DOCUMENT RESUME

ED 381 549 TM 022 832

AUTHOR Glaser, Robert; Silver, Edward

TITLE Assessment, Testing, and Instruction: Retrospect and Prospect.

INSTITUTION National Center for Research on Evaluation, Standards, and Student Testing, Los Angeles, CA.; Pittsburgh Univ., Pa. Learning Research and Development Center.

SPONS AGENCY Office of Educational Research and Improvement (ED), Washington, DC.
REPORT NO CSE-TR-379
PUB DATE Jun 94
CONTRACT R117G10027
NOTE 40p.
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Educational Assessment; Educational Change; Educational History; Educational Improvement; Educational Practices; Educational Testing; Elementary Secondary Education; *Instructional Effectiveness; *Selection; Standardized Tests; *Student Placement; *Teaching Methods; Test Construction; *Test Use; Track System (Education)
IDENTIFIERS Reform Efforts

ABSTRACT

Some of the deficiencies and abuses associated with past testing practices are reviewed, and some of the present and future possibilities for educational assessment are explored. At this time, assessment and testing in American schools are caught between the rhetoric of reform and the intransigence of long-established practices. Use of measurement of intellectual abilities for educational purposes has followed two lines of historical development: testing for selection and placement and assessment of educational outcomes. Mounting evidence of the negative consequences resulting from use of selection testing for differential placement in academic tracks calls for questioning the wisdom of such test use in schools. Assessment of school achievement has become increasingly institutionalized and separated from instruction. Ways to ensure that testing and teaching interact to work for the improvement of instruction are the focus of much current interest. New forms of assessment responsive to the needs of instruction are the goals of future test development. (Contains 94 references.) (SLD)

CRESST

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

Assessment, Testing, and Instruction: Retrospect and Prospect

CSE Technical Report 379

Robert Glaser and Edward Silver
CRESST/Learning Research and Development Center
University of Pittsburgh

- UCLA Center for the Study of Evaluation

in collaboration with:

- University of Colorado
- NORC, University of Chicago
- LRDC, University of Pittsburgh
- The RAND Corporation

Assessment, Testing, and Instruction: Retrospect and Prospect

CSE Technical Report 379

Robert Glaser and Edward Silver
CRESST/Learning Research and Development Center
University of Pittsburgh

June 1994

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

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

ASSESSMENT, TESTING, AND INSTRUCTION: RETROSPECT AND PROSPECT[1],

Robert Glaser and Edward Silver

CRESST/Learning Research and Development Center
University of Pittsburgh

In recent years, testing and assessment have been much on the minds of educational policy makers at local, state, and national levels. There has been increasing interest in the results of testing and assessment and increasing concern about the nature and form of student assessment and the uses made of the results (Linn, Baker, & Dunbar, 1991). Interestingly, in the current debates about poor educational outcomes and the need for education reform, assessment and testing have been viewed both as part of the problem and as part of the solution. On the one hand, assessment and testing are often portrayed as a major cause of current educational woes, as is illustrated in the following excerpt from the report of the National Commission on Testing and Public Policy (1990): "Current testing, predominantly multiple choice in format, is over-relied upon, lacks adequate public accountability, sometimes leads to unfairness in the allocation of opportunities, and too often undermines

---

[1] To appear in L. Darling-Hammond (Ed.), *Review of research in education, Volume 20*. Washington, DC: American Educational Research Association. We acknowledge the valuable contribution of our consulting editors, Jeremy Kilpatrick and Bob Linn, both of whom made helpful comments on an earlier draft.

The chapter is based on a paper, "Testing and Assessment: O Tempora! O Mores!", originally prepared by the first author for the 31st Horace Mann Lecture at the University of Pittsburgh, October 1990. Preparation of that lecture and this chapter were sponsored in part by the National Research Center on Student Learning at the Learning Research and Development Center and funded by the Office of Educational Research and Improvement of the U.S. Department of Education. Preparation of this chapter was also supported by a grant from the Ford Foundation for the QUASAR project; additional support was provided by the National Center for Research on Evaluation, Standards, and Student Testing (CRESST). The opinions expressed herein are those of the authors and do not necessarily reflect the views of the Ford Foundation or CRESST.

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

vital social policies" (p. ix). On the other hand, others have argued that assessments linked to high standards for student achievement are valuable in helping establish a more fruitful educational climate and more equitable educational outcomes: "Alternative forms of assessment, forms currently within reach, can adequately reflect today's educational goals and, if properly used, serve as positive tools in creating schools truly capable of teaching students to think" (L.B. Resnick & Resnick, 1992, p. 38).

Since testing and assessment figure prominently in current discussions about the improvement of American education, it seems a propitious moment to examine some deficiencies and abuses associated with past practices in educational measurement and to consider present and future possibilities. Moreover, it is critical to examine how forms of measurement interact with the forms of instruction that are also being called for in current reform discussions. At this point in time, assessment and testing in American schools are caught between the extensive rhetoric of reform and the intransigence of long-established practices.

Considering Testing and Assessment in Settings of Use

Testing and assessment, as they have been institutionalized in contemporary educational systems, represent the product of earnest attempts of prior generations to meet the conditions of earlier times. However, the conditions of today and tomorrow demand different measurement and educational solutions. Just as attempts to balance the budget, to control pollution, and to assist disadvantaged populations may involve outdated processes and can produce dysfunctional consequences, so it is with some of our efforts to improve education through testing and assessment.

Sound educational policy requires the appropriate use of measures of human ability and attainment. What matters is how tests and assessments are designed, how they are used, and how the consequences of implementation affect educational policy and practice. The measures used are derived from some theory of human ability, either tacitly assumed or explicitly described. However, their operational significance emerges as they are used to serve these ends. Often overlooked or deemphasized is the fact that measures are employed in *settings of use*. The setting—the context of testing—is crucial. It either enhances the value of measures or impoverishes them; their original

intent is either well or poorly realized. This interaction between testing or assessment and the surrounding setting or system is of fundamental significance.

Use of measurement of intellectual abilities for educational purposes has followed two lines of historical development: (a) testing for selection and placement and (b) assessment of educational outcomes. Selection placement testing is coordinate with work on individual differences in human intelligence; it extends the concept of testing developed by Binet, in France, early in this century, as measurement of the ability or aptitude to profit from schooling. The history of assessment of educational outcomes in the United States began with the thrust toward universal education, the rise of the idea that education is the key to success, and the consequent political demands for instruments by which schools could be made accountable.

These two uses of testing and assessment reflect different social and technical histories and different goals and purposes as well. Selection testing attempts to measure human abilities prior to a course of instruction so that individuals can be appropriately placed, diagnosed, certified, included, or excluded. In contrast, assessments of educational outcomes are designed to measure the results of a course of learning in relation to intended or unintended consequences of education. The demarcation between the two is not always clear because the results of testing and assessment are often the basis for decisions about a subsequent set of educational experiences. Nevertheless, selection testing is primarily used to predict success in learning. What is important is a selection test's predictive validity. In contrast, the goal of assessing educational outcomes is to describe the nature of performance that results from learning. In this case, the content validity and the nature of acquired performance are both critical. Selection tests are designed to capture capabilities that develop from educational and background experiences in school and out of school. Assessments of educational outcomes attempt to measure school achievement directly. A look at past experience in school settings in the use of selection testing and assessment of educational outcomes can help us understand their prese[illegible]eneficial or pernicious influences.

Selection Testing

Selection testing is thought to have begun in China around 2200 B.C. with proficiency testing to determine qualifications for government service (DuBois, 1965). The system was gradually refined over several millennia, and, "despite a concentration on literary rather than managerial skills, the system was to serve as a model for a number of efforts at standardizing competition for civil service positions in Europe and the United States during the 19th century" (McArthur, 1987). Ironically, the Chinese abandoned their civil service examination system in 1905, just as it was being widely copied elsewhere (DuBois, 1965). In Europe and the United States, selection testing flourished in the first half of the 20th century.

A primary influence on the development of selection testing was the pioneering work of Binet in developing a measure of intelligence. Binet's major contribution was the use of tasks that were closer to those that might be encountered in everyday life than the elemental responses, such as discrimination and reaction time, that had been used by Galton, Cattell, and a few other researchers who sought an understanding of individual differences in human mental functioning (Carroll, 1978). By 1905, Binet had spent about 15 years accumulating data on individual differences. His writings indicate that he had not been able to approximate a satisfactory definition of the nature of intelligence but that he remained convinced of the need for an instrument to measure this quality (Curtis & Glaser, 1981). His work in developing a test of intelligence spawned the growth of testing to manage the selection and sorting of individuals for a wide variety of societal purposes. For much of the 20th century, educational settings have provided significant application contexts for selection testing.

Differential Placement in Schools

As testing gained general acceptance for the management of individuals, various testing programs for differential selection and placement were instituted in school systems. Schools, faced with increasingly diverse student populations, were viewed as natural settings in which to apply testing technology. In fact, Binet's pioneering work on the measurement of intelligence was closely associated with a practical, school-related goal. The minister of public education of France wanted some means to ensure that

instructional resources were not squandered on children who would have difficulty learning. In the early 1900s, Binet and his colleague Simon developed tests—based on the approach taken in constructing Binet's intelligence test—for use in Paris in Binet's laboratory school to identify children unlikely to succeed in normal classes and therefore in need of special instruction (T. H. Wolf, 1973).

The advent of compulsory schooling placed increased pressure on the educational system, and selection testing provided a means to release some of the pressure on the system. By 1926, compulsory education laws had increased the percentage of high school-aged children who attended school to four times what it had been in 1910 (Pintner, 1931). Another source of pressure on the educational system was the need to deal with the increasing number of European immigrants. In a country whose population included a large segment of non-native-born individuals, compulsory education laws also meant that "Americanization" became one of the primary functions of schooling. Throughout the first quarter of this century, the influx of southern and eastern Europeans presented new challenges for educators and the country as a whole. Fears were roused about changes in the American way of life, and, as new waves of immigration swelled, labor unions began to worry about surplus workers. In this context, tests offered scientific legitimization for those fears. The test scores of immigrants were lower than their native-born counterparts, and concern was expressed about a declining level of national intelligence (Pintner, 1931).

As school populations continued to swell and diversify with compulsory education and immigration, differences in the levels of ability and aptitude among students continued to plague educators. Learning disability, with various etiologies, was offered as an explanation for this uneven educational progress, and school systems began to use test scores for ability grouping. The use of testing to identify children with special needs fairly quickly became a common practice in educational settings. Testing provided a convenient and powerful instrument of social control for those in the late 19th and early 20th centuries who sought to use tests as a means to create the "one best system" of education (Tyack, 1974). As D. Wolf, Bixby, Glenn, and Gardner (1991) note, the notion of differentiated instruction to accommodate the needs of diverse learners had appeal even to most progressive educators. The possibility of

adapting to differences between students' achievement and their rates of learning encouraged the widespread adoption of such instruction. The use of tests for selective purposes was considered a significant factor in the successful management of instruction.

Although arguments related to educational quality and social justice were often used as rationales for the widespread use of testing for selecting and sorting students in order to determine educational opportunities, many have subsequently argued that these practices were, in fact, motivated by racial, ethnic, and gender politics (e.g., Gould, 1981; Kamin, 1974; Mercer, 1989). Linda Darling-Hammond (1994) provides a particularly compelling example:

> Terman found many inequalities in performance among groups on his IQ test, adapted from Binet's work in France. Most, but not all of them, seemed to confirm what he, and presumably every other "intelligent" person already knew: that various groups were inherently unequal in their mental capacities. However, when girls scored higher than boys on his 1916 version of the Stanford-Binet, he revised the test to correct for this apparent flaw by selecting items to create parity among genders in the scores (Mercer, 1989). Other inequalities—between urban and rural students, higher and lower SES [socioeconomic status] students, native English speakers and immigrants, whites and blacks—did not occasion such revisions, since their validity seemed patently obvious to the test-makers.

Despite some objections to the widespread use of tests to determine educational opportunity, testing for selection and placement became an institutionalized practice in our schools and in other parts of society early in the 20th century. In fact, selection testing was used in primary and secondary education settings, in selection for university education, in qualification for government civil service positions, and for placement in jobs in the military. Lee Cronbach (1975), in a review of 50 years of controversy over testing, later commented on the scene in this way:

> William James had warned psychology that to understand man was not to write his biography in advance, but the testers came very close in their estimate as to how much education a man could use and what careers he could thrive in. More serious, when the tests determined who would enter the college preparatory program and before that determined who would go into the "fast" section of an early grade, the tests began to determine fates. (p. 11)

Even in current educational practice, tests are frequently used to sort students into instructional "tracks" that provide differential opportunities and have differential expectations. And the general failure of this process to increase educational outcomes has been amply demonstrated.

Dysfunction in Contemporary Selection Testing

The dysfunctionality of the practice of educational tracking has been documented in accounts of the inequitable distribution of educational opportunity. The crisis in exclusionary practice is illustrated in one study of tracking that reported that "although disproportionately white classes were found to be about equally likely to be identified as low- or high-ability, disproportionately minority classes were *seven times more likely to be identified as low-ability than as high-ability* (Oakes, 1990, p. 23). Studies have shown that tracking, rather than allowing students access to instruction that maximizes educational outcome and increases life chances, relegates disproportionate numbers of poor and minority students to the lower instructional tracks, where little is expected of them, where they receive little meaningful instruction, and where they find themselves blocked from access to further educational opportunities (Oakes, 1990; Oakes & Lipton, 1990; Rosenbaum, 1980).

In studies of the academic tracking of students for mathematics instruction, data regarding instructional practices suggest that students assigned to the lower tracks of many high schools tend to receive less actual mathematics instruction, less homework, and more drill and practice of low-level factual knowledge and computational skill than students assigned to middle and higher tracks (Oakes, 1985). These instructional practices may contribute to increased performance on tasks requiring only basic factual knowledge or on routine computational skills, but they are clearly unlikely to lead to improvements on more complex tasks requiring mathematical reasoning and problem solving, which are precisely the kinds of tasks that children in the highest instructional tracks more frequently encounter and the ones that are viewed as foundational for future success in further education or in employment (National Research Council, 1989).

Thus, the practice of selecting for and selecting out is pervasive in schooling. Some individuals are included into further education as a result of

these measures, and some are excluded and placed in watered-down instruction that does little to enhance educational opportunities. On the surface the process of selection may appear to cultivate the best talent; however, at the core it sets in motion an involuted causal cycle of undeveloped ability. Many students are excluded from certain educational experiences presumably because they lack attributes that promise successful learning, but these are the very same attributes that could be learned in appropriate educational environments.

The irony is that individuals are thus excluded from the very learning experiences that could develop the capabilities they need in order to be included. For example, low-achieving children's background experiences frequently do not expose them to the information and modes of cognizing that are useful in school learning and future educational access and that can be learned if they participate in environments in which this knowledge and ability are exercised. Tests and other criteria used to make readiness and retention decisions identify a disproportionate number of poor and minority children as not ready for regular schooling and so place them in lean curricula that are not likely to promote learning skills (Shepard, 1991).

A series of studies by Lorrie Shepard and her colleagues concerning school practices in the identification and placement of children with learning disabilities, in which tests play a part along with other information, provides a particularly disturbing example of how the well-intended uses of selection testing can have perverse consequences (Shepard, 1989; Shepard & Smith, 1983). Given current testing technology, errors of measurement that are due to the limits of test reliability, for example, would be expected to result in misidentification; one might expect that there would be valid cases missed, as well as normal children mislabeled. What one finds instead is that the category of learning disabled (LD), in fact, is swamped by overidentification. Only about 40% of the schoolchildren who tested as learning disabled were legitimately LD by either strong or weak signs of their abilities and achievements as displayed in the typical classroom. The remaining cases included educable mentally retarded (EMR) and emotionally disturbed children, children who had a language background other than English, and those who were slow learners, had minor behavior problems, or were average learners in districts of high socioeconomic status (Shepard, 1983; Shepard,

Smith & Vojir, 1983). As a group, the putatively LD students were often indistinguishable from other low achievers, except that minority children were overidentified in this category and the children labeled LD may have had parents who visited school often to show concern for their children.

Schools are under pressure from parents, teachers, and administrators to label as eligible for special services children who do not readily fit the categories but who otherwise might not get a fair chance in the educational system. Ironically, school authorities often deliberately misidentify children in order to obtain services for them. As Shepard (1983) has argued:

> Against this pressure there is no countervailing force, no professional reason suggesting that clinicians should resist labeling a child LD when the diagnosis is invalid. In fact, specialists are eager to provide a service. After all, they entered special education as a helping profession. Furthermore, the professional literature has increasingly drifted toward a service-oriented definition of LD. The more confused experts become about the etiology of the disorder . . . the more it is suggested that the handicap be defined in terms of instructional need. (p. 7)

Rejecting the label of LD seems tantamount to denying children help. Although attempts are made to make this label nonstigmatizing, LD implies that the cause of the learning problem is in the child rather than in the instructional system.

Indeed, there are several unhappy sides to this well-intended situation. Students who are labeled as unable to profit from the mainstream have been placed in classrooms that have limited resources and space, and occasions for developing the knowledge and skills that are needed for success in the regular classroom are limited. In general, efforts to redesign instruction have been little considered when placement in a separate part of the system is possible. Adjustments within existing teaching and instructional systems are difficult to make in any case, but testing practices keep the structure intact. In essence, selection decisions can be manipulated so that the instructional system is preserved, and effective learning is precluded for many students who could be assisted in profiting from the system. Furthermore, the special grouping of children with different learning abilities means that teachers' experience with a range of learning styles in their classrooms is narrowed. Thus, in this scenario the positive contributions of testing to the redesign of

instruction are often overshadowed by the contributions of testing to maintaining organizational continuity (Cronbach, 1984).

Academic tracking has been a common context for the use of selection testing in American education. As has been shown, the application of testing to select students into and out of educational experiences has often been more successful in replicating inequities than in providing students with opportunities to overcome obstacles and improve their life chances. The net result of the use of testing for sorting and selecting is that challenging curricula and high educational expectations are rationed to a very small proportion of students. It does not appear that the testing, selection, and tracking cycle provides even children identified as "gifted" or "advanced" with rich instructional opportunities to ensure high educational achievement (Slavin, 1990). The impoverished educational opportunities that result from this situation for most American students manifest themselves in the poor performance of these students, across all tracks and ability levels, in national and international assessments of educational achievement (McKnight et al., 1987; Mullis, Owen, & Phillips, 1990).

The determinist potential of the use of selection testing had been warned against by some in the late 19th and early 20th centuries (e.g., Rankin, 1931), but the wisdom of sorting students into distinct programs failed to receive the scrutiny it deserved at that time (D. P. Resnick, 1982). Again today, the mounting evidence of negative consequences resulting from the use of selection testing for differential placement in academic tracks compels us to question the wisdom of continuing such uses of testing in schools.

Assessment of Achievement

In contrast to selection testing, which was developed on the basis of explicit, although narrow, conceptions of aptitude and conceptions of intelligence, the theory underlying the assessment of school achievement is less explicit. Techniques for measuring achievement and the growth of competence, as they developed historically, tended to rely on the psychometric technology that emerged in the context of selection and aptitude testing. Thus, achievement testing has generally lacked adequate psychological theories of human competence and performance, which are needed for the assessment of achievement.

In recent decades, significant improvements have been made in assessment, from the perspectives of both technology and underlying theory. The technology has improved somewhat as demands for content validity have become insistent (e.g., demands for diagnosis and mastery testing, for national assessment and local accountability, and for data that describe the accomplishments and competences of learners rather than rank them) (Cronbach, 1970). The art and practice of achievement assessment has progressed through such ideas as criterion-referenced testing and anchor-point performance reference, and, more recently, authentic assessment, portfolio procedures, curriculum-embedded assessment, and analysis of cognitive process requirements of subject matter have come under consideration. The underlying psychological theory has matured from the behavioral theories of the mid-20th century that generated behavioral objectives but could not adequately describe complex processes of thought, reasoning, and problem solving, to more cognitive accounts of complex human performance, thereby laying the foundation for a theory and psychometrics of performance measurement (Bennett & Ward, 1993; Mislevy, Yamamoto, & Anacker, 1992; Shepard, 1992). Increasingly, long-encouraged (e.g., Glaser, 1981) innovative procedures and situations that assess high levels of competence and reasoning abilities realizable in schoolchildren and adults are being introduced. Nevertheless, at present, much of this work is experimental, and the most common practices in the current assessment of achievement in the national educational system have changed little in the last 50 years.

Pupil Comparisons and Program Accountability

The assessment of achievement in American schools began as early as 1845, when the Boston School Committee, under pressure from Horace Mann, the state commissioner of education, instituted a comprehensive survey of pupils' attainments to justify the appropriations provided to them by the state of Massachusetts (Kilpatrick, 1992). The examiners published a table ranking the schools of Boston in order of the achievement of their pupils on a series of written tests in various subject areas. By the 1870s, written achievement examinations were being used in many states and large school districts, and the results were often presented in newspapers (Tyack, 1974). Examinations

also developed into "high-stakes" events for students, since by the end of the century promotion from grade to grade, which had previously been based solely on teacher recommendation, began to depend on success or failure on a written examination (Engelhart, 1950).

Given the high stakes associated with the assessments, educators began to worry about variability and inconsistency in administration of these achievement measures, and they began to request standardization and comparative information. Spurring this effort was a series of studies that showed that the grading of traditional oral and essay examinations was unreliable and often unfair. The first published national subject examinations that established norms for grade-level performance appeared in the 1890s, and further development of standardized tests followed quickly thereafter. Early in the 20th century, achievement tests of all sorts were developed and commercially marketed; these tests were designed so that they could be adopted by many school systems using different materials and methods. By the time the United States entered World War I, there were more than 200 achievement tests available for use in primary and secondary schools (D. P. Resnick, 1982). The periodic administration of standardized assessments became common practice in larger American schools as a means to monitor teacher and program performance and to compare class and grade achievement within and between school districts. As the measurement of educational achievement became institutionalized, the technologies for norming and establishing content that was not especially tailored to the goals and standards of particular schools became established.

The first Stanford Achievement Test appeared in 1923. Its publication, like the appearance of Thorndike's (1904) textbook on educational measurement a few years earlier, was a landmark in the history of modern educational measurement, and it foreshadowed the future. The test was comprised of a battery of standardized achievement measures that spanned several elementary school subjects. It displayed many of the characteristics of tests today: It was constructed by trained professionals; its content was drawn from a survey of representative courses of study in all parts of the country; its items were tried out experimentally; and it was administered to thousands of schoolchildren, in many different school systems, to obtain comparative

samples of performance and norms. With its publication, standardized achievement tests passed quickly into maturity.

Certainly, it was recognized, even at the time of introduction, that such assessments of achievement were fundamentally measures of recall and recognition. Other aspects of learning, however, remained extremely difficult to measure. The constraints of task format and of assessment administration in the school structure frustrated attempts to assess complex skills and perpetuated reliance on simpler measures. Moreover, the move toward nationally developed and nationally normed achievement tests shifted the practice of assessment of educational outcomes away from its roots in assessing outcomes through *examinations* tied to particular curricula. This shift took firm hold and has remained largely in place to this day, as is expressed in the observation of D. P. Resnick and Resnick (1985) that American students are the most frequently tested and least often examined students in the world.

The 1960s marked a period of expanded educational assessment. The press for increased access to educational opportunity and heightened interest from the federal government in management by objectives led to accountability testing at national and state levels. Federal legislation obliged the states to assume responsibility for the provision of equal educational opportunity. In 1965, the Elementary and Secondary Education Act (ESEA), Title I, called for financial assistance and special services for low-income students and districts and required performance data on students receiving assistance and evaluation data on outcomes of the funded programs (D. P. Resnick, 1980).

At the time, there was enthusiasm for indicators that would measure progress toward the goal of providing all students with a good education. Until the 1960s, federal data on education had been dominated by information on enrollments and graduation rates; however, with the concern in the 1960s regarding civil rights and the Soviet Union's launching of Sputnik I, questions were raised about the quality and content of American education. The federal government responded with two initiatives. The first was the Equality of Educational Opportunity Survey (EEOS), which provided information on the achievement of more than 600,000 children in elementary and secondary schools. The analysis of the EEOS data, known as the Coleman Report (Coleman et al., 1966), documented the enormous variation in achievement of

12th graders and showed that graduation rates revealed little about what graduates learned in school. More revealing forms of assessment seemed necessary.

The second federal initiative of the 1960s was the creation of the National Assessment of Educational Progress (NAEP), which provided for periodic assessment of achievement in specific school subjects for students at different age levels as well as for the comparison of trends in their achievement levels over time. NAEP was designed primarily to supply a much-needed indicator of the quality of education in the way that other indicators, such as unemployment statistics and the gross national product, provide information about the economy. As the first of the successive waves of NAEP scores appeared, concerns were raised about clear differences in the scores of various populations.

Over the past decade, as is well known, efforts to devise useful assessment and accountability measures have proliferated along with state programs. The federal initiative expanded so that the NAEP could provide state data on mathematics, reading, and possibly other subjects to all states who so desired. In May 1990, the National Assessment Governing Board approved a document endorsing the establishment of three national levels of subject matter achievement: basic, proficient, and advanced. The way in which standards of this kind are being developed is currently under study (National Academy of Education, 1993; Phillips et al,, 1993).

Dysfunctionality in Uses of Assessment for Accountability and School Improvement

As can be seen from the brief history presented here, the measurement of achievement and the measurement of school accountability have been linked since the earliest appearances of achievement measures, and this linkage continues. Achievement measurement has become increasingly institutionalized and has been a focal point of attention on indicators of school effectiveness. However, much less attention has been paid to how assessments might be used to shape and improve learning and schooling. Standardized assessment and the conditions of instruction and schooling have coexisted largely as decoupled systems. Aside from teacher-made classroom tests, the integration of assessment and learning as an interacting system has been too

little explored. As a National Institute of Education (1979) report of a conference on assessment and instruction noted: "Current testing procedures are not helpful to teachers or students in their day-to-day efforts to teach and learn" (p. v), and "present day testing programs are largely extraneous to everyday classroom teaching" (p. 359). Given this disconnection between assessment and instruction, researchers and educators alike have called for changes that would result in test formats being more aligned with instructional tasks and test results being more useful for instructional decision making (Glaser, 1986; Linn, 1983; Nitko, 1989; Silver & Kenney, in press).

Although the measurement of school achievement has not been powerfully linked to instruction in positive ways, many have noted some negative linkages. For example, standardized assessments of school achievement have been criticized because they can be seen to symbolize the wrong outcomes as being of central import in schooling. An important "function of testing is to signal to students, teachers, and the general public those aspects of learning that are valued" (Silver & Kilpatrick, 1988, p. 180). In the area of mathematics, the National Research Council (1989) has reacted strongly to limitations in the symbolic value of currently available achievement measures:

> As we need standards for curricula, so we need standards for assessment. We must ensure that tests measure what is of value, not just what is easy to test. If we want students to investigate, explore, and discover, assessment must not measure just mimicry mathematics. (p. 70)

The misalignment of the content of achievement assessments and important curricular goals and standards in mathematics has also been seen as related to the limitations of the multiple-choice format used in most commercially produced assessments. Given the current interest in promoting complex reasoning and problem solving, the fact that these assessments have tended not to include questions in which students are required to produce their own answers, to display the processes used to obtain an answer, to explain the thinking or reasoning associated with their response, or to exhibit alternative approaches to or interpretations of a problematic situation has severely limited the extent to which they are seen as related to important curricular goals (Silver, 1992).

Achievement assessments have also been criticized for their negative effects on classroom climate and instructional practice. Evidence that externally mandated assessments can limit and negatively affect the quality of mathematics instruction has been accumulated from many sources (e.g., Madaus, West, Harmon, Lomax, & Viator, 1992; Salmon-Cox, 1982; Smith, 1991). In general, research has found that teachers are influenced by their perceptions of the content of externally mandated assessments, especially when the assessment results are viewed as having important consequences for themselves or for their students. In particular, research suggests that teachers tend to narrow their instruction by giving a disproportionate amount of their time and attention to teaching the low-level specific content most heavily assessed rather than teaching underlying concepts or overarching principles or unassessed or less assessed areas (e.g., geometry, statistics) that are also expected to be part of the curriculum (Madaus et al., 1992).

Externally mandated assessments can also affect classroom-level activity in other ways. Teachers often create or use multiple-choice and short-answer assessments, thereby evoking and evaluating performances from their students only in forms identical to those used on external assessments. The widespread use of multiple-choice assessment has contributed to a "dumbing down" of instruction, in which skills tend to be taught in the form required for performance on the assessment rather than for more realistic or natural applications (Darling-Hammond & Wise, 1985). One recent study reported that this tendency is more prevalent for teachers of high-minority classes than for teachers of low-minority classes in urban school districts. For example, 74% of the teachers of high-minority classes reported beginning test preparation activities at least 1 month before an externally mandated assessment, and more than 30% reported spending at least 20 hours of class time in preparation; however, only 32% of the teachers of low-minority classes reported beginning preparation 1 month or more before an assessment, and only 9% reported spending 20 or more hours (Madaus et al., 1992).

Under current circumstances, assessment-driven instruction creates a dilemma for many good teachers for whom teaching within a narrowly circumscribed, assessment-defined space is not acceptable. Externally mandated assessments cannot be ignored, since the teachers, their students, and their schools are likely to be judged on the basis of student achievement.

Yet, teaching only the content of these assessments is also unacceptable. Several reports (e.g., Livingston, Castle, & Nations, 1989; McNeil, 1988) suggest that reform-minded teachers attempt to overcome the inadequacies of the system through a kind of "double-entry" curriculum and instruction in which they attempt to give sufficient attention to the narrow goals embodied in the external assessments without sacrificing instructional attention to deeper conceptual understandings or broader curricular goals. Although some teachers find ways to teach high-level content despite the pressures of external assessment, many do not. And the problems appear to be worst in urban school districts, where more than 60% of mathematics and science teachers report that externally mandated assessments have a negative impact on their curriculum or instruction (Madaus et al., 1992). Clearly, there is a need to address the current dysfunctionality by improving the interaction between assessment and instruction to ensure that these two facets of educational activity work in harmony rather than at cross purposes.

Assessment and Instruction: A Look Toward the Future

At present, policies to ensure that testing and teaching interact to inform each other for the improvement of instruction are being actively considered. Subject matter specialists, test developers, teachers, and school policy makers are devoting increased effort to the design and use of assessment in the context of instruction. For example, the possibility of achieving a symbiotic relationship between mathematics assessment and instruction has been envisioned by the National Research Council (1989): "Assessment should be an integral part of teaching. It is the mechanism whereby teachers can learn how students think about mathematics as well as what students are able to accomplish" (p. 69).

In addition to what may be possible at the level of classroom assessment, a few external assessments have been created and implemented in the area of mathematics with the intention of being sensitive to the high-level instructional goals advocated in forward-thinking curriculum frameworks and instructional guides (e.g., California State Department of Education, 1989; Silver & Lane, 1993). Moreover, impressive prototypes of new forms of extended assessment tasks that measure high-level goals in subjects such as mathematics have been developed (e.g., National Research Council, 1993).

Progress is also being made in how the results of students' performance can be reported. We can soon expect that learning assessments will not provide merely a score, a label, a grade level, or a percentile but also instructional scoring that makes apparent to the student and to the teacher the requirements for increasing competence (Glaser, 1986). The kinds of detailed qualitative analyses of students' performance provided by Magone, Cai, Silver, and Wang (in press) represent one prototype of instructionally useful approaches that go beyond simply assigning responses to score levels. This perspective on assessment, however, demands that more explicit attention be paid to matters of instruction.

One way to begin to realize wiser, more constructive uses of educational measurement is to determine the kinds of educational settings and social values to be served. Because testing and assessment are social and political as well as technical artifacts, we need to acknowledge their setting and the values and aims that would be entailed in their most beneficial uses. Concepts of education need to support and be supported by improved uses of tests and assessments. The current educational system operates according to a selective model, and its approaches to testing and assessment have developed accordingly. In the selective mode of education, there is minimal variation tolerated in the conditions of learning and a narrow range of available instructional options. Thus, the educational system primarily benefits those whose backgrounds and out-of-school support systems are best matched with the expectations of formal schooling, with its narrow conception of learning and ability. Testing and assessment are important management components of this selective mode of education, and they play a key role in schooling's frequent reproduction of inequities existing in the larger societal context.

Toward New Conceptions of Education and the Role of Assessment

Many have called for a radical reformulation of the basic assumpt ns of education. For example, Gardner (1990) has called for a new model of education termed *individually configured excellence*. In the context of a modern theory of human development, individually configured excellence bears some resemblance to the older notion of *adaptive education*, in which a selective model of education is replaced by one designed to support inclusiveness (Glaser, 1972, 1977). The notions of adaptive education and

individually configured excellence argue that the primary function of schooling is not to select and sort students into rigidly defined ability categories but to identify and nurture sources of competence in individual students. Such conceptions of education assume that schooling can provide for a range of opportunities adjusted to individuals and their backgrounds, talents, interests, and prior performance as they move toward achieving the goals of education required for general societal literacy and significant life opportunities. Information about progress toward instructional goals would be available both to the teacher and the student as learning proceeds. The effect of the student's choice of or assignment to a learning opportunity would be evaluated on the basis of the progress that she or he makes in realizing the goals of competence and potential for future learning. The role of assessment in such an education is to help teachers and students in attaining the goal of helping all children "use their minds well" (D. Wolf et al., 1991).

In many ways, these current conceptualizations of education bear close familial resemblance to earlier notions of progressive education. As Linda Darling-Hammond (1993) has noted:

> The criticisms of current education reformers—that our schools provide most children with an education that is too rigid, too passive, and too rote-oriented to produce learners who can think critically, synthesize and transform, experiment and create—are virtually identical to those of the Progressives at the turn of the century, in the 1930s, and again in the 1960s . . . Indeed, with the addition of a few computers, John Dewey's 1900 vision of the 20th-century ideal is virtually identical to current scenarios for 21st-century schools. (p. 755)

Although similar to many earlier education ideals, the current conceptualizations of education reform do have several distinctive features, among which is a compelling research base on the nature of human learning and performance in complex intellectual domains and a recognition of the increasingly complex character of our contemporary pluralistic society.

The appropriateness of an adaptive mode of education is suggested by the findings of several decades of cognitive research that has pointed to the constructive nature of human learning, the complex nature of expertise related to specific subject areas, the power of intuitive conceptions, and the limitations of school knowledge for application in nonschool settings. Bolstered by this research knowledge, proponents of reform have begun to

consider strategies for promoting greater instructional and curricular emphasis on thinking and reasoning and have begun to develop new descriptions of competence and proficiency for many school subjects that emphasize such themes as thinking, reasoning, complex performance, and problem solving in addition to knowledge and skills (e.g., National Council of Teachers of Mathematics, 1989). These descriptions of competence and proficiency support a view of adaptive education, and they stand in opposition to prevailing traditional views that have long undergirded the selective mode of education. Assessment is central to the tension between selective and adaptive modes of education, since the distinction is concerned not only with the identification of appropriate goals for the school curriculum but with the nature of appropriate means of attaining these goals and measuring the extent of attainment. Many (e.g., L. B. Resnick & Resnick, 1992) have identified the mismatch between the goals of the "thinking curriculum" and current tests and testing practice.

An adaptive mode of education is especially relevant to today's aspirations for schooling and the requirements of education for our nation. If a more relevant and more substantive version of education is to be made available to all children in school, then some changes will clearly be needed in the delivery of instruction. In her review of research on teaching high-level thinking and reasoning skills, L. B. Resnick (1987) concluded that developing higher order cognitive abilities requires shaping a disposition to thought through participation in social communities that value thinking and independent judgment. This suggests a view of classrooms as communities of collaborative, reflective practice in which students are challenged to think deeply about and to participate actively in engaging the subjects they are learning (Bruer, 1993). Applying this view to mathematics classrooms, Silver, Kilpatrick, and Schlesinger (1990) have argued that communication would become a more central feature: "Within communities, the need for communication is obvious. Within mathematical communities, communication in the form of discussion, argument, proof, and justification is natural" (p. 23). In such communities, students would be expected not only to listen but to speak mathematics themselves as they discuss observations and share explanations, verifications, reasons, and generalizations. In such classrooms, students would have opportunities to see, hear, debate, and

evaluate mathematical explanations and justifications. These classrooms, as Silver et al. (1990) have noted, become places in which "the emphasis is less on memorizing procedures and producing answers and more on analyzing, reasoning and becoming convinced" (p. 38).

All students can benefit from learning in classroom communities in which high-level thinking and communication are emphasized. In mathematics, for example, some general examples of such teaching have been provided (e.g., National Council of Teachers of Mathematics, 1991), some examples have been described of teaching in fairly privileged settings (e.g., Lampert, 1986), and some cases have also been provided by the QUASAR project, which works with schools serving students in economically disadvantaged communities (Silver, 1993; Silver, Smith, & Nelson, in press). Thus, evidence is beginning to accumulate regarding the effectiveness of such forms of education with diverse student populations. Moreover, a recent examination of the educational practices used with linguistically and culturally diverse student populations found that collaboration and communication were key elements of effective instructional practice at all levels of the educational system, especially when the curriculum contained a blend of both challenging and basic academic material (Garcia, 1991). Thus, the kinds of classroom communities described above represent a new vision of education that is compatible with the precepts of adaptive education or individually configured excellence and aimed at allowing equitable access to high-quality instruction and challenging content for all students.

An example of how assessment becomes linked to this conception of education can be seen in an excerpt drawn from the QUASAR project (Silver et al., in press). In the first year of the project, teachers at one of the participating urban middle schools administered the following open-ended task to help students prepare for the administration of the QUASAR Cognitive Assessment Instrument.[2]

---

2 The QUASAR Cognitive Assessment Instrument consists of a set of open-ended tasks that assess mathematical reasoning, mathematical problem solving, the understanding of mathematical concepts, and communication of mathematical explanations or justifications (see Lane, 1993, or Silver & Lane, 1993, for additional information on the instrument).

Busy Bus Company Problem.

Yvonne is trying to decide whether she should buy a weekly bus pass. On Monday, Wednesday, and Friday she rides the bus to and from work. On Tuesday and Thursday she rides the bus to work, but gets a ride home with her friends. Should Yvonne buy a weekly bus pass? Explain your answer.

Busy Bus Company Fares

One Way: $1.00
Weekly Pass: $9.00

At a subsequent meeting, the teachers met to discuss their students' performance, which had some surprising aspects. In particular, many students indicated that Yvonne should purchase the weekly pass rather than paying the daily fare, which teachers believed to be the more economical choice. Curious about this unexpected answer to what the teachers believed to be a rather straightforward question—a multistep arithmetic story problem involving multiplication of whole numbers—they decided to discuss the problem in class and ask students to explain their thinking. The ensuing discussion with students provided an interesting illustration of their application of out-of-school knowledge and problem-solving strategies to a mathematics problem. Many students argued that purchasing the weekly pass was a much better decision because the pass would allow many members of a family to use it (e.g., after work and in the evenings), and it could also be used by a family member on weekends. Students' reasoning about this problem—situated in the context of urban living and the cost-effective use of public transportation—demonstrated to the teachers that there was more than one "correct" answer.[3] This experience made it clear to the teachers that if their goal was assessing what students know and are able to do, then it was essential that students not only provide answers but also explain their thinking and reasoning.

In applying to become part of the QUASAR project, the teachers at this middle school had noted, to their dismay, that their "mathematics instruction did not consider, nor did it utilize, the cultural background of their student population"; also, it did not capitalize on the array of problem-solving skills

[3] The task developers intended this task to have more than one correct solution, depending on the nature and quality of the explanation and reasoning provided. Thus, the children's response was not as surprising to the task developers as it was to the teachers.

students brought with them from their home environments. The experience with the Busy Bus Company problem caused them to pay explicit attention to students' nonschool knowledge and problem solving. Moreover, the students' responses illustrated to the teachers that increasing the relevance of school mathematics to the lives of children involves more than merely providing real-world contexts for mathematics problems; real-world solutions for those problems must also be considered. The forms of reasoning and problem solving that are developed and used in out-of-school settings can be brought into close contact with the forms of reasoning and problem solving being developed in school mathematics, thereby providing students with opportunities to come to understand the conditions that optimize the application of each form.

As the above excerpt from QUASAR suggests, instructional approaches compatible with notions of adaptive education or individually configured excellence are likely to address a widely noted deficiency of conventional instruction related to the tendency of students to perceive school instruction as involving domains that are disconnected from sense making and the world of everyday experience. This phenomenon has been extensively studied in the area of mathematics (e.g., Nunes, Schliemann, & Carraher, 1993; L. B. Resnick, 1988; Schoenfeld, 1991). In one relevant recent study (Silver, Shapiro, & Deutsch, 1993), middle school students were asked to provide interpretations for an answer to a division problem intended to be about a real-world situation: The Clearview Little League is going to a Pirates game. There are 540 people, including players, coaches, and parents. They will travel by bus, and each bus holds 40 people. How many buses will they need to get everyone to the game? The general finding was that students' responses dealt more with technical concerns than with sense making. Many proposed answers that involved a fraction of a bus (even though they knew that buses do not have fractional parts) apparently because the technical process of computation produced a fractional answer. Students' dissociation of sense making from mathematical activity was evident not only from the responses they provided but from the explanations they did not give, since reports from the students' teachers suggested that some children engaged in more sense making than was evident in their written responses. Apparently, students did not perceive their "sensible" answers (e.g., using a minivan could serve as a practical

representation for a "fractional part" of a full bus) as being valid in the context of responding to a mathematics problem. The requirement that mathematics should make sense was apparently not a feature of students' mathematics instruction.

The results reported by Silver et al. (1993) also identified another deficiency of conventional mathematics teaching: Students had difficulty providing explanations of their reasoning or justifications for their answers. Explanations and interpretations, in oral or written form, are not a regular feature of instructional activities in mathematics classrooms, and this has serious implications for the assessment of students' understanding. Thus, this example actually illustrates a way in which the better integration of assessment and instruction could have mutual benefits. Not only can assessment be improved through the use of tasks more closely aligned with important curricular objectives, but instruction can also be improved when the results of student performance on assessment tasks aligned with high-level objectives reveal instructional insufficiencies that lead to superficial understanding.

Toward New Forms of Assessment Responsive to the Needs of Instruction

Educational measurement is likely to result in the promotion and improvement of learning only when at least two certain conditions are met. First, the outcomes being tested must be recognized and accepted as important objectives of the instructional program. If this is not the case, the assessment program either can be disregarded as being peripheral or can deflect teaching and the educational program from central goals. Second, achievement assessment must be planned and implemented as an integral part of the curriculum and program of instruction. Only insofar as assessments are constructed or selected in terms of the instructional program and the results are available for formative planning and change can their greatest value be realized.

Many efforts have brought together teachers and researchers who design classroom situations in which students engage in cooperative efforts to learn and build their knowledge. Didactic teaching in these programs is largely replaced by learning opportunities in which understanding, efforts at problem solving, and the communication and appropriate use of knowledge can be

displayed and observed (Brown & Palincsar, 1984; Cobb, Wood, & Yackel, 1991; Silver et al., in press). In these contexts, students are able to monitor their performances and observe the performances of more competent individuals more consistently than is possible in situations in which learning and problem solving proceed individually and silently and the end product or the answer is of singular importance. Assessment in these environments is intimately tied to ongoing performances that indicate functional achievement in subject matter domains. Such activities as taking on problems, devising problems for others, and discussing levels of understanding provide displays and integral informal assessments of achievement. Because students are offered wider opportunities for learning and the assessment of their skill and knowledge is integrated with their studies, the limitations of conventional tests are avoided (Brown, Campione, Webber, & McGilly, 1992; Silver & Lane, 1993.

These and other current developments provide great promise for the future development of a system of educational measurement that is truly linked with and supportive of instructional programs aimed at educational excellence and equity. If assessments integral to instruction become common practice, the nature of testing and assessment surely will change, and we offer a few glimpses of the ways in which this might occur. Embedded in this vision of new ways of thinking about measurement and its uses, one will find many of the concepts and criteria that have been expressed by others in recent writings on educational measurement (California Mathematics Council, 1989; Frederiksen & Collins, 1989; Gardner, 1992; Linn et al., 1991; L. B. Resnick & Resnick, 1992; Wiggins, 1989): access, fairness, transparency and openness, consequential or systemic validity, cognitive significance, content quality, self-assessment, and socially situated assessment. These and other related notions provide not only a sense of the possibilities but a set of important criteria to use in determining the utility and value of alternative assessment proposals and activities.

As settings are designed in which teaching and testing are integrated, the utility of current learning for future learning will be emphasized. Testing for selection and exclusion will not obscure the view that promoting learning skill and optimizing individual competence is a primary objective of schooling. In order to lessen the exclusionary aspects of testing, tests should be designed to survey possibilities for student growth rather than to designate students as

ready or not ready to profit from standard instruction. For example, the use of norm-referenced tests as the sole measure of eligibility for special programs may be replaced by the use of integrated programs of assessment and instruction that enable teachers to recognize and support children's strengths so they can achieve in more powerful curricula. As one teacher of special education students writes:

> Perhaps if schools were to drop their screening procedures, to stop sorting out children on the basis of test results, and to refrain from predicting success or failure for entering students, they would be free to accept all children as learners with unique and interesting abilities. Staffs and small groups of teachers could work together to support each other's strengths, and thus support children's strengths, instead of dwelling on problems. (Martin, 1988, p. 501)

In this vision of the future, testing is seen as being less about sorting and selecting and more about offering information on which students and teachers can build.

As assessment and instruction are more closely linked, achievement measurement will be integral to learning rather than imposed by some external shaper of students' fates. Assessment will be tied to the curriculum, so that it examines what has been taught and practiced and is thereby more representative of meaningful tasks and subject matter goals. Assessment tasks will increasingly provide worthwhile instructional experiences that illustrate the relevance and utility of the knowledge and skill that is acquired and its application to different settings.

Assessments in which students participate in group activity are likely to increase in the future. Performance in a social setting where students contribute to a task and assist others has the advantage of encouraging students to question and develop their definitions of competence. In such assessment, as in instruction using group approaches, the student can observe how others reason and can receive feedback on his or her own efforts. In this context, not only performance, but also the facility with which a student adapts to help and guidance, can be assessed.

Since the forms of instruction envisioned will be oriented toward more complex curricular objectives, assessments will increasingly use various kinds of open-ended questions in which students write about their approach to a problem, the questions that come to mind, and explanations of their

solutions. It will no longer be assumed that a test item is a measure of higher order skills simply because it is more difficult. More and more, the nature of assessment will necessitate analysis of the cognitive aspects of a task and the performances that it entails.

The closer ties between assessment and instruction imply that the nature of the performances to be assessed and the criteria for judging those performances will become more apparent to students and teachers. Knowledge and skills will be measured so that the processes and products of learning are openly displayed. There will be fewer examples of indirect measurement procedures that take advantage of formats for multiple choice or controlled scoring. The performance criteria by which students are judged will be evident so the criteria can motivate and direct the process of learning. Teaching toward the assessment will be the point of instruction rather than a significant difficulty to be guarded against.

As performance criteria become more openly available, students will become better able to judge their own performance without necessary reference to the judgments of others. Instructional and assessment situations will provide coaching and practice in ways that help students reflect on their performances. Occasions for self-assessment will enable students to set incremental standards by which they can judge their own achievement and develop self-direction for attaining higher performance levels.

Our analysis of testing, assessment, and instruction from the perspective of settings of use would be remiss if it ignored the matter of consequences. The intended and unintended effects of an assessment on the ways teachers and students interpret the results, frame educational objectives, and allocate their time warrants serious examination. Newer forms of assessment that lead to negative consequences for teachers and their students must be rejected in the same way that older forms are being rejected. An assessment program must be judged now and in the future in terms of its effectiveness in helping teachers to maximize student learning. If an assessment leads to emphasis on certain topics, teaching materials, and kinds of performance, this consequence must be taken into account in judging the value of the assessment. Evidence must be produced to demonstrate that changes in assessment result in classroom activities that are conducive to improved student learning. In fact, this is likely to be a far more important topic for

future research on the relationship between instruction and assessment than research that leads only to improvements in the technology of alternative assessment.

Coda

Certain practices of yesteryear are dysfunctional in American education today. We now need tests that are not pessimistic about the abilities of low-achieving students and do not assign them to spare and diluted forms of education that constrain their opportunities to learn and, indeed, fail to elaborate the very skills needed for further success in school and in later life. As Lorrie Shepard (1991) has pointed out: "Although intended to be helpful, the practice of assigning poor achievers to special places where they receive bad instruction is analogous to sending debtors to prison in Victorian England" (pp. 292-293). Rather, we need practices that promote learning by offering alternative opportunities.

Assessments of the outcomes of schooling must be designed and used in ways that take account of modern knowledge of human cognition and allow us to develop educational environments in which usable knowledge is achieved by all students and high levels of competence are attained by many. Reaching these goals will be nearly impossible if we continue to carry the ballast of practices that were designed for a time gone by. Developments in the field of cognitive psychology offer educational measurement and evaluation a new perspective on the design of innovative assessments (Lane, 1993; Mislevy, 1993; Snow & Lohman, 1989), the characterization and measurement of skilled performance (Glaser, 1986; Tatsuoka, 1990, 1993), and the variety of technical issues associated with the measurement of complex educational outcomes (Bennett & Ward, 1993; Magone et al., in press; Shepard, 1992). The conceptual and technical groundwork is being laid, but much remains to be done.

There is good reason for optimism that the oft-postponed wedding of assessment and instruction will occur. If teachers can be empowered to use new forms of assessment to improve their teaching, and if they, together with educational policy makers, can devise systemic approaches that integrate assessment into efforts to improve learning and instruction, perhaps the time for change in assessment practice to enhance its usefulness for instructional decision making and the display of standards of competent performance will be at last upon us.

References

Bennett, R. E., & Ward, W. C. (Eds.). (1993). *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Brown, A. L., Campione, J. C., Webber, L. S., & McGilly, K. (1992). Interactive learning environments: A new look at assessment and instruction. In B. R. Gifford & M. C. O'Connor (Eds.), *Changing assessments: Alternative views of aptitude, achievement and instruction* (pp. 121-211). Boston: Kluwer.

Brown, A. L., & Palincsar, A. S. (1984). Reciprocal teaching of comprehension fostering and monitoring activities. *Cognitive Instruction, 1,* 175-177.

Bruer, J. T. (1993). *Schools for thought: A science of learning in the classroom.* Cambridge, MA: MIT Press.

California Mathematics Council. (1989). *Assessment alternatives in mathematics: An overview of assessment techniques that promote learning.* Berkeley, CA: EQUALS.

California State Department of Education. (1989). *A question of thinking: A first look at students' performance on open-ended questions in mathematics.* Sacramento, CA: Author.

Carroll, J. B. (1978). On the theory-practice interface in the measurement of intellectual abilities. In P. Suppes (Ed.), *Impact of research on education: Some case studies* (pp. 1-105). Washington, DC: National Academy of Education.

Cobb, P., Wood, T., & Yackel, E. (1991). A constructivist approach to second grade mathematics. In E. von Glasersfeld (Ed.), *Radical constructivisim in mathematics education* (pp. 157-176). Dordrecht, The Netherlands: Kluwer.

Coleman, J. S., Campbell, E. G., Nelson, C. J., McPartland, J., Mood, A. M., Weinfeld, F. D., & York, R. L. (1966). *Equality of educational opportunity.* Washington, DC: U.S. Government Printing Office.

Cronbach, L. J. (1970). [Review of *On the theory of achievement test items*]. *Psychometrika, 35,* 509-511.

Cronbach, L. J. (1975). Five decades of public controversy. *American Psychologist, 30,* 1-14.

Cronbach, L. J. (1984). *Essentials of psychological testing* (4th ed.). New York: Harper & Row.

Curtis, M. E., & Glaser, R. (1981). Changing conceptions of intelligence. In D. C. Berliner (Ed.), *Review of research in education*, (Vol. 9, pp. 111-148). Washington, DC: American Educational Research Association.

Darling-Hammond, L. (1993). Reframing the school reform agenda: Developing capacity for school transformation. *Phi Delta Kappan*, *74*, 752-761.

Darling-Hammond, L. (1994). Performance-based assessment and educational equity. Symposium on equity and educational testing and assessment (1993, Washington, DC). *Harvard Educational Review*, *64*, 5-30.

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

DuBois, P. H. (1965). A test-dominated society: China, 1115 B. C.–1905 A. D. In C. W. Harris (Ed.), *Proceedings of the 1964 Invitational Conference on Testing Problems* (pp. 3-11). Princeton, NJ: Educational Testing Service.

Engelhart, M. D. (1950). Examinations. In W. S. Monroe (Ed.), *Encyclopedia of educational research* (pp. 407-414). New York: Macmillan.

Frederiksen, J. R., & Collins, A. (1989). A systems approach to educational testing. *Educational Researcher*, *18*(9), 27-32.

Garcia, E. (1991). *Education of linguistically and culturally diverse students: Effective inst uctional practices* (Educational Practice Report 1). Santa Cruz: University of California, National Center for Research on Cultural Diversity and Second Language Learning.

Gardner, H. (1990). The difficulties of school: Probable causes, possible cures. *Daedalus: Journal of the American Academy of Arts and Sciences*, *119*, 85-113.

Gardner, H. (1992). Assessment in context: The alternative to standardized testing. In B. R. Gifford & M. C. O'Connor (Eds.), *Future assessments: Changing views of aptitude, achievement, and instruction* (pp. 77-119). Boston: Kluwer.

Glaser, R. (1972). Individuals and learning: The new aptitudes. *Educational Researcher*, *1*(6), 5-13.

Glaser, R. (1977). *Adaptive education: Individual diversity and learning.* New York: Holt, Rinehart & Winston.

Glaser, R. (1981). The future of testing: A research agenda for cognitive psychology and psychometrics. *American Psychologist*, *36*, 923-936.

Glaser, R. (1986). The integration of instruction and testing. In E. Freeman (Ed.), *The redesign of testing in the 21st century: Proceedings of the 1985 ETS invitational conference* (pp. 45-58). Princeton, NJ: Educational Testing Service.

Glaser, R. (1987). The integration of instruction and testing: Implications from the study of human cognition. In D. C. Berliner & B. V. Rosenshine (Eds.), *Talks to teachers* (pp. 329-341). New York: Random House.

Gould, S. J. (1981). *The mismeasure of man.* New York: Norton.

Kamin, L. (1974). *The science and politics of IQ.* New York: Wiley.

Kilpatrick, J. (1992). A history of research in mathematics education. In D. A. Grouws (Ed.), *Handbook of research on mathematics teaching and learning* (pp. 3-38). New York: Macmillan.

Knapp, M. S., Shields, P. M., & Turnbull, B. J. (1992). *Academic challenge for the children of poverty. Summary report.* Washington, DC: United States Department of Education, Office of Policy and Planning.

Lampert, M. (1986). Knowing, doing, and teaching multiplication. *Cognition and Instruction, 3,* 305-342.

Lane, S. (1993). The conceptual framework for the development of a mathematics assessment instrument for QUASAR. *Educational Measurement: Issues and Practice, 12*(2), 16-23.

Linn, R. L. (1983). Testing and instruction: Links and distinctions. *Journal of Educational Measurement, 20,* 179-189.

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

Livingston, C., Castle, S., & Nations, J. (1989). Testing and curriculum reform: One school's experience. *Educational Leadership, 46*(7), 23-25.

Madaus, G., West, M. M., Harmon, M. C., Lomax, R. G., & Viator, K. A. (1992). *The influence of testing on teaching math and science in Grades 4-12* (NSF Report No. SPA8954759). Chestnut Hill, MA: Boston College, Center for the Study of Testing, Evaluation, and Educational Policy.

Magone, M., Cai, J., Silver, E. A., & Wang, N. (in press). Validating the cognitive complexity and content quality of a mathematics performance assessment. *International Journal of Educational Research.*

Martin, A. (1988). Screening, early intervention, and remediation: Obscuring children's potential. *Harvard Educational Review, 58,* 488-501.

McArthur, D. L. (1987). Educational assessment: A brief history. In D. L. McArthur (Ed.), *Alternative approaches to the assessment of achievement* (pp. 1-20). Boston: Kluwer.

McKnight, C. C., Crosswhite, F. J., Dossey, J. A., Kifer, E., Swafford, J. O., Travers, K. J., & Cooney, T. J. (1987). *The underachieving curriculum: Assessing U.S. school mathematics from an international perspective.* Champaign, IL: Stipes.

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

Mercer, J. R. (1989). Alternative paradigms for assessment in a pluralistic society. In J. A. Banks & C. M. Banks (Eds.), *Multicultural education* (pp. 289-303). Boston: Allyn & Bacon.

Mislevy, R. J. (1993). Foundations of a new test theory. In N. Frederiksen, R. J. Mislevy, & I. Bejar (Eds.), *Test theory for a new generation of tests* (pp. 19-39). Hillsdale, NJ: Lawrence Erlbaum Associates.

Mislevy, R. J., Yamamoto, K., & Anacker, S. (1992). Toward a test theory for assessing student understanding. In R. A. Lesh & S. Lamon (Eds.), *Assessments of authentic performance in school mathematics* (pp. 293-318). Washington, DC: American Association for the Advancement of Science.

Mullis, I. V. S., Owen, E. H., & Phillips, G. W. (1990). *Accelerating academic achievement: A summary of findings from 20 years of NAEP.* Princeton, NJ: Educational Testing Service.

National Academy of Education. (1993). *Setting performance standards for student achievement.* Stanford, CA: Author.

National Commission on Testing and Public Policy. (1990). *From gatekeeper to gateway: Transforming testing in America.* Chestnut Hill, MA: Author.

National Council of Teachers of Mathematics. (1989). *Curriculum and evaluation standards for school mathematics.* Reston, VA: Author.

National Council of Teachers of Mathematics. (1991). *Professional standards for teaching school mathematics.* Reston, VA: Author.

National Institute of Education. (1979). *Testing, teaching and learning: Report of a conference on research on testing, August 17-26, 1979.* Washington, DC: Author.

National Research Council. (1989). *Everybody counts.* Washington, DC: National Academy of Sciences.

National Research Council. (1993). *Measuring up: Prototypes for mathematics assessment.* Washington, DC: National Academy of Sciences.

Nitko, A. J. (1989). Designing tests that are integrated with instruction. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 447-474). New York: Macmillan.

Nunes, T., Schliemann, A. D., & Carraher, D. W. (1993). *Street mathematics and school mathematics.* New York: Cambridge University Press.

Oakes, J. (1985). *Keeping track: How schools structure inequality.* New Haven, CT: Yale University Press.

Oakes, J. (1990). *Multiplying inequalities: The effects of race, social class, and tracking on opportunities to learn mathematics and science.* Santa Monica, CA: RAND.

Oakes, J., & Lipton, M. (1990). Tracking and ability grouping: A structural barrier to access and achievement. In J. Goodlad & P. Keating (Eds.), *Access to knowledge: An agenda for our nation's schools* (pp. 187-204). New York: College Entrance Examination Board.

Phillips, G. W., Mullis, I. V. S., Bourque, M. L., Williams, P., Hambleton, R. K., Owen, E. H., & Barton, P. (1993). *Interpreting NAEP scales.* Washington, DC: U.S. Department of Education.

Pintner, R. (1931). *Intelligence testing: Methods and results.* New York: Holt.

Rankin, P. T. (1931). Pupil classification and grouping. *Review of Educational Research, 1,* 200-300.

Resnick, D. P. (1980). Minimum competency testing historically considered. In D. C. Berliner (Ed.), *Review of research in education* (Vol. 8, pp. 3-29). Washington, DC: American Educational Research Association.

Resnick, D. P. (1982). History of educational testing. In A. K. Wigdor & W. R. Garner (Eds.), *Ability testing: Uses, consequences and controversies, Part II: Documentation section* (pp. 173-194). Washington, DC: National Academy Press.

Resnick, D. P., & Resnick, L. B. (1985). Standards, curriculum, and performance: A historical and comparative perspective. *Educational Researcher, 14*(4), 5-20.

Resnick, L. B. (1987). *Education and learning to think.* Washington, DC: National Academy of Sciences.

Resnick, L. B. (1988). Treating mathematics as an ill-structured discipline. In R. I. Charles & E. A. Silver (Eds.), *Research agenda for mathematics*

education: Vol. 3. The teaching and assessing of mathematical problem solving (pp. 31-60). Hillsdale, NJ: Lawrence Erlbaum Associates.

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

Rosenbaum, J. (1980). Social implications of educational grouping. In D. C. Berliner (Ed.), *Review of research in education* (Vol. 8, pp. 361-401). Washington, DC: American Educational Research Association.

Salmon-Cox, L. (1982). *MAP math: End of year one report.* Pittsburgh, PA: University of Pittsburgh, Learning Research and Development Center.

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.

Shepard, L. A. (1989). Identification of mild handicaps. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 545-572). New York: Macmillan.

Shepard, L. A. (1991). Negative policies for dealing with diversity: When does assessment and diagnosis turn into sorting and segregation? In E. Hiebert (Ed.), *Literacy for a diverse society: Perspective, programs, and policies* (pp. 279-298). New York: Teachers College Press.

Shepard, L. A. (1983). The role of measurement in educational policy: Lessons from the identification of learning disabilities. *Educational Measurement: Issues and Practice*, *2*(3), 4-8.

Shepard, L. A. (1992). What policy makers who mandate tests should know about the new psychology of intellectual ability and learning. In B. R. Gifford & M. C. O'Connor (Eds.), *Changing assessments: Alternative views of aptitude, achievement and instruction* (pp. 301-328). Boston: Kluwer.

Shepard, L. A., & Smith, M. L. (1983). An evaluation of the identification of learning disabled students in Colorado. *Learning Disability Quarterly*, *6*, 115-127.

Shepard, L. A., Smith, M. L., & Vojir, C. P. (1983). Characteristics of pupils identified as learning disabled. *American Educational Research Journal*, *20*, 309-331.

Silver, E. A. (1992, August). *Mathematical thinking and reasoning for all students: Moving from rhetoric to reality.* Paper presented at the Seventh International Congress on Mathematical Education, Quebec, Canada.

Silver, E. A. (1993). *QUASAR project summary*. Unpublished manuscript, University of Pittsburgh, Learning Research and Development Center.

Silver, E. A., & Kenney, P. A. (in press). Sources of assessment information for instructional guidance in mathematics. In T. A. Romberg (Ed.), *Reform in school mathematics and authentic assessment*. Albany: State University of New York Press.

Silver, E. A., & Kilpatrick, J. (1988). Testing mathematical problem solving. In R. I. Charles & E. A. Silver (Eds.), *The teaching and assessing of mathematical problem solving* (pp. 178-186). Hillsdale, NJ: Lawrence Erlbaum Associates.

Silver, E. A., Kilpatrick, J., & Schlesinger, B. (1990). *Thinking through mathematics*. New York: College Entrance Examination Board.

Silver, E. A., & Lane, S. (1993). Assessment in the context of mathematics instruction reform: The design of assessment in the QUASAR project. In M. Niss (Ed.), *Assessment in mathematics education and its effects* (pp. 59-70). London: Kluwer.

Silver, E. A., Shapiro, L. J., & Deutsch, A. (1993). Sense-making and the solution of division problems involving remainders: An examination of students' solution processes and their interpretations of solutions. *Journal for Research in Mathematics Education, 24*, 117-135.

Silver, E. A., Smith, M. S., & Nelson, B. S. (in press). The QUASAR project: Equity concerns meet mathematics education reform in the middle school. In E. Fennema, W. Secada, & L. Byrd (Eds.), *New directions for equity in mathematics education*. New York: Cambridge University Press.

Slavin, R. E. (1990). Ability grouping in secondary schools. *Review of Educational Research, 60*, 471-499.

Smith, M. L. (1991). Put to the test: The effects of external testing on teachers. *Educational Researcher, 20*(5), 8-11.

Snow, R. E., & Lohman, D. F. (1989). Implications of cognitive psychology for educational measurement. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 263-331). New York: Macmillan.

Tatsuoka, K. K. (1990). Toward an integration of item-response theory and cognitive error diagnoses. In N. Frederiksen, R. L. Glaser, A. M. Lesgold, & M. G. Shafto (Eds.), *Diagnostic monitoring of skill and knowledge acquisition* (pp. 453-488). Hillsdale, NJ: Lawrence Erlbaum Associates.

Tatsuoka, K. K. (1993). Item construction and psychometric models appropriate for constructed responses. In R. E. Bennett & W. C. Ward (Eds.), *Construction versus choice in cognitive measurement: Issues in*

constructed response, performance testing, and portfolio assessment (pp. 107-133). Hillsdale, NJ: Lawrence Erlbaum Associates.

Thorndike, E. L. (1904). *An introduction to the theory of mental and social measurement*. New York: Science Press.

Tyack, D. (1974). *The one best system: A history of American urban education*. Cambridge, MA: Harvard University Press.

Wiggins, G. (1989). A true test: Toward more authentic and equitable assessment. *Phi Delta Kappan, 70*, 703-713.

Wolf, T. H. (1973). *Alfred Binet*. Chicago: University of Chicago Press.

Wolf, D., Bixby, J., Glenn, J. III, & 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.