. Occasionally school boards issue a mandate that "all students must be above the national norm," which, absurdly, is the same as requiring that 100 percent of students be above the 50th percentile.



Limitations of Standardized Tests

Current political pressures exaggerate the inherent flaws in paper-and-pencil tests. Why should the public fear that test scores are going up without a real gain in achievement (Cannell 1987)? Because large-scale testing programs are limited by political and practical considerations that undermine their fidelity to important learning outcomes.

The negotiation of content. First, the content of such tests must be negotiated. Most state-developed tests go through a consensus-building process: teams of curriculum experts and teachers agree to the content of the test. Publishers of standardized tests follow similar procedures; in addition, they do content analyses to ensure that test objectives are well matched to widely used textbooks. These procedures are sensible but they have a homogenizing effect, limiting both the breadth and depth of content coverage. The most imaginative and challenging problem-solving tasks offered uniquely by a single district are likely to be negotiated out of the test content domain. Ironically, the textbooks that commercial test framers look to for verification of coverage are governed by the same sort of consensus model, to ensure market appeal, and are in turn matched to the outlines of standardized tests (Tyson-Bernstein 1988).

The narrowing of content. Test construction is further constrained by the emphasis on basic skills, limiting the "height" as well as the depth and breadth of permissible content. Even advocates of high-stakes testing acknowledge that the tests do not cover the full range of important instructional objectives. Instead, the argument goes, "creative teachers can efficiently promote mastery of content-to-be-tested and then get on with other classroom pursuits" (Popham 1987, p. 682). The extent of this narrowing—that is, whether 10 percent or 50 percent of desirable content is sacrificed—will depend on the particular test and the accountability purpose or market that shaped it. For example, minimum competency or promotional gates tests are even narrower than standardized tests of basic skills.

Multiple-choice formats. Practical considerations also dictate test format in ways that further constrain content. Given the huge numbers of examinees and time limits, test developers use multiple-choice formats to the exclusion of tasks where students would produce or explain a correct answer. Essay tests of subject matter knowledge are rare. Although multiple-choice questions can elicit important conceptual distinctions from students, tests composed entirely of such items do not measure a respondent's ability to organize relevant information and present a coherent argument. Time limits and the need to survey as much content as possible also preclude problem sets aimed at the interrelation of facets of the same topic.

Other format problems. Similarly, reading passages are strikingly shorter and less complex than the texts students use for daily instruction. Moreover, to avoid confusing students by frequent changes in format, each test uses a small number of item types—despite evidence that students may appear to know a concept or skill when it is measured in one format but not know it if measured another way (Shepard 1988). The effects of these format limitations for the content validity of the test will increase over time, if the same test or a similar test is given each year.

Even advocates of high-stakes testing acknowledge that standardized tests do not cover the full range of important instructional objectives.

Teaching to the Test

Despite these numerous drawbacks, the public gives standardized test scores great weight. When the scores have serious consequences—and they often do—teachers will teach to the test. Indeed, it is often the explicit intention of policymakers to force teachers to address essential skills. But teaching to the test cheapens instruction and undermines the authenticity of scores as measures of what children really know, because tests are imperfect proxies even for the knowledge domains nominally covered by the tests; and they also omit important learning goals beyond the boundaries of the test domain.

Researchers have found ample evidence that testing shapes instruction. In one study, for example, elementary teachers reported taking time away from science and social studies to devote more time to tested math skills (Salmon-Cox 1982, 1984). Darling-Hammond and Wise (1985) also found that tested areas were taught at the expense of untested areas; McNeil (1988) documented that minimum proficiencies dictated by competency tests overwhelmed class time.

In addition, Darling-Hammond and Wise (1985) found that even within the bounds of test-driven content there was "dumbing down" of instruction. Teachers taught the precise content of the tests rather than underlying concepts; and skills were taught in the same format as the test rather than as they would be used in the real world. For example, teachers reported giving up essay tests because they are inefficient in preparing students for multiple-choice tests.

Conceiving instruction in the format of multiple-choice items has other far-reaching negative consequences: it leads to endless drill and practice on decontextualized skills. The notion that learning comes about by the accretion of little bits is outmoded learning theory. Current models of learning based on cognitive psychology contend that learners gain understanding when they construct their own knowledge and develop their own cognitive maps of the interconnections among

concepts and facts. Thus, real learning cannot be spoon-fed, one skill at a time.

Logically, then, to become adept at thinking and reasoning, students need practice in solving real problems and comprehending complex texts. Not surprisingly, students given instruction aimed at conceptual understanding do better on skills tests than students drilled on the skills directly (Carpenter et al. 1988). Thus, the practice of postponing higher-order thinking goals until low-level skills have been mastered is harmful. Low-achieving students suffer most from a proficiency-driven curriculum because they are consigned indefinitely to dull and repetitive skills instruction that does not enable them to grasp underlying concepts (Levin 1987).

Moreover, test-defined instruction has the effect of driving out good teachers and "deskilling" those who remain (McNeil 1988). Teachers with the least content knowledge might feel secure in leading students through worksheets fashioned after the tests, even if they become less skilled as teachers in the process. Teachers who know the most about content, however, are insulted when external mandates prevent them from using their own expertise to devise instruction. They must either find a way to resist the deskilling, as in McNeil's magnet schools, capitulate, or get out of teaching.

Finally, teaching to the test devalues the meaning of the test results themselves. By having students practice on "remarkably similar" items, teachers can improve their test performance, but gains from this type of teaching do not necessarily generalize to independent measures of the same content. Thus, test scores can go up without a commensurate gain in achievement. To really assess what students know, legislators and school boards would need a completely new test for each administration.

Substantively Better Assessments

Twenty years ago, standardized tests served as reasonable indicators of student learning. In today's political climate, tests are inadequate and misleading as measures of achievement. Assessment tasks should be redesigned—indeed, are being redesigned—to more closely resemble real learning tasks. Tests should require more complex and challenging mental processes from students. They

Authentic Assessment in California

California Assessment Program Staff

California's commitment to authentic assessment rests on a clear vision of a powerful curriculum built on a proper understanding of the nature of learning and knowledge. Thinking—a knowledge-based, discipline-oriented function—is the centerpiece of our reform curriculum. All students are encouraged to think, engage in real-world problem solving, and share in the rich, challenging curriculum that respects the integrity of the disciplines yet emphasizes the connections among them.

The right events are in alignment to redesign statewide assessment to support this curriculum: educators realize that what you test is what you get, support from them is strong, and the state is funded to revise all assessment instruments at five grade levels in all major content areas. By sampling students within schools, we can implement a performance assessment that will reward the right kinds of instruction and have the desired impact on local programs. Not incidentally, California has 15 years of experience with matrix-sampling—the key to reducing the high cost of performance assessment.

California already has a performance measure in place at two grade levels, the teacher-developed California Assessment Program (CAP) writing assessment. The use of matrix sampling enables us to assess eight types of writing at each grade level. At grade eight, for example, students write an autobiographical incident, a report of information, an evaluation, a problem solution, a firsthand biography, a story, a speculation about causes or effects, or an observation. CAP plans to develop an integrated English-language arts assessment worthy of the literature-based curriculum now being implemented throughout the state. Direct assessment methods, both oral and written, will be extended to measure students' understanding of text and their response to literature. Schoolwide pilots of portfolio assessment of reading and writing across the curriculum are under way, aiming to develop several models tailored to diverse school settings.

Guided by the California Mathematics Framework and the recently published Curriculum and Evaluation Standards of the National Council of Teachers of Mathematics, CAP's new mathematics assessment will use a variety of approaches. Strategies will include student portfolios and performance tasks that give students an opportunity to persist in complex problem-solving situations and to pursue alternative approaches. Among the procedures in the planning stage are "curriculum assessment modules," group tasks spanning three to five days, during which time students will be stopped periodically and interviewed about their work.

Science educators are experimenting with performance tasks, buoyed by their direct observations of "practical" assessment in England. They plan to directly measure key process skills—observing, comparing, communicating, organizing, relating, referring, and applying—both individually and in groups and through both oral and written tasks.

Perhaps our greatest challenge is designing an assessment to match the bold new History-Social Science Framework. To support this curriculum, teachers will be encouraged to use a full range of technology and oral, written, and performance measures, including mock trials, debates, simulations, and field trips.

Our timeline calls for statewide performance testing by 1991. Much has been accomplished on this long journey toward authentic assessment, and we are optimistic that we can meet our goal.

Authors' note: In our efforts we have relied upon and have been generously assisted by leaders at the Assessment of Performance Unit (APU) in Great Britain, Alverno College, the Coalition of Essential Schools, the Learning Research and Development Center at the University of Pittsburgh, and personnel at various projects in Canada, Australia, and New Zealand.

The CAP staff can be reached at the State of California Department of Education, California Assessment Program, 721 Capitol Mall, P.O. Box 944272, Sacramento, CA 94244-2720.

Students given instruction aimed at conceptual understanding do better on skills tests than students drilled on the skills directly.

should acknowledge more than one approach or one right answer and should place more emphasis on uncoached explanations and real student products. Structured formats should be changed often enough that there can be no benefit to practicing a skill in one particular format. Further, the dimensions of the test domain must be expanded so that teaching to the test does not imply teaching only a subset of learning goals.

At a recent conference on Assessment in the Service of Learning, Robert Glaser (1988) identified several indicators of competence in a domain of knowledge. Assessment should collect these types of evidence:

- *Coherence of knowledge*. Beginners' knowledge is spotty and superficial, but as learning progresses, understanding becomes integrated and structured. Thus assessment should tap the connectedness of concepts and the student's ability to access "interrelated chunks."
- *Principled problem solving*. Advanced learners ignore the surface features of a task and recognize underlying principles and patterns needed to solve the problem.
- *Knowledge use*. Complete understanding also means knowing the conditions that mediate the use of knowledge.
- *Automatized skills*. Basic component skills must be automatized so as to be integrated into total performance. (This is the only indicator that resembles today's by-rote measures of skills.)
- *Metacognitive or self-regulatory skills*. Assessment should determine whether students are able to monitor their own understanding, use strategies to make questions comprehensible, evaluate the relevance of accessible knowledge, and verify their own solutions.

The best way to check for these indicators is to make assessment measures resemble learning tasks. In current practice, some of the best examples of more learning-like tests are writing assessments. So long as language mechanics are not overemphasized and ideas and coherence of expression submerged, the test product represents important learning; and practice for the test is practice with real writing. Another way to assess achievement is by collecting portfolios of students' work. In the Vermont assessment proposal, for example, students and teachers would select samples of each student's best written work in English, social studies, and science. Besides encouraging more thoughtful activities in the classroom, such an approach can "enhance teacher professionalism by enabling classroom teachers to get involved in developing and scoring the assessment" (Rothman 1988).

Classroom Assessment

Better content is not the only answer to improved assessment. Policymakers must also understand and preserve the important distinction between classroom assessment and accountability testing. Cole (1988) explained the incompatibilities between measurement to serve accountability and policy goals and measurement for instructional purposes. Large-scale assessments must be formal, objective, time-efficient, cost-efficient, widely applicable, and centrally processed. Most important, results must be in a form useful to policymakers, which usually means reducing complexity to a single score. In contrast, assessments designed to support instruction are informal, teacher-mandated, adapted to local context, locally scored, sensitive to short-term change in students' knowledge, and meaningful to students. They provide immediate, detailed, and complex feedback; and they incorporate tasks that have instructional value in themselves.

Classroom assessment is conducted in a climate of greater trust than are standardized tests. Classroom tests and observations do not have to meet the same standards of accuracy. Errors made in judging individual students are less serious and more easily redressed as teachers gather new evidence. Although single teacher tests are probably less reliable (in a statistical sense) than a one-hour standardized test, the accumulation of data gathered about individual pupils in the course of a school year has much more accuracy.

Classroom assessment places no political pressure on teachers; it can address the full range of learning goals. The low-pressure environment allows the teacher to ask potentially ambiguous questions that elicit higher-order thinking. In classroom tests, teachers can continue to reach for the conceptually more important questions because they have the luxury of responding to student questions during the assessment, allowing them to restructure the problem as they understand it, and accepting more than one answer as correct. In contrast, accountability tests cannot be experimental in nature; questions aimed at more ambitious constructs are left out if they might not withstand legal scrutiny.

The differences between accountability and instructional assessment are so fundamental and necessary that it may not be desirable to merge the

Teaching to the test cheapens instruction and undermines the authenticity of scores as measures of what children really know.

Assessment tasks should be redesigned to more closely resemble real learning tasks.

two purposes. Instead, efforts to pursue an agenda like Glaser's, to develop formal measures of integrated understandings, and to train teachers in informal methods will have greater integrity if such procedures are *not* turned into accountability devices. While the substance of accountability tests must be improved by keeping conceptions of real learning in mind, their susceptibility to distortion should not be allowed to contaminate classroom uses of student data.

Safeguarding Accountability Assessment

To restore the credibility of accountability tests, we must remove both the incentives and the means to distort scores. For example, rewarding high test scores with bonuses to schools or merit pay to teachers is a clear invitation to teach to the test. Because teaching to the test can raise scores more dramatically than can instruction designed to improve achievement (Shepard 1988), the incentive system could reward the worst practices.

We must institute procedural safeguards to signify that scores will not be used punitively and thereby protect the meaning of the test data. For example, the use of scientific sampling to report for an entire state reduces the threat to individual classrooms, because classroom results are not reported. Similarly, a fall testing date removes the ownership for scores from the teachers who supervise test administration, thereby reducing the incentive to redirect instruction to the test. When sampling is used, policymakers relinquish the ability to rank every teacher in the state by test scores but gain believable data. If school and district results are announced to the media, data should also be reported on the wealth of each community and

Performance Testing in Connecticut

Joan Boykoff Baron

Connecticut has two statewide student assessment programs. The Connecticut Assessment of Educational Progress (CAEP) Program (including Connecticut's Common Core of Learning Assessment) uses sampling procedures to periodically examine the effectiveness of state programs in 11 mandated subject areas. The Connecticut Mastery Tests annually monitor the achievement of every student in grades 4, 6, and 8 in mathematics and language arts. Both programs emphasize higher-order thinking skills and performance testing.

Since its inception in 1971, the Connecticut Assessment of Educational Progress has conducted 17 assessments. Although results are reported only on a statewide aggregate basis, local districts may test all or a sample of their students and obtain those results directly from the private vendor who scores the tests. The tests have a multiple-choice section and, because of the small numbers of students tested statewide, usually include a variety of performance tasks, where students are asked to complete exercises that integrate content and process and to generate solutions to multi-step problems requiring the application of their knowledge. Where appropriate, students are given opportunities to plan, self-correct, and use suitable tools and apparatus. The tasks included in each assessment were adapted from numerous sources and were customized for each content area. For example:

- The 1983-84 business and office education assessment incorporated realistic tasks comparable to those performed in entry-level positions in accounting, secretarial, and general office settings.
- The 1984-85 science assessment tested students' ability to design experiments, control variables, construct circuits, and use microscopes.
- As part of the 1986-87 assessments in French, German, Italian, and Spanish, students wrote letters and held conversations with state-trained interviewers.

For each task, the State Department of Education developed a scoring protocol with observable behavioral descriptions specified for different levels of performance. Connecticut teachers and members of business and industry received training in the use of these protocols to rate the performance of students statewide; they served as resources later, when the protocols were made available to teachers for classroom use. By carefully delineating levels of performance and providing adequate training, we achieved high inter-rater reliability, giving Connecticut citizens confidence that different teachers would rate a student's performance the same way.

The Connecticut Mastery Tests include performance testing to assess the state's approximately 100,000 students at grades 4, 6, and 8. Students in all three grades produce a direct writing sample. As part of the language arts tests, students take notes and use them to answer listening comprehension questions in response to tape-recorded messages. On one portion of the mathematics test, all 8th graders are required to use calculators, allowing realistic problem solving that would otherwise be too time-consuming to include on a statewide test.

Currently under development is a performance assessment of Connecticut's Common Core of Learning that portrays a standard for an educated high school graduate. Scheduled for implementation in science and mathematics in a sample of high schools during 1990-91, these assessments will focus on the integration of knowledge, skills, and attitudes and will employ principles of active and collaborative learning. They may be of short or extended duration and will include the development of portfolios, simulations, extended projects, and exhibitions. A key element of the entire endeavor will be the assessment of student attitudes, attributes, and interpersonal skills in authentic contexts.

We have found that the inclusion of nontraditional and real-world performance tasks on statewide tests often increases the amount of instructional time teachers and students spend in practicing and perfecting these and related tasks. This focus on instructional time is paramount if youngsters are to gain the flexibility and adaptability necessary for today's complex world.

Joan Boykoff Baron is Education Consultant, Connecticut Department of Education, P.O. Box 2219, Room 340, Hartford, CT 06145.

the skills of incoming students, to avoid bragging by rich districts and punishment of poor districts.

The most important safeguard is not to give the same test repeatedly, year after year. Although using alternate forms of a minimum skills test does not prevent the teaching of too narrow a set of goals, new forms at least preclude rehearsing the specific vocabulary or word problems found on any one form of a test.

Another procedure designed to protect breadth of content is matrix sampling, used by the California Assessment Program. Every pupil at a grade level takes a test that is in fact a randomly assigned subpart of a larger test. Scores cannot be computed for individual pupils. However, the resulting scores reported for schools, districts, and the State of California are based on a broad content domain, making it less likely that teaching to the test will limit instruction to a narrow set of skills. Carrying this idea further, large-scale assessments could cover many broad domains by using a cycle of topics so that different subject areas could be comprehensively assessed each year.

Ironically, the National Assessment of Educational Progress began almost 20 years ago with many of these substantive and procedural features, including the assessment of 10 subject areas in 4-year cycles. These features have gradually been altered in response to cost constraints and the shift

The dimensions of the test domain must be expanded so that teaching to the test does not imply teaching only a subset of learning goals.

The more we focus on raising test scores, the more instruction is distorted, and the less credible are the scores themselves.

from a monitoring purpose to accountability. The proportion of performance tasks compared to paper-and-pencil measures has decreased. The number of subject areas has been reduced in favor of more frequent assessments in the basics. Originally, data were reported at the item level, of use only to curriculum experts. Today results are aggregated into a single score for each subject area to produce a national report card. In 1968 a sampling design based on states was purposely avoided to prevent invidious comparisons and to ensure local cooperation with the assessment. In keeping with today's climate, however, state-by-state comparisons have been mandated by Congress beginning with the 8th grade mathematics assessment of 1990.

Expanding Boundaries

Accountability testing in the 1980s is having a pernicious effect on education. Standardized tests have always been fallible, limited measures of learning goals. Today, the joint effect of limited content and inordinate weight attached to test results is magnifying the drawbacks. Tests are being asked to do too much. The more we focus on raising test scores, the more instruction is distorted, and the less credible are the scores themselves.

The content boundaries of current assessments should be expanded to include conceptual understanding and problem-solving abilities. However, substantive improvements alone will not correct current ills. If policymakers want assessment data to reflect accurately what students know, they must remove incentives to teach to the test.□

References

Cannell, J.J. (1987). *Nationally Normed Elementary Achievement Testing in America's Public Schools: How All Fifty States Are Above the National Average.* Daniels, W.V.: Friends for Education.

Carpenter, T.P., E. Fennema, P.L. Peterson, C. Chiang, and M. Loef. (April 1988). "Using Knowledge of Children's Mathematical Thinking in Classroom Teaching: An Experimental Study." New Orleans: Paper presented at the annual meeting of the American Educational Research Association.

Cole, N.S. (1988). "A Realist's Appraisal of the Prospects for Unifying Instruction and Assessment." In *Assessment in the Service of Learning: Proceedings of the 1987 ETS Invitational Conference.* Princeton, N.J.: Educational Testing Service.

Darling-Hammond, L., and A.E. Wise. (1985). "Beyond Standardization: State Standards and School Improvement." *The Elementary School Journal* 85: 315-336.

Glaser, R. (1988). "Cognitive and Environmental Perspectives on Assessing Achievement." In *Assessment in the Service of Learning: Proceedings of the 1987 ETS Invitational Conference.* Princeton, N.J.: Educational Testing Service.

Levin, H.M. (March 1987). "Accelerated Schools for Disadvantaged Students." *Educational Leadership* 44: 19-21.

McNeil, L.M. (1988). "Contradictions of Control, Part 3: Contradictions of Reform." *Phi Delta Kappan* 69: 478-485.

Popham, W.J. (1987). "The Merits of Measurement-Driven Instruction." *Phi Delta Kappan* 68: 679-682.

Rothman, R. (October 26, 1988). "Vermont Plans to Pioneer with 'Work Portfolios.'" *Education Week* 8, 1: 11.

Salmon-Cox, L. (September 1982). "MAP Math: End of Year One Report." Pittsburgh: Learning Research and Development Center.

Salmon-Cox, L. (September 1984). "MAP Reading End of Year Report." Pittsburgh: Learning Research and Development Center.

Shepard, L.A. (April 1988). "Should Instruction Be Measurement-Driven?: A Debate." Paper presented to the annual meeting of the American Educational Research Association, New Orleans.

Tyson-Bernstein, H. (1988). "America's Textbook Fiasco: A Conspiracy of Good Intentions." *American Educator* 12: 20-27, 39.

Lorrie A. Shepard is Professor, University of Colorado at Boulder, School of Education, 236 Education Bldg., Campus Box 249, Boulder, CO 80309-0249.