What Research Tells Us About Good Assessment

As our experience with alternative measures grows so does our knowledge, but we still have far to go to verify that these promising new approaches promote quality education.

JOAN L. HERMAN

Educational assessment is in a process of invention. Old models are being seriously questioned; new models are in development. Open-ended questions, exhibits, demonstrations, hands-on experiments, computer simulations, and portfolios are a few examples. The promise of the new approaches is alluring and is being effectively advanced at the national, state, and local levels all over the country.

Although the new designs offer potential, what we know about them is relatively small compared to what we have yet to discover. How close are we to having the assessments required? Here I summarize the research supporting current beliefs in testing, identify qualities of good assessment, and review the current knowledge on how to produce such measures.

Does Assessment Support Change?

Interestingly, much of the research supporting the power of testing to influence schooling is based on traditional standardized tests and concludes that such tests have a negative impact on program quality. A number of researchers — using teacher surveys, interview studies, and extended case studies — have found that accountability pressures encourage teachers and administrators to focus planning and instructional effort on test content and to devote more and more time to preparing students to do well on the tests (Dorr-Bremme and Herman 1983, Herman and Golan 1991, Kellaghan and Madaus 1991, Shepard 1991, Smith and Rottenberg 1991). Insofar as standardized tests assess only part of the curriculum, many of these researchers conclude that the time focused on test content has narrowed the curriculum by overemphasizing basic-skill subjects and neglecting higher-order thinking skills. Herman and Golan (1991), among others, have noted that such narrowing is likely to be greatest in schools serving at-risk and disadvantaged students, where there is the most pressure to improve test scores.

Cheerier pictures emerge, however, when assessments model authentic skills. Studies of the effects of California's 8th grade writing assessment program, for example, indicate that it encourages teachers both to require students to write more and to give them experience in practicing writing in a wider variety of genres.

Beyond impact on instruction, studies of some states and districts have found improved student performance over time associated with new assessment programs (Chapman 1991, Quellmalz and Burry 1983). One district in southern California, for instance, involved its teachers in the development of an analytic scoring scheme for assessing students' writing and trained a cadre of teachers from each school in its use. Over the next several years, the district witnessed an improvement in students' writing, which it attributed to the districtwide standard, the focus it provided for teachers' instruction, and the district's attention to writing instruction.

This latter point is important in interpreting both the district and the state stories: change in assessment practices was one of several factors that potentially influenced teachers' and students' performance. The California Writing Project and a number of statewide training efforts occurring at the same time gave teachers effective models of writing instruction and stressed the importance of giving students ample opportunities to write.

Pressure to improve tests scores, in the absence of serious, parallel supports for instructional improvement, however, is likely to produce serious distortions. In 1987, John Cannell, at that time a pediatrician in West Virginia, was surprised to read that the students in his state had performed above the national average on the statewide assessment program. If the largely disadvantaged students in West Virginia were scoring above the national average, who, he wondered, might be scoring below it? When he contacted the states and many large districts, almost all reported scoring above the national norm sample — a finding that was essentially replicated by the National Center for Research on Evaluation, Standards, and Student Testing (CRESST) using more rigorous

Instructional strategies must match up with assessment techniques, research finds. Here, Danville High School geometry students practice cooperative problem-solving techniques.

methods (Linn et al. 1990).

How can all students be performing "above average," a clear contradiction in the meaning of performance? Shepard concludes that the answer in large part lies in teachers' directly teaching to the test, often providing daily skill instruction in formats that closely resemble tests. Shepard and her colleagues have found that such improvements in test scores do not generalize to other measures of achievement (Koretz et al. 1991). In other words, superficial changes in instruction to improve test performance are not likely to result in meaningful learning. As a result, scores no longer represent broader student achievement, but only the content and formats included on the tests.

Ellwein and Glass (1987), looking at the effects of minimum competency testing and other assessment-based reforms, illuminate other potential distortions when serious consequences follow from test results (Glass and Ellwein 1986). When policymakers and others try to raise standards based on test results, "safety nets are strung up (in the form of exemptions, repeated trials, softening cut-scores, tutoring for retests) to catch those who fail"; and, further, "standards are determined by consideration of politically and economically acceptable pass rates, symbolic messages and appearances, and scarcely at all by a behavioral analysis of necessary skills and competencies" (Glass and Ellwein 1986). Shaped by political realities, as well as concerns for equity and future consequences, test-based standards often become diluted and have little or no influence on teachers and their instruction or on students and their learning.

What Is Good Assessment?

These findings aside, a number of current policy initiatives show continuing optimism in the power of *good* assessment, finding the problem with the assessments used, not with the basic model of accountability.

Good assessment is built on current theories of learning and cognition and grounded in views of what skills and capacities students will need for future success. To many, good assessment is also defined by what it is not: standard, traditional multiple-choice items.

According to cognitive researchers, meaningful learning is reflective, constructive, and self-regulated (Bransford and Vye 1989, Davis and Maher 1990, Marzano et al. 1988, Wittrock 1991). To *know* something is not just to have received information but to have interpreted it and related it to other knowledge one already has.

Recent studies of the integration of learning and motivation also highlight the importance of affective and metacognitive skills (McCombs 1991, Weinstein and Meyer 1991). For example, research suggests that poor thinkers and problem solvers differ from good ones not so much in the skills they possess as in their failure to use them in certain tasks. Competent thinkers or problem solvers also possess the disposition to use the skills and strategies as well as the knowledge of when to apply them.

The role of the social context of learning in shaping cognitive ability also has received recent attention. It has been noted that real-life problems often require that people work together as a group. Further, groups may facilitate learning by modeling effective thinking strategies, scaffolding complicated performances, providing mutual constructive feedback, and valuing the elements of critical thought (Resnick and Klopfer 1989).

Can We Ensure Quality?

Our new understandings of the nature and context of learning have supported the movement toward alternative assessments. A CRESST project illustrates the current enthusiasm; it has located more than 171 examples, representing the active efforts, conservatively, of 19 state departments, more than 30 school districts, and a dozen other groups.

Assuring the quality of the new assessments poses significant R&D problems. Face validity, that an assessment appears to be assessing complex thinking, is not sufficient. Essential is the notion that students' performance represents something of importance, something beyond the specific task assessed.

At the simplest level, validity indicates whether test scores accurately reflect the knowledge, skills, and abilities they are intended to measure. For traditional multiple-choice measures, concerns for validity have focused on issues of reliability (stability and consistency of performance) and patterns of relationships that may suggest whether the assessment is tapping the intended construct.

While these traditional notions are still applicable, Linn and colleagues (1991a) call for additional criteria for judging the quality of an assessment:

• *Consequences.* The consequences of an assessment influence how people respond to its results and, as Cannell's (1987) findings suggest, can rebound to influence the validity of the results themselves. This overarching criterion requires that we plan from the outset to appraise the actual use and consequences of an assessment.

• *Fairness.* Does the assessment equitably consider the cultural background of those students taking the test? Winfield and Woodard (in press) warn that standardized performance assessments are at least as likely to disadvantage students of color as traditional measures. With Winfield and Woodard, Linn and colleagues (1991a) point to additional equity problems stemming from students' "opportunity to learn" that which is assessed: Have all students had equal occasions to comprehend the complex thinking and problem-solving skills that are the targets of these new assessments?

• *Transfer and generalizability.* Do the results of an assessment support accurate generalizations about student capability? Are they reliable across raters, consistent in meaning across locales? Research on these issues raises perplexing questions about feasibility.

• *Cognitive complexity.* We cannot tell from looking at an assessment whether it actually assesses higher-level thinking. Schoenfeld (in press) cites a telling example: An award-winning teacher, whose reputation was based on his students' Regents exam performance, had drilled his students on the geometry proofs likely to appear on the exam.

• *Content quality.* The tasks selected to measure a given content domain should themselves be worthy of the time and efforts of students and raters. The content must reflect the best current understanding of the field and important aspects of a discipline that will stand the test of time.

• *Content coverage.* Coverage raises issues of curriculum match and whether the assessment tasks represent a full curriculum. As Collins and colleagues (1990) have noted, if there are gaps in coverage, teachers and students may underemphasize those topics and concepts excluded from assessment.

• *Meaningfulness.* One rationale for more contextualized assessments is that they will result in worthwhile educational experiences and in greater motivation for performance. However, additional evidence is needed to support this theory, as is investigation into the relationship between alternative assessments and student motivation to do well on them.

• *Cost/efficiency.* With more labor-intensive, performance-based assessments, greater attention will need to be given to efficient data collection designs and scoring procedures.

How Far Along Are We?

Currently, most developers of the new alternatives (with the exception of writing assessments) are at the design and prototyping stages, some distance from having validated assessments. The CRESST database project, for example, indicates that few have yet collected data on the technical quality of their assessments or about their integrity as measures of significant student learning.

Knowing how to reliably score essays and other open-ended responses is one area of relative technical strength. Research on writing assessment informs us that: (1) raters can be trained to score open-ended responses reliably and validly; (2) validity and reliability can be maintained through systematic procedures — including specified scoring schemes, sound training procedures, and ongoing reliability checks throughout the rating process; and (3) rater training reduces the number of required ratings and costs of large-scale assessment (Baker 1991, p. 3). Studies Baker reviews from the literature in the military further support the feasibility of large-scale performance assessments and the feasibility of assessing complex problem solving and team or group performance.

Trials in progress in various states, districts, and schools provide similar data. Vermont's experiments with portfolios, Connecticut's and California's pilots of hands-on math and science assessment, and Maryland's integrated assessment also indicate that it is logistically possible to administer these assessments on a large scale, schemes can be devised to score them, and teachers can be trained to reliably score them.

The generalizability of these scores — however reliable the scoring process — remains a challenge as indicated by, for example, the research of Shavelson and his colleagues on hands-on assessments in math and science (1990a,b; 1991). They essentially asked, "How many tasks does one need to get a stable estimate of a student's problem-solving capability in a given topic area?" Their answer varied from one data set to another, but the range is telling: from approximately 8 to 20 tasks were needed to obtain reliable individual level estimates. Further, they (1991) found great variability across content or topic areas within a given discipline: at least 10 different topic areas may be needed to provide dependable measures of one subject. Given the time required for administering a single hands-on experiment, these findings give pause for thought.

Also giving pause for thought are findings from Shavelson and others

which suggest that the context in which you ask students to perform influences the results. Shavelson looked at how students' performance on science experiments compared with that on simulations and on journals, all intended to measure the same aspects of problem solving. Similarly, Gearhart et al. (1992) compared how students' performance in writing was judged when based on their writing portfolios, classroom narrative assignments, and responses to a standard narrative prompt. Both studies showed substantial individual variation across the various tasks.

A study by Linn and colleagues (1991b) of the comparability of writing results across different state assessments addresses similarly thorny issues, and ones particularly germane to discussions about a national system of tests to assess progress toward national standards. Under current proposals, national standards are to be articulated and states would develop tests, tied to their state curriculums, to assess students' progress toward those standards. The results might be used for student certification, college admissions, and/or job applications, as well as to evaluate the quality of schooling at the state, district, and school levels. Because of the high stakes potentially associated with students' performance, concerns for equity demand concern for comparability of results from the different assessments.

Linn and colleagues (1991b) used the results of statewide writing assessments to examine the comparability of results from 10 states. When trained raters used their state's scoring schemes to score student papers from a different state, the results showed relatively high correlations between students' scores on the different scoring schemes. The student essays rated as the best, average, and poorest tended to be the same regardless of the specific scheme used. Such a high level of agreement of the relative ordering of student performance, according to Linn, is necessary but not sufficient for any system intending to compare results within a state to a common national standard. Also required is agreement on the absolute standard of mastery; in this area, Linn found rather substantial differences in the level of scores assigned to the same papers by different states, meaning variations in leniency and in absolute standards for performance. Assuring comparability of results, in short, will require more work.

What Remains to Be Done?

The following example illustrates both the exciting progress being made across the country and internationally and the problems that will need to be addressed if assessment is to meet its promise.

Building on past experiences with assessment in the service of accountability and on an expanded set of criteria for productive assessment, researchers at CRESST are developing new approaches to assessment, generating appropriate psychometric theory to undergird them, and exploring the process and impact of new alternatives in educational practice. For example, the center's content assessment project has produced a prototype for assessing the depth of student understanding in specific subjects (Baker et al. 1991).

Starting with students' understanding of American history, the project developed an approach that asks students to read primary source materials (for example, the Lincoln-Douglas debates) and then write an essay explaining the issues raised in the reading (explain the causes of the Civil War). Essays are then rated for quality of understanding using a scoring scheme that provides holistic and analytic ratings.

On the positive side, this project has demonstrated that it is possible to:

- design comparable, parallel tasks, based on prespecified design characteristics (the same scheme can be used to assess, for example, Civil War history, immigration history);
- use uniform scoring schemes across disciplines;
- use the same assessment to derive holistic information for large-scale assessment and diagnostic information for improving classroom practice.

But importantly, these studies also indicated that student performance on the new kinds of measures is dismally low, a finding shared by most states and districts that have tried such assessments; and that teachers need substantial training and follow-up support in both suitable assessment techniques and appropriate instructional strategies.

In conclusion, progress is being made to clarify the potential of the new alternatives, but substantial challenges remain. We must ensure that assessment supports, and does not detract from, quality education. Assessment practices themselves must be accountable to criteria that define quality assessments. These criteria force attention not only to technical issues but also to consequences of an assessment and to students' opportunity to learn that which is assessed.

Finally, changes in assessment are only part of the answer to improved instruction and learning. Schools need support to implement new instructional strategies and to institute other changes to assure that all students can achieve the complex skills that these new assessments strive to represent. □

References

Baker, E. L. (April 1991). "What Probably Works in Alternative Assessment." In

Authentic Assessment: The Rhetoric and the Reality. Symposium conducted at the annual meeting of the American Educational Research Association, Chicago.

Baker, E. L., M. Freeman, and S. Clayton. (1991). "Cognitively Sensitive Assessment of Subject Matter." In *Testing and Cognition,* edited by M. C. Wittrock and E. L. Baker. New York: Prentice-Hall.

Bransford, J. D., and N. Vye. (1989). "A Perspective on Cognitive Research and Its Implications in Instruction." In *Toward the Thinking Curriculum: Current Cognitive Research,* edited by L. B. Resnick and L. E. Klopfer. Alexandria, Va.: ASCD.

Cannell, J. J. (1987). *Nationally Normed Elementary Achievement Testing of America's Public Schools: How All 50 States Are Above the National Average.* Daniels, W.Va.: Friends for Education.

Chapman, C. (June 14, 1991). "What Have We Learned from Writing Assessment That Can Be Applied to Performance Assessment?" Presentation at ECS/CDE Alternative Assessment Conference. Breckenridge, Colo.

Collins, A., J. Hawkins, and J. Frederiksen. (April 1990). "Technology-Based Performance Assessments." Presented at the annual meeting of the American Educational Research Association, Boston.

Davis, R. B., and C. A. Maher. (1990). "Constructivist View on the Teaching of Mathematics." *Journal for Research in Mathematics Education.* Reston, Va.: NCTM.

Dorr-Bremme, D., and J. Herman. (1983). *Assessing Student Achievement: A Profile of Classroom Practices.* Los Angeles: UCLA, Center for the Study of Evaluation.

Ellwein, M. C., and G. Glass. (April 1987). *Standards of Competence: A Multi-Site Case Study of School Reform.* Los Angeles: UCLA, Center for Research on Evaluation, Standards, and Student Testing.

Gearhart, M., J. Herman, E. L. Baker, and A. K. Whittaker. (1992). *Writing Portfolios at the Elementary Level: A Study of Methods for Writing Assessment. CSE Technical Report #337.* Los Angeles: UCLA, Center for the Study of Evaluation.

Glass, G., and M. C. Ellwein. (December 1986). "Reform By Raising Test Standards." *Evaluation Comment.* Los Angeles: UCLA, Center for the Study of Evaluation.

Herman, J., and S. Golan. (1991). *Effects of Standardized Testing on Teachers and Learning — Another Look. CSE Technical Report #334.* Los Angeles: Center for the Study of Evaluation.

Kellaghan, T., and G. Madaus. (November 1991). "National Testing: Lessons for America from Europe." *Educational Leadership* 49, 3: 87-93.

Koretz, D., R. Linn, S. Dunbar, and L. Shepard. (1991). "The Effects of High Stakes Testing on Achievement." Presented at the annual meeting of the American Educational Research Association, Chicago.

Linn, R., M. Graue, and N. Sanders. (Fall 1990). "Comparing State and District Test Results to National Norms: The Validity of Claims that 'Everyone Is Above Average.' " *Educational Measurement: Issues and Practice:* 5-14.

Linn, R., E. Baker, and S. Dunbar. (1991a). "Complex, Performance-Based Assessment: Expectations and Validation Criteria." *Educational Researcher* 20, 8: 15-21.

Linn, R. L., V. L. Kiplinger, C. W. Chapman, and P. G. LeMahieu. (1991b). "Cross-State Comparability of Judgments of Student Writing: Results from the New Standards Project Workshop." Los Angeles: UCLA, Center for the Study of Evaluation.

Marzano, R., R. Brandt, and C. S. Hughes. (1988). *Dimensions of Thinking: A Framework for Curriculum and Instruction.* Alexandria, Va.: ASCD.

McCombs, B. L. (1991). "The Definition and Measurement of Primary Motivational Processes." In *Testing and Cognition,* edited by M. C. Wittrock and E. L. Baker. Englewood Cliffs, N.J.: Prentice Hall.

Quellmalz, E., J. Burry. (1983). *Analytic Scales for Assessing Students' Expository and Narrative Writing Skills. CSE Resource Paper No. 5.* Los Angeles: Center for the Study of Evaluation.

Resnick, L. B., and L. E. Klopfer. (1989). "Toward the Thinking Curriculum." In *Toward the Thinking Curriculum: Current Cognitive Research,* edited by L. B. Resnick and L. E. Klopfer. Alexandria, Va.: ASCD.

Schoenfeld, A. H. (In press). "On Mathematics as Sense-Making." In *Informal Reasoning in Education,* edited by D. N. Perkins, J. Segal, and J. Voss. Hillsdale, N.J.: Erlbaum.

Shavelson, R., G. P. Baxter, and J. Pine. (October 1990a). "What Alternative Assessments Look Like in Science." Presented at Office of Educational Research and Improvement Conference, Washington, D.C.

Shavelson, R., P. Mayberry, W. Li, and N. M. Webb. (1990b). "Generalizability of Military Performance Measurements." *Military Psychology.*

Shavelson, R., X. Gao, and G. Baxter. (November 1991). "Design Theory and Psychometrics for Complex Performance Assessment." Los Angeles: UCLA, Center for Research on Evaluation, Standards, and Student Testing.

Shepard, L. (November 1991). "Will National Tests Improve Student Learning?" *Phi Delta Kappan:* 232-238.

Smith, M. L., and C. Rottenberg. (1991). "Unintended Consequences of External Testing in Elementary Schools." *Educational Measurement: Issues and Practice* 10, 4: 7-11.

Weinstein, C., and D. Meyer. (1991). "Implications of Cognitive Psychology for Testing." In *Testing and Cognition,* edited by M. C. Wittrock and E. L. Baker. Englewood Cliffs, N.J.: Prentice Hall.

Winfield, L., and M. D. Woodard. (In press). "What About the 'Rest of Us' in Bush's America 2000?" *Education Week.*

Wittrock, M. C. (1991). "Testing and Recent Research in Cognition." In *Testing and Cognition,* edited by M. C. Wittrock and E. L. Baker. Englewood Cliffs, N.J.: Prentice Hall.

Author's Note: The work reported herein was supported under the Educational Research and Development Center Program cooperative agreement R117G10027 and CFDA catalog number 84.117G as administered by the Office of Educational Research and Improvement, U.S. Department of Education. The findings and opinions expressed here do not reflect the position or policies of the Office of Educational Research and Improvement or the U.S. Department of Education.

Joan L. Herman is Associate Director, University of California-Los Angeles National Center for Research on Evaluation, Standards, and Student Testing, Center for the Study of Evaluation, UCLA Graduate School of Education, 405 Hilgard Ave., Los Angeles, CA 90024-1522.