Large-Scale Assessment in Support of School Reform: Lessons in the Search for Alternative Measures

CSE Technical Report 446

Joan L. Herman
National Center for Research on Evaluation,
Standards, and Student Testing (CRESST)
University of California, Los Angeles

October 1997

Center for the Study of Evaluation
National Center for Research on Evaluation,
Standards, and Student Testing
Graduate School of Education & Information Studies
University of California, Los Angeles
Los Angeles, CA 90095-1522
(310) 206-1532

The work reported herein was supported in part 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 article do not reflect the policies of the Office of Educational Research and Improvement or the U.S. Department of Education.

LARGE-SCALE ASSESSMENT IN SUPPORT OF SCHOOL REFORM: LESSONS IN THE SEARCH FOR ALTERNATIVE MEASURES

Joan L. Herman

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

Assessment has long been a cornerstone of educational reform in the United States, fueled by beliefs in meritocracy, accountability, and the value of programmatic efforts to improve teaching and learning. Thirty years ago, for example, the passage of the original Elementary and Secondary Education Act of 1965 brought the federal government into local schools for the first time in support of quality education for disadvantaged students, and with it the requirement that schools receiving this funding administer standardized tests to determine eligibility and to evaluate the effects of their programs. Fifteen or so years ago, minimum competency testing enjoyed a groundswell of popularity across the country, mandated by states to assure that all students would attain minimum standards of competence. More recently, the Goals 2000 legislation (1994), advocated by the President and passed by Congress, encouraged states to set rigorous standards for student performance and to assess students' progress toward their attainment; and even more recently, a national summit of the nation's governors similarly affirmed the need for their states to establish high standards for and rigorous assessment of student accomplishment (National Governors' Association, 1996).

What's new in today's assessment, thus, is a belief not in the power or necessity of assessment per se but rather in the types of assessment that are being used and the explicit policy and practical purposes those assessments are expected to serve. Thirty years ago, assessment meant norm-referenced testing and an exclusive reliance on multiple-choice measures that ranked students, schools, and locales on general skill areas relative to one another. Fifteen years later, when minimum competency tests were added to the mix, they too were primarily multiple-choice, but tended to be criterion-referenced rather than norm-referenced. That is, these tests were designed to assess whether students had mastered specific objectives and to describe how students performed relative to expected competencies rather then in reference to the performance of others.

Most recently, there has been great enthusiasm for alternative assessments, which ask students to create their own responses rather than simply selecting them, assessments that many believe best represent the kinds of skills students will need for future success. And indeed today, of the 48 states that already conduct statewide assessments or are in the process of developing such assessment systems to support state goals for education, the great majority include both multiple-choice and alternative type assessments (Bond, Braskamp, & Roeber, 1996).

National and state educational policy, as well as local initiatives, thus reflect continuing belief in the power of assessment to support their educational goals. The underlying logic appears relatively simple: (a) Assessments can communicate meaningful standards to which school systems, schools, teachers, and students can aspire. (b) These standards provide direction for schools' instructional efforts and for students' learning. (c) Results from the assessments provide accurate feedback on performance, including insights on curriculum strengths and weaknesses for various levels of the educational system—individual student, school, state, etc. (d) Educators and students will use this feedback to improve their teaching and learning practices. (e) Coupled with appropriate incentives and/or sanctions, assessment will motivate students to learn better, teachers to teach better, and schools to be more educationally effective. Following this logic, assessment can provide valuable focus to the system and has the potential to be a powerful and beneficial engine of change.

Such use of large-scale assessment, of course, is neither unique to the United States nor of recent origin: Civil service examinations in China enjoy a 3000-year history (DuBois, 1966); and the British eleven-plus examination, which is used to determine students' transfer from primary to secondary education, is among the most visible high-stakes tests in the world (Egan & Bunting, 1991). What is unusual in the United States situation is the use of large-scale assessment to promote system accountability and improvement (Feuer & Fulton, 1994) for educational systems that historically have been committed to local control. Also unique is an apparent general public dissatisfaction with student outcomes and public schooling that has fueled attention to accountability and demands for change (Keeves, 1994).

This article provides a historical perspective on current interest in alternative assessment in the United States and identifies critical qualities that

good assessment should exemplify. The paper then reviews research results regarding the technical quality and consequences of using this form of assessment for large-scale accountability purposes and concludes with implications for future practice.

Alternative Assessment: A Rose by Many Names

A bit of context may be helpful in considering the findings below. Many terms have been advanced when discussing alternatives to conventional, multiple-choice testing. These include alternative assessment, authentic assessment, and performance assessment and, in fact, run the gamut from portfolios of student work or extended student projects that may consume an entire school year or years to open-ended questions that resemble multiple-choice test items where the response options have been omitted. In this article, the terms *alternative, authentic* and *performance assessment* are used more or less synonymously to mean variants of performance assessments that require students to generate rather than choose a response. Alternative assessment by any name requires students to actively accomplish complex and significant tasks, while bringing to bear knowledge, recent learning, and relevant skills to solve realistic or authentic problems.

In the context of large-scale assessment in the United States and internationally, alternative assessment has typically meant having students solve problems and/or compose a response over a span ranging from 20 minutes to a few classroom periods. Some examples: having students design a bookcase within given function, space and cost constraints and explain how their design meets the given parameters (California State Department of Education, 1992); asking students to study the motion of falling maple seeds, design experiments to explain their spinning flight patterns, and interpret results in terms of scientific concepts such as laws of motion, aerodynamics, and air resistance (Lomask, Baron, Greig, & Harrison, 1992); planning and carrying out appropriate tests to determine the starch content of three unknown substances (Tamir & Doran, in press); writing a letter in French comparing the benefits of living in the country and living in a big city (Carroll, 1975); asking students to read historical documents and then use the perspective in these documents with their prior knowledge to explain a major historical issue to a peer (Baker, Freeman, & Clayton, 1991); based on a semester of work, asking students to make a presentation on their proposal for disposing of nuclear waste, based on their

knowledge of science and taking into consideration social, political, and environmental issues (Herman, Osmundson, & Pascal, 1996).

Portfolios also have experienced popularity as a form of alternative assessment in a number of statewide and local district systems. Typically in these systems, students are asked to select examples of their work from that conducted over the course of a semester or of the school year, and the body of work so assembled is then evaluated based on a standard scoring rubric. The mathematics portfolios required in the state of Kentucky, for example, require students to select five to seven pieces of their work that show the breadth and depth of their understanding of mathematics concepts and principles. They are scored against a rubric that defines performance characteristics for ratings of "novice," "apprentice," "proficient" or "distinguished."

Recent Historical Perspective: Does Assessment Support Change?

Interestingly, much of the rationale underlying the United States' use of alternative assessment to influence instruction and school learning is based on research showing adverse reactions to traditional standardized tests and evidence that such tests have had a negative impact on the quality of curriculum and classroom learning. A number of researchers, using surveys of teachers, interview studies, and extended case studies, have found that mandated, public testing does indeed encourage teachers and administrators to focus their planning and instructional effort on test content, to mimic the tests' multiple-choice formats in their classroom curriculum, and to devote more and more time to preparing students to do well on the tests (Corbett & Wilson, 1991; Dorr-Bremme & Herman, 1986; Herman & Golan, 1991; Kellaghan & Madaus, 1991; Shepard, 1991; Smith & Rottenberg, 1991). These researchers found that not only did teachers tend to emphasize the content and format of the tests throughout the year, but any number actually stopped their regular instruction weeks before the test in order to intensively drill their students on specific test preparation activities.

Insofar as traditional standardized tests can at best assess only part of the curriculum, many of these researchers concluded that the focus on test content had narrowed the curriculum in a number of ways, by encouraging teachers to (a) overemphasize the basic skills subjects and lower levels of cognitive skill stressed by the tests; (b) neglect complex thinking and problem-solving skills that are not

well assessed with multiple-choice formats; and (c) give short shrift to content areas such as science, social studies and the arts, which often were not the subjects of testing (Darling-Hammond & Wise, 1985; Shepard, 1991). Herman and Golan (1993), among others, noted that such narrowing was likely to be greatest in schools serving at-risk and disadvantaged students, because test scores in these schools were typically very low, and educators in these schools were likely to be under great pressure to improve their scores.

The problem was not one of just what was tested and taught, but how it was tested and taught, as well. Noting that worksheets posing simple, decontextualized questions about discrete pieces of knowledge appeared rampant in American schools, some observers traced the practice to the models provided by traditional standardized tests. Theorists noted that multiple-choice tests represented old, behaviorist views of learning, where knowledge was thought to be accumulated in easily digestible chunks, where students were viewed as black boxes to be filled with knowledge, where learning was thought to be a progression of discrete skills, and where meaningful context and links to students' experience were not accorded importance (Resnick & Resnick, 1992; Shepard, 1991). Thus in preparing students for traditional standardized tests and using instructional processes consistent with the test, teachers were clinging to outmoded behavioral approaches to instruction.

Some Positive Examples

Effects on instruction, however, appeared very different when tests or other assessments modeled authentic skills. Direct writing assessment—asking students to actually compose an essay rather than answer multiple-choice questions about the quality or grammar of a given piece—was a first example. Large-scale writing assessment had begun to gain popularity in the 1970s, starting with the National Assessment of Educational Progress; then gradually throughout the 1980s more and more states and locales moved to include this type of assessment in their programs. At the time, arguments for this mode of testing were based primarily on evidence of validity—evidence suggesting that multiple-choice tests did not provide accurate measures of students' ability to write (Quellmalz & Burry, 1983). But as experience with these direct measures grew, their potential for influencing teaching and learning became more apparent. Studies of the effects of California's eighth-grade writing assessment program, for example, indicated that the program encouraged teachers both to require more

writing assignments of students and to give students experience in producing a wider variety of genres, effects that most would view as positive impact on instructional practice.

Beyond impact on instruction, furthermore, studies showed that student performance in some states and districts improved over time with the institution of the 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 deserves underscoring and is very important in interpreting both the district and the California state stories: Change in assessment practices was one of several important factors that were likely responsible for changes in teachers' practices and in students' performance. The California Writing Project and a number of statewide training efforts occurring at the same time were dedicated to teachers' capacity building and made substantial investments to provide professional development. These ongoing and serious teacher capacity-building efforts acquainted teachers with and helped them to implement new models of writing instruction that advocated involving students in an extended writing process and giving them ample opportunities to write.

Is There Meaningful Improvement?

In fact, in the absence of serious teacher capacity building to support instructional improvement, pressure to improve test scores may well corrupt both the teaching and learning process and the meaning of the test scores. In 1987, John Cannell, at that time a pediatrician in one of the poorest states in the country (West Virginia), was surprised to read that the students in his state had performed above the national average on the statewide assessment (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? Dr. Cannell contacted all the states and a number of large school districts to inquire about their performance on norm-referenced achievement

tests. He found that almost all reported scoring above the national norm sample, a finding that was essentially replicated by CRESST researchers using more rigorous methods (Linn, Graue, & Sanders, 1990). How could all students be performing "above average," when the nature of the metric is that half should be above average and half below? Clearly the meaning and credibility of the test scores were in doubt.

After conducting an interview study to delve into possible reasons for these findings, researcher Lorrie Shepard concluded that the answer lay largely in the teaching phenomenon mentioned above: Teachers were directly teaching to the test (Shepard, 1990). They often provided daily skill instruction in the content and formats that closely resembled the tests and, in the worst cases, had students practice actual test items. Further, she and colleagues Daniel Koretz, Robert Linn and Stephen Dunbar found that observed improvements in test scores did not generalize to other measures of student achievement, raising a significant challenge to the validity of the standardized test results (Koretz, Linn, Dunbar, & Shepard, 1991). In other words, superficial changes in instruction to improve test performance apparently did not result in meaningful learning and achievement as might be evidenced by consistent results over various measures of student achievement. Instead, the process appeared to distort of the meaning of test performance. In such situations 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 documented other distortions that can occur in assessment-based reform models when serious consequences follow from test results (Ellwein & Glass, 1987; Glass & Ellwein, 1986). These researchers examined the effects of minimum competency testing, such as for high school graduation, and other assessment-based reforms, such as raising standards for admission, or remedial course placements at college entry. Their study concluded that when policy makers 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.

Alternative Assessment as a Key to Reform

History and prior research, in short, suggest both the potential and difficulties of using assessment as a tool to support meaningful improvement in schools. History also shows the shortcomings of using traditional multiple-choice tests to drive such improvement. With these findings as background, current policy initiatives show continuing optimism in the power of assessment to support rigorous goals for academic achievement. Continuing to cleave to a basic strategic model of accountability to force change, these initiatives have identified the problems of history as residing in the test instruments themselves—the exclusive reliance on multiple-choice testing and an often startling mismatch between the content of traditional standardized tests and many current goals for student performance. For example, it was estimated that only 26% of the Arizona Essential Skills were covered by then-mandated standardized tests being used statewide (Haladyna as reported by Smith, 1997). The Skills represented high standards, complex thinking, and integrated problem solving, mirroring the national standards being advocated by content specialists across the country.

For current assessment policy, the solution lies in alternative forms of assessment that better represent the rigorous standards being advocated nationally and the advanced knowledge and skills that students will need to be successful and productive citizens—abilities to access and use information, to solve problems, to communicate. These alternative forms of assessment are also intended to provide good models for instruction that will support meaningful learning, consistent with recent cognitive theory.

The Link to Learning and Cognition

According to today's cognitive researchers and theorists, meaningful learning is reflective, constructive, and self-regulated (Bransford & Vye, 1989; Davis, Maher, & Noddings, 1990; Glaser & Silver, 1994; 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 typically has been the focus of traditional multiple-choice tests, is not of primary importance in the assessment of meaningful learning. Instead, what is highly valued is 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 (Borkowski & Muthukrishna, 1992; Garcia & Pintrich, 1994; 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, the knowledge of when to apply them, and the ability to learn from their experiences. These too have been incorporated into alternative assessments that almost always require that students plan, organize, and execute complex tasks; attention to these dispositions and metacognitive skills also is evident in portfolio assessments that ask that students to reflect upon their work.

The role of the social context of learning in shaping students' cognitive abilities and dispositions also has received attention over the past several years, and it too has been incorporated into some alternative assessments. Groups are thought to facilitate learning in several ways, for example, by modeling effective thinking strategies, scaffolding complicated performances, providing mutual constructive feedback, and valuing the elements of critical thought (Resnick & Klopfer, 1989; Slavin, 1990). That real-life problems often require teams to work together in problem-solving situations has provided additional rationale for including group work in alternative assessments.

Modeling Good Instruction

Alternative assessments often explicitly model what is thought important in good instruction, adding a significant new twist to the role and function of large-scale assessment in school reform. Not only is accountability assumed to

motivate systems and the individuals within them to change and improve their performance, but the assessments themselves are intended to communicate how to change. Negative findings of the past regarding teachers' test preparation practices—that teachers have students directly practice test-like activities—have been turned to the positive: The assessment provides tasks that are instructionally valuable and promote learning, and if teachers mimic such tasks in their practice, they will likely improve their instruction.

Making Alternative Assessment Good Measurement

These new assessments, however, do pose significant R&D problems to assure their validity as measures of student performance. Face validity—that an assessment appears to tap the complex thinking and problem-solving skills that are intended to be assessed—is not sufficient to assure accurate measurement. As one example, Schoenfeld (1991) shares the story of a New York teacher who was given high awards for his students' advanced performance on the Regents Exam. The exam asked students to complete a series of what were ostensibly complex geometry proofs requiring complex thinking and problem solving. But it turned out that the teacher had determined which proofs were likely to appear on the exam and had drilled his students in how to solve them. As a result, despite what the assessment "looked" like, it was not possible to draw inferences from it about these students' understanding of geometry and their ability to apply complex geometric concepts and principles, because for them, the exam was an exercise in rote learning.

Validity. Basic to good student assessment is the notion that results represent important knowledge and/or capabilities, broader than the specific task(s) that happen to be chosen for assessment. In a good assessment—regardless of the type of measure, test performance generalizes to a larger domain of knowledge and/or skills and thus enables us to make accurate inferences about students' capabilities and accomplishments. For example, when an assessment asks students to conduct an experiment to determine optimal environmental conditions for sow bugs to thrive, we probably are interested not so much in whether the students can create good conditions for sow bugs as in whether students can apply their knowledge of biology and the scientific method to solve problems. We typically want to use the students' performance on this specific task as an indicator of broader scientific knowledge and problem solving ability. We intend and expect the test performance to represent something more than the

specific object included on the assessment and the specific time and testing occasion on which the assessment was administered.

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 the test is intended to measure. For traditional multiple-choice measures, concerns for validity have focused on issues of reliability and patterns of relationships that suggest whether the assessment is tapping the intended construct and whether it provides accurate information for specified decision purposes. For example, does a student's performance on a standardized test of problem solving coincide with classroom observations of his/her capability and with his/her success in subsequent courses emphasizing problem solving? Does using test results as part of the evidence for placement produce accurate decisions? Any test in and of itself is neither valid nor invalid; rather, current theory requires that we accumulate evidence of the accuracy of that assessment for particular purposes.

Expanded criteria for judging validity. While these technical concerns for validity are still critical, Linn, Baker, and Dunbar (1991) have called 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 meaning 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? Does it provide a level playing field that enables all children to show what they know and can do? History suggests a number of areas in which fairness must be assured.

- 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 shall return, raises perplexing questions about feasibility.

- Cognitive Complexity—We cannot tell from simply looking at an assessment whether or not it actually elicits higher level thinking skills. Instead, we need evidence that an assessment actually measures the complex thinking and problem solving it is intended to measure.

- 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 that 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 under-emphasize 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.

Issues and Status in Establishing Technical Quality

We explore below evidence about the current state of the art in these areas, first reviewing evidence about the technical quality issues related to reliability, generalizability, fairness, and content quality, followed by recent investigations of the consequences, including the implementation, impact, and costs of these forms of assessment.

Reliability: The Essential Foundation

Reliability is the necessary but not sufficient prerequisite to the technical quality of any measure. In order to provide meaningful information for any

purpose, we need to know that the measurement instrument provides consistent results—that the score reflects something meaningful and is not subject unduly to fluctuation from irrelevant sources. For example, imagine that a student takes a science problem-solving assessment today and a parallel assessment tomorrow or next week, and the student has not studied, been taught science or had any related experiences in the interim. One expects the student's score to be similar on both occasions because the underlying science capability has not changed. If, however, the scores are quite different from one occasion to the next, then the measurement does not have a consistent meaning on both occasions and thus does not provide trustworthy information. Just as we expect the scale to show the same weight this Monday and next Monday, if we've not gained or lost weight in the interim, and just as we expect our height to be same regardless of whose yardstick is used, we want our measures of student achievement to be reliable and consistent.

We've learned a lot about how to assure reliable scoring. Because student-constructed responses are a defining feature of alternative assessment, scoring requires humans, not machines, to read and judge the responses. Reliability of scoring thus is the base issue in reliability—one that hardly occurs in the automated world of multiple-choice testing, and yet is the foundation upon which all other decisions about technical quality rest. Raters judging student performance should be in basic agreement as to what scores should be assigned to students' work, within some tolerable limits (which measurement experts report as "measurement error"). Do raters agree on how an assessment ought to be scored? Do they assign the same or very similar scores to a particular student's response? If the answers to these questions are not affirmative, then student scores are a measure of who does the scoring rather than the quality of the work. The score cannot well represent student capability because it reflects the scorers' idiosyncrasies as much as the skills and abilities of the individuals being assessed.

While not without challenge, assuring the reliability of scoring is an area of relative technical strength in performance assessment. Largely from research on writing assessment, we have accumulated considerable knowledge about how to reliably score essays and other open-ended responses. According to Baker (1991), the literature shows that (a) raters can be trained to score open-ended responses consistently, particularly if well-documented scoring rubrics and benchmark or anchor papers exemplifying various score points are used, and scorers are given

ample opportunities to discuss and apply the rubric to student response samples requiring increasingly complex discriminations; (b) applying systematic procedures throughout the scoring process can help to assure consistency of scoring—for example, procedures such as having raters qualify for scoring by demonstrating their consistency, conducting periodic reliability checks throughout the scoring period, and retraining scorers as necessary; 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. International studies similarly show that it is feasible—although not without challenge—to use alternative assessments in cross-national studies (Wolf, 1994), as do the experiences of any number of countries in using alternative assessments for student selection decisions (Feuer & Fulton, 1994; Kellaghan & Madaus, 1991; Madaus, 1988).

But reliable scoring can be difficult to achieve. The alternative assessment trials that have been undertaken in various states, districts, and schools over the last several years provide similar data on some of these feasibility issues, but also document the challenges of reaching agreement in scoring in new areas of performance. On one extreme, the Iowa Tests of Basic Skills direct writing assessment demonstrates that it is possible to achieve very high levels of agreement—better than 90% exact agreement on student scores—with highly experienced, "professional" raters and tightly controlled scoring conditions, and scoring criteria that have an established history (Hoover & Bray 1995). On the other end of the spectrum, in Vermont's early experiments with portfolios and Arizona's recent integrated assessment program, there was insufficient reliability to permit the public release and intended use of the assessment system results (Koretz, McCaffrey, Klein, Bell, & Stecher, 1993; Smith, 1997). In the Vermont case, which used as scorers teachers from throughout the state, the percentage of exact rater agreement in the first year of statewide implementation was essentially that which would be expected by chance alone (Koretz, McCaffrey, et al., 1993). Correlations between the scores given by different raters to the same pieces of student work, a second way of looking at rater reliability, were similarly discouraging: They ranged from .28 to .60, depending on whether individual

dimension scores or aggregate or overall summary scores were the subject of analysis.

The New Standards Project, a national project that is developing high-stakes assessments in concert with professional development and technical assistance activities in an attempt to raise standards of student performance, experienced similar difficulties in its early trials with math and language arts assessment. Here, the reliabilities generally ranged from .60 to .75, based on scorers who were highly experienced as teachers, with limited prior experience in scoring (Resnick, Resnick, & DeStefano, 1994).

The Vermont case is telling not only in demonstrating the difficulties of achieving reliability of scoring in the early years of a new assessment, but also in pointing out the potential conflict in various purposes of large-scale, alternative assessment. Vermont's statewide portfolio assessment was intended to serve both accountability purposes and teacher capacity building and instructional improvement purposes. In order to serve the first purpose, reliability of scoring is a high priority, because without such consistency, scores cannot be reported at the school and district or regional levels. But in order to serve the second purpose, Vermont wanted to involve as many teachers as possible across the state in the scoring process. The two purposes thus pulled in two different directions: tightly controlled scoring with highly experienced raters to serve accountability uses versus a highly inclusive process involving as many teachers as possible in the scoring to support new instructional practices.

The imperative of consensus. Available data, however, do suggest that problems in achieving reliability in scoring decrease in time as states and locales work out the bugs in their rubrics, fine-tune their procedures, and develop a core of knowledgeable, experienced raters who, in turn, can support the development of consensus on standards for scoring with a widening pool of educator/scorers. The Pittsburgh portfolio assessment experience provides one such example, and one which demonstrates the power of consensus derived from common understandings of curriculum and instruction priorities (LeMahieu, Gitomer, & Eresh, 1994). Achieving consistency in scoring of portfolios has been particularly challenging, in large part because although portfolios are assembled according to the same general specifications, the specific examples of work contained in an individual portfolio can vary from student to student and from class to class. Thus in contrast to direct writing assessments where all student papers are written in

response to the same prompt or assignment, where that assignment and likely responses to it are well understood by scorers, and where all students are responding under the same conditions, portfolio scorers must grapple with a wide range of nonstandard assignments that vary in difficulty and that were produced under varying conditions. As a result, the scoring judgment is a complex interaction between the nature, appropriateness, and difficulty of the student's assignment and the quality of the student response itself.

In Pittsburgh's 1992 portfolio assessment, students across the district in Grades 6-12 created portfolios by selecting four writing samples from those they completed over the course of a year. The four pieces were selected according to a set of guidelines to include an important piece, a satisfying piece, an unsatisfying piece, and a free pick. The scoring rubric emerged from a decade of discussions of student writing, conducted first in the context of developing new approaches to curriculum and instruction and a three-year dialogue on rubric development (Camp, 1990, 1993; Gitomer, 1993). The discussions produced a rich evaluative framework that was commonly understood among participants, and most of the scorers for the districtwide portfolio assessment had been participants in this process. The rubric featured three dimensions: accomplishment in writing; use of processes and strategies for writing; and growth, development, and engagement as a writer. Despite the great variety of student-selected work, scoring reliabilities ranged from .84 to .87 at the middle school level and from .74 to .80 for high school students; and at the middle school level, raters agreed within one score point over 95% of the time (agreements at the high school level ranged from 87%-91%), far above what would be expected by chance. The authors credit the effort's success to the strength of the shared interpretative framework, carefully nurtured and developed over time in the course of continuing, critical conversations. At the same time, researchers studying the Advanced Placement Studio Art assessment (Myford & Mislevy, 1994) noted the difficulty of establishing such a shared interpretative framework.

What Do the Scores Mean?

While history shows that reliability of scoring is a tractable problem, the literature makes clear the difficulties of generalizing from performance on specific measures to inferences about student capabilities in larger domains. The consistency of students' performance on alternative assessments designed to measure the same underlying capability is a significant problem. For example,

student performance appears very sensitive to changes in assessment format, meaning that the context in which you ask students to perform—as much as student capability, which is the object of assessment—influences the results you find. In one such demonstration, Shavelson and colleagues (Shavelson, Baxter, & Pine, 1991) analyzed how students' performance on science experiments compared with their performance on computer simulations of the same experiments and with an analysis of their journal entries from their laboratory work, all intended to measure the same aspects of science problem solving. Similarly, Gearhart, Herman, Baker, and Whittaker (1993), 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, with all three assessments intended to measure students' writing capability. The results from both studies showed substantial individual variation across the various assessment tasks. Two thirds of the students who were classified as capable on the basis of portfolios, for example, were not so classified on the basis of the direct writing prompt (Herman, Gearhart, & Baker, 1993). By the same token, doing well based on observations of laboratory work did not predict good performance on the simulation. 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.

Predictably, however, when tasks are tightly defined and the questions strictly parallel except for format, results are more consistent. Brenda Sugrue and colleagues used a carefully crafted task to investigate the components of science problem solving in various formats. Students' performance was similar regardless of whether the assessment task was a hands-on problem or a paper-and-pencil facsimile of the problem—that is, regardless of whether students worked with actual batteries, wires, and light bulbs to create a circuit, or simply were presented with visual representations of these objects, their written explanations were similar. However, performance on tasks eliciting different types of performance—for example, selecting the right combination of materials to achieve the brightest light versus explaining why that combination achieved such a result—produced different results (Sugrue, 1996).

How many tasks are needed to get a reliable estimate? That a given student's performance will vary depending on which specific tasks are included on the assessment means that multiple tasks are needed to achieve a stable

measure of that student's capability. A number of researchers have used generalizability theory to examine this general issue in a number of content areas, including writing, mathematics, and science assessment (Dunbar, Koretz, & Hoover, 1991; Linn, Burton, DeStephano, & Hanson, 1995; Moerkerke, 1996; Shavelson et al., 1991; Shavelson, Baxter & Gao, 1993; Shavelson, Mayberry, Li, & Webb, 1990). These researchers have analyzed sources of measurement error in student scores, looking at the variability attributable to raters, tasks, and the interaction of raters, task and students, to estimate the combination of numbers of raters and tasks that are needed to produce reliable results. The research has consistently shown that although raters are an important source of measurement error, variability due to task sampling (i.e., the particular tasks included on the test) is a far greater problem. Individual student performance varies over and interacts with tasks—that a student does well on one science problem-solving task, for example, does not mean she or he will do well on a second problem-solving task, and the students who perform well on one task may well not be the same students who perform the best on the second task. This variability makes the number of tasks included in an assessment a critical determinant of its reliability.

So how many tasks does one need to get a minimally stable estimate of a student's capability in a given content area? Although the results vary somewhat depending on the specific study, the range is telling: When items are not tightly specified, analyses show that from 8 to 20 tasks are needed to obtain reliable individual estimates, with most studies in the 15 to 17-task range (Dunbar et al., 1991; Linn et al., 1995; Shavelson et al., 1990, 1991, 1993). And these numbers of tasks only achieve about a .8 level of reliability, a level considered minimum by some but one that risks significant reclassification errors (Rogosa, 1994).

Furthermore, Shavelson et al. (1993) remind us of the problem of great variability in performance across topic areas within a given discipline (e.g., numbers, operations, measurement within mathematics; chemistry biology, physics within science). The researchers estimated that at least 10 different topic areas may be needed to provide dependable measures of a student's performance in one subject area.

However, it does appear that careful specification procedures, which set parameters for the nature of the tasks, their content, and format, and which use common scoring schemes, can substantially reduce the number of items needed to achieve minimum reliability. Based on a standard specification to generate

explanation tasks, Baker's results indicate that only three to five items would be needed to produce stable estimates of an individual's understanding of history (Baker, 1994). Even so, with each item taking at least two class periods, the time demands are substantial. Similarly, Moerkerke (1996) found that the use of specifications enabled developers to produce parallel assessments that had similar difficulty levels.

Assuring accuracy of decisions relative to a standard. Generalizability of measurement, as reviewed above, is one approach to examining the dependability of a measure. But how might the case be when standards-based assessments are used to make decisions about whether students have achieved a particular standard, and when achieving the standard has important consequences for students, such as determining school graduation or university entry? Here, the technical accuracy of the decision becomes a critical issue—that is, if we are certifying someone as capable on the basis of assessment results, have we made an accurate decision about that individual's capability? Recognizing that any score is only an estimate of a student's actual capability, Bob Linn and colleagues have used the concept of standard error of measurement to examine this question. The researchers ask *How many items would it take to be 95% confident that a decision is correct, that is, that a student who scores a 3, if that is the cut point for certification, actually has a true score of 3?* With student performance classified on a 4-point scale, the technical question revolves around the number of items needed to achieve a standard error of measurement (SEM) of .25 or less, and the assumption that a student's true score likely (95% probability) falls within two SEM of the observed score. Based on the New Standards Project 1993 mathematics trial data,[1] 9 to 25 items would be required to achieve the .95 confidence level (Linn et al., 1995)

Analyzing results over time. Whether the analysis issue be based on generalizability or SEM, the underlying issue is similar: There is great variability in student performance by task. This not only raises important issues for interpreting individual scores, but moreover raises particularly thorny challenges when examining student progress or comparing results from year to year. While we know that using different tasks will influence performance, concerns for memorability of tasks, test security, and practice effects make it inappropriate to

[1] New Standards is attempting to create an assessment system to certify whether students have achieved rigorous academic standards; project leaders intend the results to be used for graduation, college admission and job entry, among other things.

use the same tasks to make such comparisons. Solutions to these vexing comparability, equating, and progress assessment issues are under study (Bryk & Raudenbush, 1992; Muthén et al., 1995; Rogosa, 1995; Rogosa & Sanger, 1995; Seltzer, Frank, & Bryk, 1994). Adding their views from a nontechnical perspective, educators in Maryland and Kentucky questioned the validity of test score gains reported in both those states (Koretz, Barron, Mitchell, & Stecher, 1996; Koretz, Mitchell, Barron, & Keith, 1996)—that is, teachers questioned whether the gains mirrored real gains in learning as opposed to superficial score gains from practice and greater familiarity with the test formats.

Cognitive validity of results. Analyses of generalizability and of standard errors of measure speak to the dependability and consistency of a score but do not directly address whether the results have the meaning as intended. Returning to the scale example raised earlier, a scale may consistently weigh us—for instance, showing the same weight when we step on and off it, over and over again—but that does not necessarily mean that the weight it shows is accurate. Our scale could be off by a kilo or two, for example, and each time show us to weigh two kilos more than our true weight. Similarly with measures of student capability: Results are not necessarily what they seem to be. Just because a test is intended to measure mathematics problem solving does not mean its results can be so interpreted. And in fact, a number of observers have noted a paucity of significant disciplinary content in many performance assessments (Herman, 1996; Wolf, 1992).

Robert Glaser and Gail Baxter have developed a framework and an explicit methodology for examining the match between an assessment's intentions and the nature of cognition that is actually assessed (Baxter, Elder, & Glaser; 1996; Baxter & Glaser, 1996). Guided by the expert-novice literature and current understandings of the relationship between competence and quality of cognitive activity (Glaser, 1991), their framework highlights four types of cognitive activity that differentiate different levels of competence: problem representation, solution strategies, self-monitoring, and explanation. Applying the framework to a small sample of science problem-solving assessment tasks from state and local programs, and using protocol analysis, observation, and analysis of work, the researchers then analyzed the nature and quality of cognitive activity actually elicited by the tasks compared to the objectives of the test developers. In essence, they asked: If an assessment task claims to measure complex science problem

solving, is there evidence that scores well represent the level of students' competence in this specific instance? They identified three prevalent situations. In the first group were tasks that elicited appropriate cognitive activity, and the nature and quality of observed activity correlated well with student performance scores on the task. An assessment purported to measure scientific problem solving; the task indeed presented students with a new problem and required that students identify and pursue solution strategies, self-monitor their progress, and explain their rationales; and the scoring appropriately reflected the level of competence observed. In the second group were tasks that elicited appropriate cognitive activity, but the scores did not match the level of observed activity, because the scoring system either was not aligned with task demands or was not sensitive to the differential quality of cognition in students' performances. For example, a task required students to sort and categorize a collection of animals, but the scoring scheme did not take into account the quality of scientific explanation evidenced, and thus students who used scientifically arbitrary classification schemes received scores as high as those of students who used scientifically-based schemes. In the third group were tasks where students could bypass the intended cognitive aspects—that is, tasks ostensibly measuring problem solving that students could answer without engaging in any of the intended activity. The New York example mentioned earlier belongs to this category.

It is noteworthy that two of the three represent situations where results are invalid—that is, score results do not support inferences about students' problem-solving capability. That all of the assessments analyzed by Baxter and Glaser were part of prominent pilot programs that were being administered in large numbers statewide or districtwide in fact underscores the need for attention to content quality and cognitive validity. The framework developed by these researchers clearly has implications for assessment development as well as for revision and validation of results.

Issues and Status of Fairness in Alternative Assessment

An essential ingredient in any assessment, fairness requires attention to a variety of measurement and use issues. Historically, concerns for equity and fairness have centered on assuring objectivity and avoiding bias. We want to be sure that scores are a function of students' capability and not a function of who

the students are, their gender, ethnic or cultural background, or other personal or social characteristics that are irrelevant to the capabilities being assessed. Safeguarding that no students get special advantage, in fact, is an important rationale underlying the development the standardized administration and scoring of multiple-choice tests (National Academy of Education, Committee on Educational Research, 1969). Considerable attention also has been given to sensitivity reviews to guard against any potentially offensive or culturally unfair test content and to statistical techniques, such as differential item functioning, to detect possibly biased items (see, e.g., Camilli & Shepard, 1994; Holland & Wainer, 1993).

Objectivity and Bias

Attention to reliability of scoring, described earlier, represents similar concerns for objectivity and avoiding bias in scoring—assuring that the score reflects the performance and not who does the scoring or other features or characteristics that are relevant to what is being assessed. Because alternative forms of assessment generally require human judgment for scoring, additional sources of bias may creep into the scoring process, and there need to be safeguards against these. For example, the Pittsburgh portfolio assessment project previously mentioned (LeMahieu et al., 1994) explicitly examined the effects of gender and race of both scorers and students on scoring. With regard to gender, the researchers found that females' performance was scored higher than males, and that female scorers tended to give higher scores than their male counterparts, but importantly there was no interaction between sex of rater and sex of student—for example, female scorers did not score female students differentially higher than male students or vice versa. Similarly, with regard to race, while White students received higher ratings than African American students and African American raters gave lower scores than White raters, the race of the rater did not interact with the race of the student.

Differences in performance by race or gender (depending on subject area) of course are not unique to performance assessment: Results from traditional standardized tests historically have shown substantial gaps between the performance of Whites and that of economically disadvantaged minority students and, for some subjects, between boys and girls. Bolger and Kellaghan (1990), for example, found that boys outperformed girls on multiple-choice tests of

mathematics, Irish, and English achievement but, interestingly, found no such differences in short-answer assessments of each subject area.

On the issues of bias, the analytic question is whether observed differences reflect actual difference in competence or some bias in the assessment situation that unduly advantages some groups over others. Does the assessment put minority students at a disadvantage relative to their majority counterparts because of cultural content that is not essential to the skills and understandings that the assessment is intended to measure? Similarly, does some non-essential aspect of the task give unfair advantage to boys over girls or vice versa? Researcher Linda Winfield (1995) warns 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, it is likely that the tasks selected will be those more familiar to middle-class, Caucasian students.

In the minds of teachers, based on data from statewide assessment programs in Arizona, Kentucky, and Maryland, performance assessments do unfairly disadvantage some students (Koretz, Barron, et al., 1996; Koretz, Mitchell, et al., 1996; Smith, 1997). Of particular concern are the language demands of many alternative assessments, which often ask students to engage in significant reading and/or writing even though the object of measurement is not language arts skills. For example, for a math problem-solving assessment that requires reading ability to comprehend the problem, do the reading demands detract from some students' ability to show their mathematics capability? Virtually all teachers reported that the emphasis on writing in Maryland's statewide assessment program made it difficult to judge the mathematical competence of some students (Koretz, Mitchell, et al., 1996). At particular risk are students with limited English proficiency (August, Hakuta, & Pompa, 1994).

Bias in Opportunity to Learn

A number of observers also have highlighted fairness issues stemming from students' "opportunity to learn" that which is assessed (Darling-Hammond, 1995; Herman & Klein, 1996; Linn et al., 1991). Their concerns are particularly acute in high-stakes assessments where results carry serious consequences for students and schools. It is unfair, for example, to hold students accountable for achieving standards for which they have had little or no instructional exposure. Similarly it

is unfair to use assessment results to compare schools or students—for example, to determine which schools are the best, which students are to be admitted to college, gain job entry, etc.—when the assessments are well aligned with the curriculum for some students but quite inconsistent with the instruction that is provided to other students.

The equity issues are compounded because there is good reason to believe that economically disadvantaged, minority students are likely to have less access to the kinds of thinking curriculum that would prepare students to do well on performance assessments. A number of researchers have noted that these are the students who were most likely to have been subjected to a "drill and kill," basic skills curriculum, driven by strong accountability pressure in their schools to improve scores on traditional standardized tests (Darling-Hammond, 1995; Herman & Golan 1993; Madaus, 1991; Shepard, 1991). There also is evidence that children in economically disadvantaged communities are less likely than their more advantaged suburban peers to have available some of the resources that are essential to instructional opportunity—for example, teachers possessing appropriate subject matter background and instructional materials in line with new curricular thinking (Herman & Klein, 1996). Smith (1994) similarly observes that districts serving poor children are less likely to have the professional development resources to support teachers' capacity to support the kinds of learning valued by new forms of assessment.

Comparability and Equity in Portfolio Assessment.

Portfolio assessment in particular makes apparent these problems of equity in opportunity and comparability of results. In recent years, large-scale portfolio assessment has gained popularity because of its potential to bridge the worlds of public accountability and classroom practice and as the ultimate example of assessment productively integrated with instruction. In contrast to more contrived, "drop from the sky" assessments (Hoover, 1996), portfolios are the products of ongoing classroom work. Targeted on agreed-upon standards for student performance but still permitting teachers and students a great deal of flexibility and choice, portfolio assessments challenge teachers and students to focus on meaningful work, support the assessment of long-term projects over time, encourage student-initiated revision, and provide a context for presentation, guidance, and critique of student progress. However, the very strengths of portfolio assessment—its flexibility and direct integration with classroom practice—at the

same time present a number of measurement weaknesses, if indeed portfolios are to provide accurate information about student performance (Gearhart & Herman, 1995). Obvious problems of comparability arise from the variability and generalizability of tasks that are included in students' portfolios and the differential conditions under which the work was produced. Findings from the Vermont statewide portfolio assessment clearly indicated, for example, that teachers vary widely in their classroom portfolio practices. There was substantial variation in the amount of time students spent on their portfolios, and in classroom policies on revision. Some teachers encouraged revision, others discouraged it, still others required it. Policies on feedback and support similarly were variable; in some classrooms, getting feedback from others was permissible, in other classrooms it was explicitly forbidden (Koretz, Stecher, Klein, & McCaffrey, 1994; Koretz, Stecher, Klein, McCaffrey, & Deibert, 1993).

Perhaps the most vexing among these challenges to comparability of results is summed up by the words of a Vermont teacher after having scored portfolios for several days: "Whose work is this anyway?" The question naturally arises because portfolios contain the products of classroom instruction, and good classroom instruction, according to current theory, engages communities of learners in a supportive learning process (Camp, 1993; Wolf, Bixby, Glenn, & Gardner, 1991; Wolf & Gearhart, 1993). Thus, under optimal instructional conditions, the product(s) being assessed are not the result of a single individual but rather of an individual working in a social context. The better the instructional process, the more an individual student's work is likely to have benefited from others.

How to infer an individual's competence from such supported performance is one important aspect of the problem, but perhaps more important is the differential support that students receive within and across classrooms. While the Vermont study documents differential help with portfolio work across classrooms, Gearhart et al. (1993) report substantial variation within classrooms as well. The researchers asked teachers such things as how much structure or prompting they provided individual students, what types of peer or teacher editorial assistance occurred, and what resources and time were available for portfolio work. Patterns differed across teachers. Within classrooms, not surprisingly, teachers tended to provide more help to lower ability students, students who most needed it.

What this means is that the quality of a student's work is a function of variable and potentially substantial external assistance (another source of measurement error) as well as of the student's capability. As a result, the validity of inferences we can draw about individual student competence based solely on portfolio work is highly suspect. Because for some students the work has been highly assisted and for other students little assisted, comparisons of students' capability based on such work is unfair.

Inferences from group work. Inferring individual performance from assessments that include group work raises similar issues. Research by Noreen Webb (1993) suggests that an individual's performance in the context of group activity may or may not represent his or her capability. Webb found that low-ability students achieved higher scores on the basis of group work than on the basis of the same work produced individually, suggesting that group assessments are likely to overestimate the performance of these students. Further, Webb also has identified a number of variables in group composition and process, such as ability level and gender of group members, and experience in group process, that influence performance. To the extent that such variables are irrelevant to the knowledge and skills that are the targets of the assessment, the results of group assessments may be biased against some students—for example, students who were not members of optimally composed groups.

These issues in portfolio and group assessment are not grave concerns for classroom assessment where teachers can judge students' performances with knowledge of their context, and where teachers draw on a variety of sources of evidence to make inferences and decisions about students. The problems are troubling indeed for large-scale assessment programs where performance is judged by those outside the classroom, where results tend to stand in isolation, and where comparability of data is essential.

Consequences of Alternative Assessment

Despite the many technical challenges, alternative assessment does appear to have value in supporting instructional reform, based on accumulating evidence from implementation studies. Lorraine McDonnell, for example, conducted comparative case studies in Kentucky and North Carolina to analyze the extent to which the two states' assessment systems promote those states' curricular reform goals and encourage classroom teaching practices consistent with those

goals (McDonnell & Choisser, forthcoming). The two systems are quite different in assessment design and incentive structure, with Kentucky being the more radical of the two. Based on state standards for what students need to know and be able to do, the Kentucky Instructional Results Information System (KIRIS) is composed of portfolios, performance events (including group and hands-on activities), and short-answer and multiple-choice questions in five subjects at each of three grade levels (Grades 4 or 5 depending on assessment, 8, and 11). Assessment results are used to give schools cash rewards for success in meeting performance goals—amounting to $1300 to $2600 per teacher—and threaten take-over if school goals are not met or adequate progress based on test results is not accomplished. North Carolina's system, in contrast, is relatively low-stakes and less a departure from its previous assessment system. Students in Grades 3-8 are tested at the end of each year in reading, mathematics and social studies using a combination of multiple-choice and short-answer items. Students in Grades 4 and 6-8 also produce a writing sample.

Despite these differences, McDonnell and Choisser found that educators in the two states had similar reactions to their state assessment programs. Teachers and principals took the assessments very seriously, were generally supportive of the reform goals embodied by the assessments, but were ambivalent about the assessments themselves. Teachers saw value for students in that the tests encouraged teachers to engage students in activities such as writing and problem solving that otherwise would be absent or less frequent in the curriculum, and they viewed the assessments as more complete measures of student accomplishment than previous tests. At the same time, they questioned the validity of the assessments for some students and were concerned about the stress the assessments place on them and their students (see also Koretz, Baron, et al., 1996). McDonnell and Choisser's analysis of instructional artifacts, teacher logs, and surveys also showed that the content and process of teachers' classroom practices generally conformed with the goals of the reform, although in both states there was evidence that teachers lacked thorough understanding of the meaning of the reform and the specific kinds of learning that were required by the assessments.

A variety of impacts. Similar pictures of mixed support and generally beneficial impact on curriculum and instruction emerge from studies of statewide systems in Maryland and Vermont (Koretz, Mitchell, et al., 1996; Koretz, Stecher,

et al., 1993). Koretz, Mitchell, et al. (1996), however, point out that from some vantage points, there is negative spillover from some of the positive influences on curriculum. For example, while Maryland teachers reported instructional changes consonant with the goals of the state reform (i.e., increases in writing for a variety of purposes, analysis of text, literary comprehension, mathematical communication, data analysis, use of graphs and tables, and meaningful problem solving), these increases also meant less instructional time spent in areas not assessed. Some teachers worried about lack of attention to basic skills, such as grammar and number facts, among other things. As the discussion below indicates, vocal parents and members of the community sometimes share these concerns.

Aschbacher's action research (1994) also shows that teachers' involvement in the development and implementation of alternative assessments has diverse, positive influences on teaching practices, at least when combined with training and follow-up technical support:

- Two thirds of the teachers reported substantial change in the ways they thought about their own teaching. As two teachers explain,

  > "I have begun to look at teaching from a different vantage point. I can see more possibilities" (p. 20);
  >
  > "The portfolios seem to mirror not only the students' work but the teacher's as well. As a result, I have found the need to re-work, re-organize, and re-assess my teaching strategies" (p. 22).

- Two thirds reported at least some increase in their expectations for students—more thinking and problem solving and/or higher levels of performance from students.

- For the majority, the experience of working with alternative assessments reinforced the importance of purpose or goals.

Clark and Stephens (1996) document similar effects of assessment in Australia. Their study shows that the implementation of the Victoria Certificate of Education supported systematic reform of mathematics education and effectively influenced curriculum and teaching.

Effects depend on local conditions. Mary Lee Smith's case studies on the consequences of the Arizona State Assessment Program (ASAP) remind us that such changes may be highly dependent on the culture, philosophy and leadership,

and other local conditions evident in individual schools (Smith, 1997). Her study reveals schools where changes were dramatic in direct response to ASAP, schools that were changing anyway and would have done so with or without the program, and schools where no changed occurred or was possible. In the first category was a school whose teachers were predisposed to the constructivist goals implicit in ASAP but with little knowledge of how to pursue them. With a supportive principal and resources for intensive professional development, these teachers were able to change teaching throughout the school. In another school, although some teachers were similarly predisposed and even knowledgeable about intended changes, a "persuasive climate of behaviorism" (p. 99) limited the impact. Nonetheless, these teachers were able to use the ASAP mandate to advance their agenda and introduce change in the tested grade. In contrast, in two other schools very little change occurred. One school was geographically remote and lacked resources for new materials or professional development, and the other was permeated by beliefs that their children were too poor and limited in English language and ability to profit from new curriculum and instructional goals. These two schools well demonstrate two critical variables that shape implementation outcomes: the will and the capacity to change (McDonnell & Choisser, forthcoming; McLaughlin, 1987).

Benefits Carry Costs

Although the literature is promising with regard to the potential positive effects of alternative assessment on curriculum and instruction, research also indicates the significant challenges and time such systems entail. For example, a majority of principals interviewed in Vermont believed that portfolio assessment generally had beneficial effects on their schools in terms of curriculum, instruction and/or effects on student learning and attitudes, but almost 90% of these same principals characterized the program as "burdensome," particularly from the perspective of its demands on teachers (Koretz, Stecher, et al., 1993). Nearly every project, in fact, reports concerns about pressure on teachers and the pervasive demands on teachers' time (Aschbacher, 1993; Koretz, Mitchell, et al., 1996; Koretz, Stecher, et al., 1993; Wolf & Gearhart, 1993): time for teachers to become familiar with the new assessments and their administration, to understand how tasks are developed and scored, to discern and apply criteria for assessing students' work, to develop the content and pedagogical knowledge they need to change their practice, to reflect upon and fine-tune their instructional and

assessment practices, etc. Such time and the professional development efforts that need to undergird it represent both important and significant costs in implementing new assessment systems.

The time demands of portfolio assessment programs appear particularly acute. The Vermont study, for example, asking about only some of these demands, found that teachers devoted 17 hours a month to finding portfolio tasks, preparing portfolio lessons, and evaluating the contents of portfolios; and 60% of the teachers surveyed at both fourth and eighth grades indicated they often lacked sufficient time to develop portfolio lessons (Koretz, Stecher, et al., 1993). Again, these time estimates represent important opportunity costs for both teachers and students.

Similarly, it is clear that most states implementing new assessments must make sizable investments in professional development for teachers—either directly through state efforts or pushed down as a local district responsibility. In the Maryland sample, for example, two thirds of the teachers had engaged in at least one professional activity to explain the Maryland assessment program or performance assessment in general; nearly half gained knowledge by participating in the assessment by either developing or scoring assessment tasks; and over half participated in at least one professional development activity dedicated to develop content or pedagogical knowledge related to the assessed outcomes (Koretz, Mitchell, et al., 1996).

Economic costs. Which of these costs should be directly ascribed to assessment as opposed to instructional or professional development components of the educational system is one of the issues that make it difficult to estimate the costs of alternative assessments. Although conceptual models for analyzing the cost of alternative assessment and for conducting cost-benefit analyses have been formulated (Catterall & Winters, 1994; Picus, 1994), definitive cost studies are yet to be completed (see, however, Picus & Tralli, forthcoming). Nonetheless, it is clear that compared to traditional multiple-choice tests, the costs of development, administration, scoring, and reporting of alternative assessment are dramatically higher (Hardy, 1995; Hoover & Bray, 1995; Stecher, 1995). For example, Koretz, Madaus, Haertel, and Beaton (1992) estimate that Advanced Placement exams, which typically take three hours and require extended essay responses, cost approximately $65 per subject test, whereas commercial, standardized tests cost from $2 to $5 per subject test.

One area where direct comparisons are perhaps easiest is scoring. Here, for example, Catterall and Winters (1994) estimate the cost of scoring a 45-minute essay as part of the California assessment system at between $3 and $5. Stecher (1995) estimates the cost of scoring a hands-on science task comprising one class period at $4 to $5 per student. In comparison, the complete battery of the Iowa Tests of Basic Skills, a nationally standardized multiple-choice test, costs about $1 per student.

Public Credibility and Support

As with any public policy, investment in alternative assessment is dependent on the support of the public and its policy makers. Of late, some segments of the public have been vocal in their opposition and, in some states, have been successful in derailing new systems. While it is axiomatic to many educators that schools must emphasize complex thinking and problem solving if students are to be well prepared for future success and life-long learning, and that good instruction and good assessment alike are constructivist, parents and the community do not necessarily agree. A prominent national survey, for example, found that the public is very concerned about students' basic skills and believes that schools should put "first things first" (Johnson & Immerwahr, 1994). Public controversies over assessments in both Kentucky and California (McDonnell, 1997) indicated as well that parents in some cases misunderstood the nature of the tests and disagreed with underlying values. For example, opposition groups questioned the academic rigor of the tests. Some also were offended by the social and cultural agenda they saw underlying the assessments because some assessment materials represented diverse viewpoints or were perceived as encouraging children to question authority—for example, a language arts question: "Think about a rule at your school . . . that you think needs to be changed. Write a letter to your principal about the rule you want changed" (McDonnell, 1996, p. 42). Also at issue was whether the assessments inappropriately intruded into family life and violated parents' rights because some questions asked students to reflect on events in their lives—for example, "Why do you think some teenagers and their parents have problems communicating?" (McDonnell, 1996, p. 42). Although the political motives and wider agenda of some in the opposition might be open to question, what was clear from these examples is that significant segments of the population did not understand the aims or purposes of the new assessments and did not feel that their viewpoints were considered in the development process. The diversity of

opinion within the public—about what is important for students to know and be able to do, and what are the goals of schooling—also is clear, as is the need to involve parents and community activity in all phases of the assessment process.

Largely because of the public controversy, the lighthouse California Learning Assessment System—and with it a $32 million, two-year investment (Picus & Tralli, forthcoming)—was abruptly discontinued. Also at issue, although less visibly so in the California case, were concerns about technical quality; and in fact, the latter were the primary rationale for the recent discontinuation of the Arizona performance assessment (ASAP; Smith, 1997).

What Next?

The last decade has witnessed an explosion of interest in performance assessment in the United States as a policy tool to support accountability and school improvement. With great enthusiasm and commitment to change, many in the United States have rediscovered what most countries outside the United States have long understood: that multiple-choice and other selected response testing cannot be the sole basis for assessment systems, and that essay and other open-response questions deserve an important role (Keeves, 1994). Set in a unique American environment that values the efficiency and psychometric qualities of multiple-choice testing and fears the litigation that might accompany high-stakes assessments that do not meet technical standards, states, local school districts and others across the country have embarked on developmental efforts to bring their vision of alternative assessment to reality.

As a result, the last five or so years have been a period of great experimentation and learning in the United States, which has produced substantial knowledge about the strengths and challenges of alternative assessment for accountability purposes. The consequences of using performance assessment in accountability systems appear to be a clear strength: Teachers and principals take the new assessments and the goals they represent seriously and move to incorporate new pedagogical practices into their teaching; teachers engage their students in the kinds of activities they see embodied in the assessment. With appropriate professional development and supportive local context, new assessment indeed can support meaningful change and improvement of practice.

The technical and logistical challenges, however, are daunting indeed. The possibility of providing accurate, reliable, individual results using performance assessments alone seems remote. Based on available evidence, the number of tasks or items required to get a stable, individual estimate makes it unlikely that any state or local system would be able or willing to invest the necessary student time or financial resources. While available methods do make possible school-level estimates that can be used for a variety of purposes, educators and the public alike clamor for individual results from their assessment systems. United States parents want to know—and demand formal, comparable evidence of—what their children are learning, and students likewise want to know their progress. Similarly, like teachers around the world, teachers in the United States seek information they can use to understand individual students and how best to support their learning.

The public controversies in a number of states and communities, furthermore, underscore a real diversity of opinion in what children ought to know and be able to do and the types of assessments that should be used to measure accomplishments. The strength of sentiment for local control and the difficulty of moving to more centralized systems that are taken for granted elsewhere in the world are evident. The costs of new forms of assessment in current times of fiscal austerity and public cutbacks also have given policy pause. In fact, two states that were in the forefront of innovation in assessment—Arizona and California—have seen their programs discontinued, and several other states that were moving in that direction have had their funding derailed. The policy engine that was steaming ahead just two years ago today appears to have slowed a bit.

The last five years serve as a clear reminder of the complexities of moving from the simple assumptions of policy to solutions that work in reality. The challenge of designing beneficial assessment systems to accommodate multiple interests and to serve multiple purposes within given constraints also is apparent. Past history makes evident the folly of relying solely on multiple-choice testing for accountability purposes, a lesson that other countries had no need to learn; more recent history suggests that alternative assessments alone—at least at this stage of their development and under the time constraints and costs that those in the United States are willing to bear—will not suffice. The obvious and sensible solution towards which most states and local districts are now moving is an

optimal combination of both. How to configure such systems represents an important and ambitious research and development agenda for our future work.

References

Aschbacher, P. R. (1993). *Issues in innovative assessment for classroom practice: Barriers and facilitators* (CSE Tech. Rep. No. 359). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

August, D., Hakuta, K., & Pompa, D. (1994). *For all students: Limited English proficient students and Goals 2000* (NCBE Occasional Papers in Bilingual Education No. 10). Washington, DC: National Clearinghouse for Bilingual Education.

Baker, E. L. (1991, April). Panel member. In *Authentic assessment: The rhetoric and the reality*. Symposium conducted at the annual meeting of the American Educational Research Association, Chicago.

Baker, E. L. (1994). Learning-based assessments of history understanding. *Educational Psychologist, 29,* 97-106.

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.

Baxter, G., Elder, A. D., & Glaser, R. (1996). Knowledge-based cognition and performance assessment in the science classroom. *Educational Psychologist, 31,* 133-140.

Baxter, G., & Glaser, R. (1996). *An approach to analyzing the cognitive complexity of science performance assessments* (CRESST deliverable to OERI). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Bolger, N., & Kellaghan, T. (1990). Method of measurement and gender differences in scholastic achievement. *Journal of Educational Measurement, 27,* 165-174.

Bond, L. A., Braskamp, D., & Roeber, E. (1996). *The status report of the assessment programs in the United States*. Oakbrook, IL: North Central Regional Educational Laboratory and the Council of Chief State School Officers.

Borkowski, J. G., & Muthukrishna, N. (1992). Moving metacognition into the classroom: Working models and effective teaching strategies. In M. Pressley, K. R. Harris, & J. T. Guthrie (Eds.), *Promoting academic competence and literacy in school* (pp. 477-501). San Diego: Academic Press.

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.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models: Applications and data analysis methods*. Newbury Park, CA: Sage.

California State Department of Education. (1992). *California Assessment Program: A sampler of mathematics assessment*. Sacramento, CA: Author.

Camilli, J., & Shepard, L. A. (1994). *Methods for identifying biased test items*. Newbury Park, CA: Sage.

Camp, R. (1990). Thinking together about portfolios. *The Quarterly of the National Writing Project and the Center for the Study of Writing, 12*(2), 8-14, 27.

Camp, R. (1993). The place of portfolios in our changing views of writing assessment. In R. E. Bennett & W. C. Ward (Eds.), *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment* (pp. 183-212). Hillsdale, NJ: Lawrence Erlbaum Associates.

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.

Carroll, J. B. (1975). *The teaching of French as a foreign language in eight countries*. New York: John Wiley.

Catterall, J., & Winters, L. (1994). *Economic analysis of testing: Competency, certification, and "authentic" assessments* (CSE Tech. Rep. No. 383). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

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.

Clark, D., & Stephens, M. (1996). The ripple effect: The instructional impact of the systematic introduction of performance assessments in mathematics. In M. Birenbaum & F. J. R. C. Dochy (Eds.), *Alternative in assessments of achievements, learning processes and prior knowledge* (pp. 63-93). Boston: Kluwer Academic.

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.

Corbett, H. D., & Wilson, B. L. (1991). *Testing, reform, and rebellion*. Norwood, NJ: Ablex Publishing.

Darling-Hammond, L. (1995). Equity issues in performance based assessment. In M. T. Nettles & A. L. Nettles (Eds.), *Equity and excellence in educational testing and assessment* (pp. 89-114). Boston, MA: Kluwer.

Darling-Hammond, L., & Wise, A. E. (1985). Beyond standardization: State standards and school improvement. *The Elementary School Journal, 85,* 315-336.

Davis, R. B., Maher, C. A., & Noddings, N. (Eds.). (1990). Constructivist view on the teaching of mathematics. *Journal for Research in Mathematics Education Monograph No. 4.* Reston, VA: National Council of Teachers of Mathematics.

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.

DuBois, P. H. (1966). Test dominated society: China 1155 BC-1905 AD. In A. Anastasi (Ed.), *Testing problems in perspective* (pp. 29-36). Washington, DC: American Council on Education.

Dunbar, S. B., Koretz, D. M., & Hoover, H. D. (1991). Quality control in the development and use of performance assessments. *Applied Measurement in Education, 4,* 298-303.

Egan, M., & Bunting, B. (1991). The effects of coaching on 11+ scores. *British Journal of Educational Psychology, 61,* 85-91.

Elementary and Secondary Education Act of 1965, 20 U.S.C. §§ 236 *et seq.,* 821 *et seq*.

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

Feuer, M. J., & Fulton, K. (1994). Education abroad and lessons for the United States. *Educational Measurement: Issues and Practice, 13*(2), 31-39.

Garcia, T., & Pintrich, P. R. (1994). Regulating motivation and cognition in the classroom: The role of self-schemas and self-regulatory strategies. In D. H. Schunk & B. J. Zimmerman (Eds.), *Self-regulation of learning and performance* (pp. 127-153). Hillsdale, NJ: Lawrence Erlbaum Associaties.

Gearhart, M., & Herman, J. L. (1995, Winter). Portfolio assessment: Whose work is it? Issues in the use of classroom assignments for accountability. *Evaluation Comment*, 1-16.

Gearhart, M., Herman, J., Baker, E. L., & Whittaker, A. (1993). *Whose work is it? A question for the validity of large-scale portfolio assessment* (CSE Tech. Rep. No. 363). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Gitomer, D. H. (1993). Performance assessment and educational measurement. In R. E. Bennett & W. C. Ward (Eds.), *Construction versus choice in cognitive measurement* (pp. 241-263). Hillsdale, NJ: Lawrence Erlbaum Associates.

Glaser, R. (1991). Expertise and assessment. In M. C. Wittrock & E. L. Baker (Eds.), *Testing and cognition* (pp. 17-30). Englewood Cliffs, NJ: Prentice-Hall.

Glaser, R. (1992). Expert knowledge and processes of thinking. In D. F. Halpern (Ed.), *Enhancing thinking skills in science and mathematics* (pp. 63-75). Hillsdale, NJ: Lawrence Erlbaum Associates.

Glaser, R., & Silver, E. (1994). Assessment, testing and instruction. *Annual Review of Psychology, 40,* 631-666.

Glass, G., & Ellwein, M. C. (1986, December). Reform by raising test standards. *Evaluation Comment*, 1-6.

Goals 2000: Educate America Act, Pub. L. No. 103-227, 108 Stat. 125 (March 31, 1994).

Hardy, R. A. (1995). Examining the costs of performance assessment. *Applied Measurement in Education, 8*, 121-134.

Herman, J. L. (1996). Technical quality matters. In R. Blum & J. A. Arter (Eds.), *Performance assessment in an era of restructuring* (pp. 1-7:1–1-7:6). Alexandria, VA: Association for Supervision and Curriculum Development (ASCD).

Herman, J. L., Aschbacher, P. R., & Winters, L. (1992). *A practical guide to alternative assessment*. Alexandria, VA: Association for Supervision and Curriculum Development (ASCD).

Herman, J. L., Gearhart, M., & Baker, E. L. (1993). Assessing writing portfolios: Issues in the validity and meaning of scores. *Educational Assessment, 1*, 201-224.

Herman, J. L., & 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.

Herman, J. L., & Golan, S. (1993). Effects of standardized testing on teaching and schools. *Educational Measurement: Issues and Practice, 12*(4), 20-25, 41-42.

Herman, J. L., & Klein, D. C. (1996). Assessing equity in alternative assessment: An illustration of opportunity-to-learn issues. *Journal of Educational Research, 89,* 246-256.

Herman, J. L., Osmundson, E., & Pascal, J. (1996). *Evaluation of the Los Alamos National Laboratory Critical Issues Forum*. Los Angeles: University of California, Center for the Study of Evaluation.

Holland, P. W., & Wainer, H. (Eds.). (1993). *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Hoover, H. D. (1996, May). *Practical and technical issues in the development of performance assessments in a large-scale testing program*. Paper presented at a meeting at the National Center for Research on Evaluation, Standards, and Student Testing, UCLA, Los Angeles.

Hoover, H. D., & Bray, G. B. (1995). *The research and development phase: Can performance assessment be cost effective?* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Johnson, J., & Immerwahr, J. (1994). *First things first: What Americans expect from the public schools*. New York: Public Agenda.

Keeves, J. (1994). Tests: Different types. In T. Husen & T. N. Postlethwaite (Eds.), *International encyclopedia of education* (2nd ed., Vol. 11, pp. 6340-6349). Oxford/New York: Pergamon/Elsevier Science.

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

Koretz, D. M., Barron, S., Mitchell, K. J., & Stecher, B. M. (1996). *Perceived effects of the Kentucky Instructional Results Information System (KIRIS)*. Santa Monica, CA: RAND.

Koretz, D. M., 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.

Koretz, D. M., Madaus, G. F., Haertel, E., & Beaton, A. E. (1992). *National educational standards and testing: A response to the National Council on Education Standards and Testing*. Santa Monica, CA: RAND.

Koretz, D. M., McCaffrey, D., Klein, S., Bell, R., & Stecher, B. (1993). *The reliability of scores from the 1992 Vermont Portfolio Assessment Program* (CSE Tech. Rep. No. 355). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Koretz, D. M., Mitchell, K. J., Barron, S., & Keith (1996). *Final report: Perceived effects of the Maryland School Performance Assessment Program* (CSE Tech. Rep. No. 409). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Koretz, D. M., Stecher, B., Klein, S., & McCaffrey, D. (1994). The Vermont Portfolio Assessment Program: Findings and implications. *Educational Measurement: Issues and Practice, 13*(3), 5-16.

Koretz, D. M., Stecher, B., Klein, S., McCaffrey, D., & Deibert, E. (1993). *Can portfolios assess student performance and influence instruction? The 1991-92 Vermont experience* (CSE Tech. Rep. No. 371). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Learning Research and Development Center and the National Center on Education and Economy. (1991). *The New Standards Project: An overview*. Pittsburgh, PA/Washington, DC: Author.

LeMahieu, P., Gitomer, D., & Eresh, J. (1994). *Portfolios beyond the classroom: Data quality and qualities* (MS # 94-01). Princeton, NJ: Educational Testing Service, Center for Performance Assessment.

Linn, R. L. (1993). Educational assessment: Expanded expectations and challenges. *Educational Evaluation and Policy Analysis, 15*, 1-16.

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

Linn, R. L., & Burton, E. (1994). Performance-based assessments: Implications of task specificity. *Educational Measurement: Issues and Practice, 13*(1), 5-8, 15.

Linn, R. L., Burton, E., DeStefano, L., & Hanson, M. (1995). *Generalizability of New Standards Project 1993 pilot study tasks in mathematics* (CSE Tech. Rep. No. 392). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Linn, R. L., Graue, M., & Sanders, N. (1990). 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.

Lomask, M., Baron, J., Greig, J., & Harrison, C. (March, 1992). *ConnMap: Connecticut's use of concept mapping to assess the structure of students' knowledge of science*. A symposium presented at the annual meeting of the National Association for Research in Science Teaching, Cambridge, MA.

Madaus, G. F. (1988). The influence of testing on the curriculum. In L. N. Tanner (Ed.), *Critical issues in curriculum* (Eighty-seventh yearbook of the National Society for the Study of Education, Part 1, pp. 83-121). Chicago: University of Chicago Press.

Madaus, G. F. (1991). The effects of important tests on students: Implications for a national examination system. *Phi Delta Kappan, 73*, 226-231.

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.

McDonnell, L. M. (1997). *The politics of state testing: Implementing new student assessments* (CSE Tech. Rep. No. 424). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

McDonnell, L. M., & Choisser, C. (forthcoming). *Testing and teaching: Local implementation of new state assessments*. Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

McLaughlin, M. W (1987). Learning from experience: Lessons from policy implementation. *Educational Evaluation and Policy Analysis. 9,* 171-178.

Moerkerke, G. (1996). Assessment for flexible learning. Performance assessment, prior knowledge state assessment and progress assessment as new tools (Academisch proefschrift). Heerlen, NL: Open Universiteit. (Handelseditie: Uitgeverij Lemma te Utrecht).

Muthén, B., Huang, L. C., Jo, B., Khoo, S. T., Goff, G., Novak, J., & Shih, J. (1995). Opportunity to learn effects on achievement: Analytic aspects. *Educational Evaluation and Policy Analysis, 17*, 371-403.

Myford, C., & Mislevy, R. (1995). *Monitoring and improving a portfolio assessment system* (Research Rep. 94-05). Princeton, NJ: Educational Testing Service, Center for Performance Assessment.

National Academy of Education, Committee on Educational Research. (1969). *Research for tomorrow's schools: Disciplined inquiry for education*. Stanford, CA: National Academy of Education/Macmillan.

National Governors' Association. (1996). *Policy statement.* National Education Summit, Palisades, NY, March 26-27. Available: http://www.summit96.ibm.com/brief/finaledpol.html

Picus, L. O. (1994). *A conceptual framework for analyzing the costs of alternative assessment* (CSE Tech. Rep. No. 384). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

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.

Resnick, L. B., & Resnick, D. P. (1992). Assessing the thinking curriculum: New tools for educational reform. In B. G. Gifford & M. C. O'Conner (Eds.), *Changing assessments: Alternative views of aptitude, achievement and instruction* (pp. 37-75). Boston: Kluwer Academic Publishers.

Resnick, L. B., Resnick, D., & DeStefano, L. (1994). *Cross-scorer and cross-method comparability and distribution of judgments of student math, reading, and writing performance: Results from the New Standards Project Big Sky scoring* (CSE Tech. Rep. No. 368). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Rogosa, D. (1994). *Misclassification in student performance levels* (Technical report prepared for the California Learning Assessment System). Stanford, CA: Stanford University.

Rogosa, D. (1995). Myths and methods: Myths about longitudinal research. In J. M. Gottman (Ed.), *The analysis of change* (pp. 3-65). Hillsdale, NJ: Lawrence Erlbaum Associates.

Rogosa, D., & Sanger, H. M. (1995). Longitudinal data analysis examples with random coefficient models. *Journal of Educational and Behavioral Statistics, 20,* 149-170.

Schoenfeld, A. H. (1991). On mathematics as sense-making: An informal attack on the unfortunate divorce of formal and informal mathematics. In J. F. Voss, D. N. Perkins, & J. W. Segal (Eds.), *Informal reasoning and education* (pp. 311-343). Hillsdale, NJ: Lawrence Erlbaum Associates.

Seltzer, M. H., Frank, K. A., & Bryk, A. S. (1994). The metric matters: The sensitivity of conclusions about growth in student achievement. *Educational Evaluation and Policy Analysis, 16,* 41-49.

Shavelson, R. J., Baxter, G. P., & Gao, X. (1993). Sampling variability of performance assessments. *Journal of Educational Measurement, 30,* 215-232.

Shavelson, R. J., Baxter, G. P., & Pine, J. (1991). Performance assessment in science. *Applied Measurement in Education, 4,* 347-362.

Shavelson, R. J., Mayberry, P. W., Li, W-C., & 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, National Center for Research on Evaluation, Standards, and Student Testing.

Shepard, L. (1991). *Will national tests improve student learning?* (CSE Tech. Rep. No. 342). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Slavin, R. E. (1990). *Cooperative learning: Theory, research and practice.* Englewood Cliffs, NJ: Prentice-Hall.

Smith, M. L. (1994, September). *How assessments work: Lessons learned in equity.* Presentation at the annual conference of the National Center for Research on Evaluation, Standards, and Student Testing, Los Angeles.

Smith, M. L. (1997). *Reforming schools by reforming assessment: Consequences of the Arizona Student Assessment Program (ASAP): Equity and teacher capacity building* (CSE Tech. Rep. No. 425). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

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

Stecher, B. (1995, April). *The cost of performance assessment in science.* Invited symposium presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Sugrue, B. (1996). *The relative validity and reliability of different assessment formats for measuring domain specific problem-solving ability.* Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Tamir, P., & Doran, R. (in press). Science process skills in six countries. *Studies in Educational Evaluation*.

Webb, N. (1993). Collaborative group versus individual assessment in mathematics: Processes and outcomes. *Educational Assessment, 1,* 131-389.

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. F. (1995). Performance-based assessments: Contributor or detractor to equity? In M. T. Nettles & A. L. Nettles (Eds.), *Equity and excellence in educational testing and assessment* (pp. 221-241). Boston, MA: Kluwer.

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.

Wolf, D. P. (1992). Good measures: Assessment as a tool for educational reform. *Educational Leadership, 49*(8), 8-13.

Wolf, D. P., Bixby, J., Glenn, J., & Gardner, H. (1991). To use their minds well: Investigating new forms of student assessment. In G. Grant (Ed.), *Review of Research in Education* (Vol. 17, pp. 31-74). Washington, DC: American Educational Research Association.

Wolf, R. M. (1994). Performance assessment in IEA studies. *International Journal of Educational Research*. *21*, 239-245.

Wolf, S. A., & Gearhart, M. (1993). *Writing what you read: Assessment as a learning event* (CSE Tech. Rep. No. 358). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing/Center for the Study of Evaluation.