DOCUMENT RESUME

ED 357 037 TM 019 743

AUTHOR Herman, Joan L.
TITLE Accountability and Alternative Assessment: Research and Development Issues.
INSTITUTION National Center for Research on Evaluation, Standards, and Student Testing, Los Angeles, CA.
SPONS AGENCY Office of Educational Research and Improvement (ED), Washington, DC.
REPORT NO CSE-TR-348
PUB DATE Aug 92
CONTRACT R117G10027
NOTE 20p.; Abridged version of a report in "Educational Leadership," 1992.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Academic Achievement; *Accountability; Cost Effectiveness; *Educational Assessment; Educational Improvement; Educational Policy; Elementary Secondary Education; Evaluation Criteria; Knowledge Level; *Research and Development; Scoring; *Student Evaluation; Test Use; *Thinking Skills
IDENTIFIERS *Alternative Assessment; *Performance Based Evaluation

ABSTRACT

The research evidence supporting current beliefs in testing is summarized, and critical qualities that good assessment should exemplify and the current state of the research knowledge on how to produce good assessment are reviewed. Research has demonstrated the difficulties of achieving meaningful improvement in schools and the shortcomings of using existing tests to drive such improvement. However, several current policy initiatives support continuing optimism about the power of good assessment. Good assessment is built on current theories of learning and cognition and is grounded in views of the skills and capacities that students will need for future success. Several alternative assessments are being proposed, linked by common threads in that students produce or do something, using complex thinking skills in real-world contexts. These new assessments pose research and development problems to ensure their quality, and expanded quality criteria must be applied. These criteria include: (1) consequences; (2) fairness; (3) transfer and generalizability; (4) content quality; (5) content coverage; (6) meaningfulness; and (7) cost and efficiency. Many issues, including those of scoring and comparison, remain to be solved, but alternative assessments offer great potential for educational improvement. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED357037

CRESST

National Center for Research on Evaluation, Standards, and Student Testing

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☒ This document has been reproduced as received from the person or organization originating it
☐ Minor changes have been made to improve reproduction quality
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy

Accountability and Alternative Assessment: Research and Development Issues

CSE Technical Report 348

Joan L. Herman

▸ UCLA Center for the Study of Evaluation

in collaboration with:

▸ University of Colorado

▸ NORC, University of Chicago

▸ LRDC, University of Pittsburgh

▸ The RAND Corporation

TM019743

BEST COPY AVAILABLE

Accountability and Alternative Assessment: Research and Development Issues

CSE Technical Report 348

Joan L. Herman

August 1992

National Center for Research on Evaluation,
Standards, and Student Testing (CRESST)
Graduate School of Education
University of California, Los Angeles
Los Angeles, CA 90024-1522
(310) 206-1532

COPYRIGHT © 1992 THE REGENTS OF THE UNIVERSITY OF CALIFORNIA

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 in this report do not reflect the position or policies of the Office of Educational Research and Improvement or the U.S. Department of Education.

ACCOUNTABILITY AND ALTERNATIVE ASSESSMENT: RESEARCH AND DEVELOPMENT ISSUES*

Joan L. Herman

National Center for Research on Evaluation, Standards, and Student Testing (CRESST)

UCLA, Center for the Study of Evaluation

America 2000....

The national education goals....

Concerns for international competitiveness....

Renewed calls for restructuring and accountability at the state, local, school levels....

Is there any doubt that assessment continues to be a cornerstone of educational reform in the 1990s? Despite growing dissatisfaction with traditional multiple-choice tests, national and state educational policy reflects continuing belief in the power of *good* assessment to encourage school improvement.

The underlying logic is relatively simple: (a) Good assessment sets meaningful standards to which school systems, schools, teachers, and students can aspire. (b) These standards provide direction for instructional efforts and models of good practice. (c) Results from the assessment provide feedback on instructional strengths, weaknesses, and prescriptions for action at all levels of the educational system. (d) Coupled with effective incentives and/or sanctions, assessment can motivate students to learn better, teachers to teach better, and schools to be more educationally effective. Following this

* An abridged version of this report appears in *Educational Leadership*, *49*(8), 1992.

logic, assessment has the potential to be a powerful and beneficial engine of change.

Are these reasonable assumptions? How close are we to having the good assessments that are required? This article summarizes the research evidence supporting current beliefs in testing, identifies critical qualities that good assessment should exemplify, and reviews the current state of the research knowledge on how to produce such measures.

Does Assessment Support Change?

Interestingly, much of the research supporting the potential power of testing to influence instruction and schooling is based on traditional standardized tests and concludes that such tests have a negative impact on program quality. A number of researchers, using surveys of teachers, interview studies, and extended case studies, have found that accountability pressures encourage teachers and administrators to focus their 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 & Herman, 1986; Herman & Golan, 1991; Kellaghan & Madaus, 1991; Shepard, 1991; Smith & Rottenberg, 1991). Insofar as traditional standardized tests assess only part of the curriculum, many of these researchers conclude that the time focused on test content has narrowed the curriculum in two ways: (a) an overemphasis on the basic skills subjects and lower levels of cognitive skill stressed by the tests; and (b) a neglect of higher order thinking skills and content areas such as science and social studies that are not the subjects of testing. 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.

Some positive examples. Cheerier pictures emerge, however, when tests or other assessments model authentic skills. Studies of the effects of California's eighth-grade writing assessment program, for example, indicate that the program encourages teachers both to require more writing assignments of students and to give students experience in producing a wider variety of genres, effects which most would view as positive impact on instructional practice. Beyond impact on instruction, furthermore, studies of some states and districts have found improvements in student performance

over time associated with new assessment programs (Chapman, 1991; Quellmalz & 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 to use the scheme. The district witnessed an improvement in students' writing performance over the next several years, an improvement it attributed to the common, districtwide standard, the focus it provided for teachers' instructional efforts, and the district's attention to writing instruction.

This latter point is an important one for emphasis in interpreting both the district and the California state stories: Change in assessment practices was one of several factors which had the potential to influence teachers' and students' performance. The California Writing Project and a number of statewide training efforts occurring at the same time provided teachers with new, effective models of writing instruction and stressed the importance of giving students ample opportunities to write.

But pressure can corrupt. Pressure to improve tests scores, in the absence of serious, parallel supports for instructional improvement, in fact, 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 (Cannell, 1987). If the largely disadvantaged students in West Virginia were scoring above the national average, who, he wondered, might be scoring below the national average? He contacted all the states and a number of large school districts to inquire about their test performances. He found almost all reported scoring above the national norm sample, a finding which was essentially replicated by CRESST researchers using more rigorous methods (Linn, Graue, & Sanders, 1990). How can all students be performing "above average," a clear contradiction in the meaning of performance? Based on the results of an interview study, researcher Lorrie 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. She and colleagues Dan Koretz, Bob Linn and Steve Dunbar have found that such improvements in test scores do not generalize to other measures of student achievement (Koretz, Linn, Dunbar, & Shepard, 1991). In other words, superficial changes in instruction to improve test performance are not likely to

result in meaningful learning and achievement. Instead, the process results in a distortion of the meaning of test performance: Test scores no longer represent broader student achievement, but only the specific content and the specific formats included on the tests.

Mary Catherine Ellwein and Gene Glass, looking at the effects of minimum competency testing and other assessment-based reforms, illuminate other potential distortions of the ideal model when serious consequences follow from test results (Ellwein & Glass, 1987; Glass & Ellwein, 1986). They conclude that 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, and the like) to catch those who fail;" and that, furthermore,"in the end, 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 & Ellwein, 1986, p. 4). Shaped by political realities, as well as important concerns for equity and future consequences, test-based standards often become diluted and therefore have little or no influence on teachers, their instructional practices, or on students and their learning.

What Does Good Assessment Look Like?

Prior research, in short, suggests the difficulties of achieving meaningful improvement in schools and the shortcomings of using existing tests to drive such improvement. Nonetheless, a number of current policy initiatives show continuing optimism in the power of *good* assessment, finding the problem with the assessments that have been used and not with the basic strategic model of accountability.

What is good assessment? Good assessment is assessment that is built on current theories of learning and cognition and that is grounded in futurists' and others' views of what skills and capacities students (and our society) will need for future success. To many people, good assessment is also defined by what it is not: Good assessment doesn't look like the assessment that is associated with negative effects; good assessment is not standard, traditional multiple-choice items.

According to today's cognitive researchers and theorists, meaningful learning is reflective, constructive, and self-regulated (Bransford & Vye, 1989; Davis & Maher, 1990; Marzano, Brandt, & Hughes, 1988; Wittrock, 1991). People are seen not as mere recorders of factual information but as creators of their own unique knowledge structures. To *know* something is not just to have received information but to have interpreted it and related it to other knowledge one already has. In addition, we now recognize the importance of knowing not just how to perform, but also when to perform and how to adapt that performance to new situations. Thus the presence or absence of discrete bits of information, which is typically the focus of traditional multiple-choice tests, is *not* of primary importance in the assessment of meaningful learning, but rather how and whether students organize, structure, and use that information in context to solve complex problems.

Recent studies of the integration of learning and motivation also highlight the importance of affective and metacognitive skills in learning (McCombs, 1991; Weinstein & Meyer, 1991). For example, recent research suggests that poor thinkers and problem solvers differ from good ones not so much in the particular skills they possess as in their failure to use them in certain tasks. Acquisition of knowledge and skills is not sufficient to make one into a competent thinker or problem solver. People also need to acquire 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 students' cognitive abilities and dispositions also has received attention over the past several years. It has been noted that real-life problems often require that people work together as a group in problem-solving situations, in contrast to the formats of traditional tests. Further, it is postulated that groups facilitate learning in several ways: modeling effective thinking strategies, scaffolding complicated performances, providing mutual constructive feedback, and valuing the elements of critical thought (Resnick & Klopfer, 1989).

Lots of Enthusiasm for a Variety of Alternative Assessments

These new understandings of the nature and context of student learning have supported the movement away from traditional, multiple-choice tests to alternative assessments, including a wide variety of strategies such as open-

ended questions, exhibits, demonstrations, hands-on execution of experiments, computer simulations, writing in many disciplines, and portfolios of student work over time. While the terms may be diverse, several common threads link these alternative assessments:

- students perform, create, produce or do something;
- the tasks require students to use complex/multiple thinking and/or problem-solving skills;
- they often provide measures of metacognitive processes and attitudes as well as the more usual intellectual products;
- the assessment tasks themselves represent meaningful instructional activities;
- the tasks themselves often are contextualized in real-world applications (hence the term *authentic*); and
- student responses are scored according to specified criteria, known in advance, that define standards for good performance.

The enthusiasm for these new alternatives is documented by the number of states, local school districts, and other groups that are pursuing their development: A CRESST project to collect existing examples of alternative assessments has located over 171 separate examples, representing the active efforts, conservatively, of 19 state departments of education, over 30 school districts, and over a dozen other groups. The collection is catalogued as a database.

Good Assessment Represents Something Important

These new assessments do pose significant R&D problems to assure their quality. Face validity—that the assessment tasks appear interesting and appear to tap higher level thinking skills—is not sufficient. Essential to good assessment is the notion that students' performance represents something of importance, something beyond the specific task which is assessed—that is, that test performance, whatever type the measure, generalizes to a larger domain of knowledge and/or skill. For example, when an assessment asks a student to conduct a hands-on experiment to determine the optimal environment for a silk worm, we probably are not so much interested in whether the student can identify a healthy environment for silk worms; instead we probably want to use the student's performance on this specific task

as an indicator of whether he or she can use the scientific method to solve problems. We intend and expect the test to represent something more than the specific object included on the assessment.

What Are Critical Qualities for Assessment Quality?

Validity is the term the measurement community has used to characterize the quality of an assessment: at the simplest level, whether test scores accurately reflect the knowledge, skills, and/or 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. For example, does a student's performance on a standardized test of problem solving coincide with classroom observations of his/her capability, with his/her success in subsequent courses emphasizing problem solving, or with future life success in handling complex problems?

While these traditional notions of validity are still applicable, Linn, Baker, and Dunbar (1991) have noted their insufficiency for the new assessment alternatives being advanced. These researchers call for an expanded set of criteria for judging the quality of an assessment:

- **Consequences**–The history of testing has many examples of good intentions gone awry. The consequences of an assessment, as mentioned above, influence how people respond to its results and, as the Cannell findings suggest, can rebound to influence the validity of the results themselves. This overarching criterion requires that we plan from the outset to assess the actual use and consequences of an assessment. Does it have positive consequences or are there unintended effects such as narrowing of curriculum, adverse effects on disadvantaged students, etc.?

- **Fairness**–Does the assessment fairly consider the cultural background of those students taking the test? Researchers Winfield and Woodard (in press) warn that standardized performance assessments are at least as likely as current traditional measures to disadvantage students of color. She worries that, because time requirements will limit the number of tasks chosen for assessment, there is greater likelihood that the tasks selected will be those more familiar to middle-class, Caucasian students. Along with Winfield and Woodard, Linn, Baker, and Dunbar (1991) point to additional equity problems stemming from students' "opportunity to learn" that which is assessed: Have all students had equal opportunity to learn the

complex thinking and problem-solving skills that are the targets of these new assessments? In the immediate future, the answer probably is "No."

- **Transfer and Generalizability**–Mentioned above, this criterion asks whether the results of an assessment support accurate generalizations about student capability. Are the results reliable across raters, consistent in meaning across locales? Research on these issues, to which we return, raises perplexing questions about feasibility.

- **Cognitive Complexity**–We cannot tell from simply looking at an assessment whether or not it actually assesses higher level thinking skills. Schoenfeld (1991) cites a telling example: A New York teacher was given high awards for his students' performance on the Regents Exam. The exam asked students to do what were ostensibly complex geometry proofs. But it turned out that the teacher had predicted what proofs were likely to appear on the exam and had drilled his students in how to solve them. As a result, the cognitive level of students' responses is moot.

- **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 selected content needs to be consistent with the best current understanding of the field and to reflect important aspects of a discipline that will stand the test of time. That an assessment reflects and draws on critical, enduring aspects of content needs to be verified.

- **Content Coverage**–Content coverage raises issues of curriculum match and whether the assessment tasks represent a full curriculum. Because time constraints are likely to limit the number of alternative assessments which can be given, adequate content coverage represents a significant challenge. As Collins, Hawkins, and Frederiksen (1990) have recently noted, if there are gaps in coverage, teachers and students are likely to underemphasize those topics and concepts that are excluded from assessment.

- **Meaningfulness**–One of the rationales for more contextualized assessments is that these assessments will assure that students are engaged in meaningful problems, resulting in worthwhile educational experiences and in greater motivation for students' performance. However, additional evidence is needed to support this theory, as is further investigation into the relationship between alternative assessments and student motivation to do well on such assessments.

- **Cost and 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 in Assuring Such Quality?

Although the development of new alternatives is a popular idea, and many are engaged in the process, most developers of these new alternatives (with the exception of writing assessments) are at the design and prototyping stages, at some distance from having validated assessments. The aforementioned CRESST database project on alternative assessments, for example, indicates that empirical data about the quality of these assessments or about their integrity as measures of significant student learning is scant.

We've learned a lot about assuring reliable scoring. One area of both relative strength and challenge relates to issues of transfer and generalizability. On the positive side, largely from research on writing assessment, we have accumulated considerable knowledge about reliably scoring essays and other open-ended responses. According to Baker (1991), generalizations from this literature include findings that (a) raters can be trained to score open-ended responses reliably and validly; (b) validity and reliability can be maintained through use of systematic procedures [including specified scoring schemes, sound training procedures, and on-going reliability checks throughout the rating process]; and (c) rater training reduces the number of required ratings and costs of large-scale assessment (p. 3). Studies Baker reviewed from the performance assessment literature in the military further support the feasibility of large-scale performance assessments, involving tens of thousands of examinees, and the feasibility of assessing complex problem-solving and team or group performance. The alternative assessment trials currently going on in various states, districts, and schools provide similar data on these feasibility issues. For example, Vermont's experiments with portfolios, Connecticut's and California's pilots of hands-on math and science assessment, and Maryland's integrated assessment also provide evidence that it is logistically possible to administer these assessments on a large-scale, that schemes can be devised to score these assessments, and that teachers can be trained to reliably score them.

What about the meaning of these scores? The generalizability of these scores—however reliable the scoring process—remains a challenging issue. Consider, for example, the research of Shavelson and his colleagues on hands-on assessments in math and science (Shavelson, Baxter, & Pine, 1990; Shavelson, Gao, & Baxter, 1991; Shavelson, Mayberry, Li, & Webb, 1990). They

essentially pose the question "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: Their analyses found that from approximately 8 to 20 tasks were needed to obtain reliable individual level estimates. Further, Shavelson et al. (1991) found great variability across content or topic areas within a given discipline (e.g., math or science): They estimated that 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 you find. For example, Shavelson looked at how students' performance on science experiments compared with that on simulations and that on journals, all intended to measure the same aspects of problem solving. Similarly, Gearhart, Herman, Baker, and Whittaker (1992), in a study of portfolio assessment, compared how students' performance in writing was judged when based on their writing portfolios, their classroom narrative assignments, and their responses to a standard narrative prompt. The results from both studies showed substantial individual variation across the various assessment tasks. What you ask students to do and the circumstances under which they are asked to do it, in short, influence their performance and, consequently, inferences about their capabilities.

Other issues in comparing the results of similar assessments. Linn, Kiplinger, Chapman, and LeMahieu's (1991) study of the comparability of writing results across different state assessments addresses similarly thorny reliability and validity issues, and ones particularly germane to current discussions about a national system of tests to assess progress toward national standards. Under current proposals, national standards are to be articulated, and states (or clusters of states) would develop their own tests, tied to their state curricula, to assess their students' progress and status with respect to the national standards. The results of such assessments might be used for student certification, for college admissions and/or for 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

on such tests, concerns for equity demand concern for comparability of results from the different assessments. (For example, at a simple level, if test results were an important factor in hiring decisions, one would not want to unfairly disadvantage a student from one state compared to a student from another state because tests in these two states were of different difficulty.)

Linn, Kiplinger, Chapman, and LeMahieu's (1991) study used the results of statewide writing assessments to examine the comparability of results from different states. Student papers from one state were scored by trained raters from other states, using these other states' scoring schemes. In total, 10 states participated in the study. The results showed relatively high correlations between students' scores on the different scoring schemes—meaning that the student essays that were rated as best, average, and poorest tended to be the same, regardless of the specific scheme used. This high level of agreement of the relative ordering of student performance, according to Linn, is a necessary but not sufficient prerequisite 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, and in this area Linn's results found rather substantial differences in the level of scores that were assigned to the same papers by different states, meaning differences in leniency and differences in absolute standards for performance. Assuring comparability of results, in short, will require substantial work.

A Complex Challenge

The development and validation of new kinds of assessments offer great potential and significant challenge. The promise is alluring and is being effectively argued and advanced at the national, state, and local levels all over the country. Yet what we know about alternative or performance-based measures is relatively small when compared to what we have yet to discover. Building on past experiences with assessment in the service of accountability and on an expanded set of criteria for good, productive assessment, the research agenda at the National Center for Research on Evaluation, Standards, and Student Testing (CRESST) focuses on the development of new approaches to assessment, the development of appropriate psychometric theory to undergird these measures, and the exploration of the process and impact of new alternatives in educational practice. Thus, for example, the

content assessment project at CRESST (Baker, Freeman, & Clayton, 1991) has developed a prototype for assessing the depth of student understanding in specific subject areas. Starting first with students' understanding of American history, the project developed a standard assessment approach that requires students to read primary source materials (e.g., the Lincoln-Douglas debates) and then write an essay to answer a complex problem (e.g., explain the causes of the Civil War). Student essays are then rated for quality of understanding using a scoring scheme which provides holistic and analytic ratings (e.g., general impression, presence of problem focus; use of principles, use of facts).

The content assessment project is illustrative of both the exciting progress that is being made by assessment projects across the country and internationally and the problems—many unrelated to technical or measurement issues—that will need to be addressed if assessment is to meet its potential. On the plus side, the research again demonstrates the feasibility of alternative assessment. It also demonstrates that it is possible to:

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

But importantly, these same studies also indicated that student performance on these new kinds of measures is dismally low, a finding shared by most of the states and districts that have tried such assessments; and a related concern: Teachers need substantial training and follow-up support in both suitable assessment techniques and appropriate instructional strategies. Teachers are not knowledgeable about the development and use of traditional or alternative assessment; and many do not, nor do they seem to know how to, teach students to solve complex problems.

In conclusion, educational assessment currently is in a process of invention. Old models are being seriously questioned, new models are in the process of development. Substantial progress is being made to clarify and

amplify the potential of these new alternatives, but substantial challenges remain to assure that assessment supports, and does not detract from, quality education. Assessment practices themselves need to be accountable to criteria that define quality assessments. These criteria force attention not only to traditional technical issues but also, and importantly, to the consequences of an assessment and to students' opportunity to learn that which is assessed. Finally, changes in assessment, at best, 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 are able to achieve the complex skills that these new assessments strive to represent.

References

Baker, E.L. (1991, April). 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., Freeman, M., & Clayton, S. (1991). Cognitively sensitive assessment of subject matter: Understanding the marriage of psychological theory and educational policy in achievement testing. In M.C. Wittrock & E.L. Baker (Eds.), *Testing and cognition* (pp. 135-153). New York: Prentice-Hall.

Bransford, J.D., & Vye, N. (1989). A perspective on cognitive research and its implications in instruction. In L.B. Resnick & L.E. Klopfer (Eds.), *Toward the thinking curriculum: Current cognitive research* (pp. 173-205). Alexandria, VA: Association for Supervision and Curriculum Development.

Cannell, J.J. (1987). *Nationally normed elementary achievement testing of America's public schools: How all 50 states are above the national average.* Daniels, WV: Friends for Education.

Chapman, C. (1991, June). *What have we learned from writing assessment that can be applied to performance assessment?* Presentation at ECS/CDE Alternative Assessment Conference, Breckenridge, CO.

Collins, A., Hawkins, J., & Frederiksen, J. (1990, April). *Technology-based performance assessments.* Paper presented at the annual meeting of the American Educational Research Association, Boston.

Davis, R.B., & Maher, C.A. (1990). Constructivist view on the teaching of mathematics (Monograph #4). *Journal for Research in Mathematics Education.* Reston, VA: NCTM.

Dorr-Bremme, D., & Herman, J. (1986). *Assessing student achievement: A profile of classroom practices* (CSE Monograph Series in Evaluation No. 11). Los Angeles: University of California, Center for the Study of Evaluation.

Ellwein, M.C., & Glass, G. (1987, April). *Standards of competence: A multi-site case study of school reform* (Report to OERI, Dept. of Education). Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing.

Gearhart, M., Herman, J., Baker, E.L., & Whittaker, A.K. (1992). *Writing portfolios at the elementary level: A study of methods for writing assessment* (CSE Tech. Rep. No. 337). Los Angeles: University of California, Center for the Study of Evaluation.

Glass, G., & Ellwein, M.C. (1986, December). Reform by raising test standards. *Evaluation Comment*. Los Angeles: University of California, Center for the Study of Evaluation.

Herman, J., & Golan, S. (1991). *Effects of standardized testing on teachers and learning: Another look* (CSE Tech. Rep. No. 334). Los Angeles: University of California, Center for the Study of Evaluation.

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

Koretz, D., Linn, R., Dunbar, S., & Shepard, L. (1991). *The effects of high stakes testing on achievement: Preliminary findings about generalization across tests*. Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Linn, R., Baker, E., & Dunbar, S. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher, 20*(8), 15-21.

Linn, R., Graue, M., & Sanders, N. (1990, Fall). Comparing state and district test results to national norms: The validity of claims that 'Everyone Is Above Average.' *Educational Measurement: Issues and Practice, 9*(3), 5-14.

Linn, R.L., Kiplinger, V.L., Chapman, C.W., & LeMahieu, P.G. (1991). *Cross-state comparability of judgments of student writing: Results from the New Standards Project workshop* (CSE Tech. Rep. No. 335). Los Angeles: University of California, Center for the Study of Evaluation.

Marzano, R., Brandt, R., & Hughes, C.S. (1988). *Dimensions of thinking: A framework for curriculum and instruction*. Alexandria, VA: Association for Supervision and Curriculum Development.

McCombs, B.L. (1991). The definition and measurement of primary motivational processes. In M.C. Wittrock & E.L. Baker (Eds.), *Testing and cognition* (pp. 62-81). Englewood Cliffs, NJ: Prentice Hall.

Quellmalz, E., & Burry, J. (1983). *Analytic scales for assessing students' expository and narrative writing skills* (CSE Resource Paper No. 5). Los Angeles: University of California, Center for the Study of Evaluation.

Resnick, L.B., & Klopfer, L.E. (1989). Toward the thinking curriculum: An overview. In L.B. Resnick & L.E. Klopfer (Eds.), *Toward the thinking curriculum: Current cognitive research* (pp. 1-8). Alexandria, VA: Association for Supervision and Curriculum Development.

Schoenfeld, A.H. (1991). On mathematics as sense-making: An informal attack on the unfortunate divorce of formal and informal mathematics.

In J. Voss, D.N. Perkins, & J. Segal (Eds.), *Informal reasoning and education*. Hillsdale, NJ: Erlbaum.

Shavelson, R., Baxter, G.P., & Pine, J. (1990, October). *What alternative assessments look like in science*. Paper presented at Office of Educational Research and Improvement Conference "The Promise and Peril of Alternative Assessment," Washington, DC.

Shavelson, R., Gao, X., & Baxter, G. (1991, November). *Design theory and psychometrics for complex performance assessment: Transfer and generalizability* (Project 2.4 Progress Report, November 30, 1991, to OERI, Grant No. R117G10027). Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing.

Shavelson, R.J., Mayberry, Paul W., Li, Weichang, & Webb, N.M. (1990). Generalizability of Job Performance Measurements: Marine Corps Rifleman. *Military Psychology, 2*, 129-144.

Shepard, L. (1990). *Inflated test score gains: Is it old norms or teaching the test?* (CSE Tech. Rep. No. 307). Los Angeles: University of California, Center for the Study of Evaluation.

Shepard, L. (1991, November). Will national tests improve student learning? *Phi Delta Kappan*, 232-238.

Smith, M.L., & Rottenberg, C. (1991, Winter). Unintended consequences of external testing in elementary schools. *Educational Measurement: Issues and Practice, 10*(4), 7-11.

Weinstein, C., & Meyer, D. (1991). Implications of cognitive psychology for testing: Contributions from work in learning strategies. In M.C. Wittrock & E.L. Baker (Eds.), *Testing and cognition* (pp. 40-61). Englewood Cliffs, NJ: Prentice Hall.

Winfield, L., & Woodard, M.D. (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 M.C. Wittrock, & E. L. Baker (Eds.), *Testing and cognition* (pp. 5-16). Englewood Cliffs, NJ: Prentice Hall.