García, G. E., & Pearson, P. D. (1994). Assessment and diversity. In L. Darling-Hammond (Ed.), *Review of research in education* Vol. 20 (pp. 337-392). Washington, DC: American Educational Research Association.

# Chapter 8

# Assessment and Diversity

GEORGIA EARNEST GARCÍA AND P. DAVID PEARSON
University of Illinois at Urbana-Champaign

Systemic reform in public education has undoubtedly been the most common educational goal of the past decade. As this movement toward full-scale overhaul of the structure, content, and process of education has gathered momentum, educational assessment has assumed an increasingly prominent position. The prominence of assessment seems to stem from its potential to fuel the entire reform process (L. B. Resnick & Resnick, 1992; Wiggins, 1989, 1992). The political logic of those who rely on assessment as an instrument of reform seems, ironically, to stem from a deep-seated belief that assessment has been responsible for many of our current educational woes: Since curriculum-narrowing, standardized, multiple-choice tests created the problems from which schools, teachers, and students now suffer (Shepard, 1989), the key to a liberating curriculum must be an assessment system in which broader, more challenging, and more authentic educational values are operationalized and promoted (Simmons & Resnick, 1993).

An even more openly political motivation appears in the public rhetoric about our comparative economic disadvantage in world markets (Brandt, 1992). Politicians and business leaders point to our lack of a centralized educational system with clearly established standards and a national assessment system that evaluates the extent to which students measure up. In their eyes, standards and accompanying assessments would drive everything from classroom instruction to instructional materials to teacher education. Accordingly, the more that key educational milestones can be tied to the attainment of standards, the greater the likelihood that reform will move ahead. This means using standards and assessments as the basis for promotion, high school graduation, college entrance, and entry into the work force. Indeed, important reform movements, such as the New Standards Project (L. B. Resnick, 1989; L. B. Resnick & Resnick, 1992) and several statewide movements (Kentucky Department of Education, 1992; Koretz, Stecher, & Deibert, 1992, 1993), have used just this logic in rationalizing their assessment-driven reform efforts.

A third motivation for assessment reform comes from a synchronically parallel but completely independent source: changing theoretical views of language, learning, and cognition. Constructivist views from cognitive psychology (see

Gardner, 1985), social constructivist views from developmental psychology (see Rogoff & Lave, 1984), and transactional views from literary theory (see Rosenblatt, 1985) have pushed more traditional transmissionist views of knowledge and learning aside in favor of highly constructive views. No longer is knowledge viewed as a commodity to be passed from one entity to another through a process called education; no longer is learning viewed as the accumulation of bodies of facts. Instead, knowledge is viewed as the residual outcome of the process that occurs when learners construct meaningful interpretations of the data that they encounter in their transactions with the world and with other learners. Learning, along with the traditional linguistic concepts of comprehension and composition, is synonymous with constructing meaning. These radical departures from transmissionist views of language and learning have sparked as much interest in developing new forms of assessment as they have in developing new conceptualizations of curriculum and instruction. Traditional measures, both formal (e.g., standardized tests) and informal (e.g., teacher-made tests), have been criticized for being out of touch with these recent developments in theory, research, and practice. In fact, most advocates of alternative assessment practices preface their description of their pet approach with an accounting of the extent to which traditional measures are inconsistent with current theory (e.g., Pearson & Valencia, 1987; L. B. Resnick & Resnick, 1992). The net result is that everyone drawn into the web of constructivist epistemological traditions—psychologists, philosophers, educators, and even measurement experts—is now interested in finding ways to evaluate how students approach, pursue, and interpret meaning construction and problem-solving tasks (see Paris, Lawton, & Turner, 1992; Presseisen, Smey-Richman, & Beyer, 1992).

A fourth motive for focusing on assessment-driven reform is to reduce or eliminate the pernicious influence tests exert on the lives and well-being of students, particularly low-income students (Presseisen et al., 1992). In comparison with teachers of moderate- and high-income students, teachers of low-income students are far more likely to rely on data from commercial tests to develop curricular plans and deliver instruction (Center for the Study of Testing, Evaluation, and Educational Policy, 1992; Dorr-Bremme & Herman, 1986; Rothman, 1992). If the consequences of our current assessment practices result in double jeopardy for low-income students, then we, as a profession, will need to redouble our own efforts to make sure that the consequences of submitting oneself to examination are positive rather than negative.

The net result of these motivations, whether they stem from intellectually pure, socially noble, or crass political origins, has been to focus the attention of the educational research and policy community on educational assessment. Our goal in this chapter is to examine past, current, and future assessments from a multicultural perspective. We will examine the well-established tradition of formal, standardized, multiple-choice assessments as well as some of the newcomers that have gained prominence in the past 5 years, such as authentic classroom assessment and

performance assessment (Wiggins, 1992). We will focus our examination on the impact of assessment on the curricular lives of students who bring diverse cultural, linguistic, and economic backgrounds to our classrooms. This means that matters of ethnicity, race, class, and language will dominate the discourse.

A word about scope: Both of the authors come from a literacy education background. Therefore, even though we have tried to examine assessment and diversity across a range of subject matters and test types, we will undoubtedly exhibit a bias, particularly in our examples, toward language and literacy assessment. We have chosen to concentrate our review on formal assessment and two of the most popular of the newcomers to the assessment scene, authentic classroom assessment and performance assessment. Our definition of formal assessment extends to most of what we would call standardized and commercially available assessments. We mainly deal with achievement tests, but we have included relevant studies about aptitude tests, intelligence tests, and criterion-referenced curriculum tests when they provide insight on problems of cultural, linguistic, or economic diversity. In our scheme, authentic classroom assessment extends to those evaluation tools that are situated in the classroom, designed by the teacher, and used to evaluate student performance within the classroom curriculum context. The common feature to all that we include in performance assessment is either a direct (live observation and judgments about performance) or indirect (judgments about artifacts of performance, for example, in a portfolio) assessment of an on-line performance.

We confess, at the outset, to a dilemma that we faced in conducting the research for this essay. Our review of formal tests differs considerably from our review of newer assessment traditions. In reviewing formal assessment, we were able to rely on a substantial body of research in which issues of student diversity have been addressed, either directly or indirectly. When we turned to authentic classroom assessment and performance assessment, we found that these traditions are so new that little research has been conducted on their efficacy or their impact on any population, let alone students from diverse cultural, linguistic, or economic backgrounds. Thus, our review of these efforts consists primarily of descriptions of their goals and practices, followed by an extrapolation of their potential advantages and disadvantages based on plausible inferences that can be drawn from other research endeavors: sociolinguistic and ethnographic studies of minority populations in school settings, studies of novel assessment practices, and the few existing studies of assessment practices for diverse populations. The logic that prevailed in the section on alternative assessment was as follows: "The paucity of research notwithstanding, if we look at school-based research in which we can find social, political, and curricular relations that mirror those found in authentic classroom and performance assessment, what can we learn about their probable impact on low-income students, students of color, and students for whom English is a second language?"

## FORMAL ASSESSMENT

### The Historical Bias of Formal Assessment

African-American, Latino, and Native American students, as well as students for whom English is a second language, do not, as a group, perform as well as Anglos on formal tests; the data from numerous studies and national assessments consistently confirm these discrepancies (see Applebee, Langer, & Mullis, 1987; Educational Testing Service [ETS], 1988; Mullis & Jenkins, 1990; National Center for Education Statistics, 1988). However, it is important to recognize that the critical factors underlying these findings are poverty and English proficiency (Pennock-Roman, 1992; Rodriguez, 1992). Differences between Anglo students and students of color are substantially reduced when comparisons are limited to students from the same income levels and similar proficiency in standard English. Having established the significance of these moderating factors, it is equally important to point out that they do not eradicate the discrepancy. In particular, Anglo students achieve higher scores than other students on formal tests that reward verbal abilities in standard English and/or knowledge of the Anglo culture (ETS, 1988; Patterson, 1989). Asian Americans, who have established a reputation as high achievers in American society and who score higher on mathematical measures than Anglo students, still score lower on verbal measures (ETS, 1988; Tsang, 1989). Gender differences are as long-standing as ethnic differences. Women, as a group, do not do as well as men on tests that reward mechanical and physical skills (Patterson, 1989) or mathematical, scientific, or technical skills (Moore, 1989).

The history of early test development reveals the roots of the discrimination problem. In 1916, Terman adapted Alfred Binet's Intelligence Test and renamed it the Stanford-Binet Intelligence Test. Binet originally developed his test at the request of the French government to identify those children who could not do the work required in a regular school setting and who would benefit from placement in "schools for the mentally subnormal" (Mercer, 1989). Terman translated Binet's test into English, changed the questions to reflect American values and knowledge, normed the test on middle-class Anglo children, and standardized the scores by age (Mercer, 1989). When girls outscored boys on the 1916 version of the test, the test designers, apparently operating under the assumption that girls could not be more intelligent than boys, concluded that the test had serious faults. When they revised the 1937 version, they eliminated those items on which girls outperformed boys. By contrast, they did not revise or eliminate items that favored urban over rural children or children of professional fathers over children of day laborers (Mercer, 1989); these cultural differences apparently matched the developers' expectations of how intelligence and achievement ought to be distributed across groups (Kamin, 1974; Karier, 1973a, 1973b; Mercer, 1989). Although the Stanford-Binet and its numerous sibling tests—the Group Test of Mental Ability, the Stanford Achievement Tests, the National Intelligence Tests, and the

Army Alpha Test—were hailed for ushering in an era of "objective" and "efficient" means of assessing students' intelligence and achievement (Moore, 1989), their historical development clearly reflected the racism, xenophobia, classism, and sexism that prevailed during this era (Chachkin, 1989; Karier, 1973a, 1973b; Mercer, 1989). Karier (1973b) points out that even as late as the 1960s, an item on the Stanford-Binet test asked students to select the drawing that was "prettier" when presented a pair of female drawings, one clearly Nordic/Anglo and the second more Mexican-American/southern European.

## Characteristics of Formal Tests

In a recent U.S. General Accounting Office report (1993) of a wide-scale survey of test use and attitudes, school districts reported that more than 80% of all systemwide tests administered were achievement tests. Despite the rhetoric of the reform movements, 71% of all tests used were classified as norm referenced and multiple choice. And 43% of all tests referred to for any testing category (e.g., achievement, aptitude, curriculum referenced, state accountability assessment) were developed by one of the three major testing companies. Thus, the standard in educational testing is a norm-referenced, multiple-choice test that samples performance or achievement in a well-defined curricular domain.

For example, in the field of reading, current versions of commercial reading achievement tests tend to include a range of short passages that represent a variety of topics and genres (narrative and expository text, poems, advertisements, letters to the editors). Students' reading achievement is measured by summing their answers to a series of multiple-choice questions, each with a single correct answer, following these short passages. On reading achievement tests, time limits are usually set after extensive field tests; the developers often use a criterion of allowing sufficient time for some desired percentage (e.g., 95%) of the sample to complete all of the test items. Other tests (e.g., the mathematical sections of aptitude tests) use time as an intentional discriminating factor. They sequence the items in increasing order of difficulty and set tight time limits so that speed as well as accuracy enters into the assessment of ability or achievement.

## Problems Arising From Test Development

### *Theoretical Inadequacy*

Standardized tests of all sorts have always had their critics (Karier, 1973a, 1973b; Mercer, 1989). Reading tests have been criticized by reading researchers on a number of counts. For at least the last decade, they have been taken to task for failing to reflect current reading theory and research (García & Pearson, 1991b; Johnston, 1984a, 1984b; Pearson & Valencia, 1987; Royer & Cunningham, 1981). A long-standing problem with survey tests is that while they are quite good at indicating a student's relative standing in a group, they provide few, if any, clues about the locus of the performance: Was a low performance due to

limited prior knowledge about the topics, difficulty in reasoning, or difficulty in decoding? In addition, analyses of students' test answers reveal that their answer selections do not always reflect the extent to which they have comprehended the text (Cicourel, 1974; García, 1991; Langer, 1987). Finally, tests have been criticized because they simulate reading but do not indicate how well students approach, process, and interpret text in the pursuit of authentic uses of literacy (Edelsky & Harman, 1988).

In response to criticisms about the theoretical underpinnings of traditional reading tests, many states, some commercial companies, and even the National Assessment for Educational Progress (NAEP) have recently changed the format and content of their tests. For example, many of the new statewide reading tests (e.g., those in Illinois and Michigan) include assessment of students' knowledge of the topics, questions based on inferencing and text structure taxonomies, and evaluation of students' awareness of reading strategies. They also present the students with longer, less-contrived passages and require that the students choose more than one correct answer (Pearson & Valencia, 1987; Wixson, Peters, Weber, & Roeber, 1987). On the 1992 NAEP state-by-state trial assessment (Valencia, Hiebert, & Kapinus, 1992), one of the options in the reading assessment involved an entirely new format. Instead of reading passages and answering multiple-choice short answer or extended response questions, students in one 12th-grade block were presented with a "NAEP reader" containing several short stories. After consulting a preview of all the stories, students chose one story, read it, and responded to a set of generic, open-ended response probes.

The change from a range of short passages to two or three longer passages presents an interesting but frustrating dilemma. On one hand, the longer passages provide all students with a more authentic context for reading and responding to questions. On the other hand, the sheer length of passages means that any inferences drawn about an individual student's achievement will be based on a drastically reduced range of passages. This practice of selecting naturally occurring passages may prove particularly problematic for low-achieving students. Because these state tests are targeted for particular grade levels, the passages tend to come from grade-appropriate sources. Unlike standardized tests, in which a fourth grader is likely to be exposed to many short passages that represent a range of difficulty (e.g., a sample of materials taken from second- to sixth-grade sources), genre, and topic, a fourth grader taking one of these new tests is likely to face two fairly challenging fourth-grade passages. Many students will simply not be able to decode the material. Coupled with format changes, some of which require that students determine how many answer options are correct or write extended responses, the frustration level of many students may be increased, resulting in low or random performance (García & Pearson, 1991b). In fact, Hiebert and Corley (1993), in comparing performance across several types of testing formats, found that the multiple correct option format proved to be the

most difficult—in terms of the absolute number of correct responses—for third-grade students at all levels of achievement.

Reading tests are not unique in attracting criticism. In a recent analysis of the latest standardized tests and textbook tests in mathematics and science, Madaus and his colleagues reported that there was an undue emphasis on low-level thinking and knowledge (Center for the Study of Testing, Evaluation, and Educational Policy, 1992; Rothman, 1992). Few of the tests challenged students with problem-solving tasks. More than 75% of the science items tested students' recall of facts and routine applications, while less than 10% evaluated students' knowledge of scientific procedures. The reviewers were disappointed to find that none of the mathematics tests surveyed conformed to the recently published standards of the National Council of Teachers of Mathematics (NCTM). Ginsburg and Allardice (1984) also point out that, similar to reading tests, standardized tests of mathematics reveal little about students' mathematical thinking. As a result, they do not give teachers the information they need to help students improve their mathematical thinking and problem solving.

### *The Mainstream Bias of Formal Testing*

*Norming bias.* The norming process, by its nature, leans toward the mainstream culture. For example, when test companies draw strict probability samples of the nation, very small numbers of particular minorities are likely to be included, increasing the likelihood that minority group samples will be unrepresentative. The mainstream bias in test development cuts even deeper than the norming process. When items are piloted, one of the statistics commonly computed is the correlation of each item with the total test score. To create a final test, those items that have the lowest correlation with the total test score are eliminated on the grounds that they provide a poor estimate of the phenomenon being measured. In other words, those very items on which low-scoring students do comparatively well disappear! If we remember that low-income and ethnic minority students are overrepresented in the set of low-scoring students, then it is almost inevitable that minority students will perform relatively poorly on final versions of tests built through this process. Ironically, this procedure is derived from item response theory, the current psychometric tool used to determine whether items are culturally biased. Even if a test is criterion referenced instead of norm referenced, the performance standards (cutoff scores) by which the students' performances are evaluated are likely to be based on professional judgments about what typical (read mainstream) students know and can do at a particular developmental level.

Test developers have used a variety of techniques to create unbiased tests (Cole & Moss, 1989; Linn, 1983; Oakland & Matuszek, 1977). Among others, they have examined item selection procedures, examiner characteristics, and language used on the tests as possible sources of bias. One of the most common methods used to control for test bias is that of examining the concurrent or predictive validity of individual tests for different groups through correlational

or regression analysis. A test is considered biased when it over- or underpredicts the performance of particular groups in relation to the performance of the mainstream group on some criterion measure, such as another test score, grade point average, or a job-related performance (O'Connor, 1989). In addition, psychometricians involved in test development have worked with educational psychologists and others to develop standards for educational and psychological testing (American Psychological Association, American Educational Research Association, and National Council on Measurement in Education, 1985) that specifically address issues of test bias and misuse. One of their standards states that the reporting of test scores should be "restricted to samples from which they are derived" and that they should be "generalizable [only] to populations the samples adequately represent" (p. xxvii).

In response to the norming bias inherent in using wide-scale tests with culturally, linguistically, and economically diverse groups, several psychologists have proposed that students' test scores be evaluated according to the norms developed for specific groups and populations. Mercer (1979) developed the System of Multicultural Pluralistic Assessment (SOMPA) in an attempt to find a culturally fair way to assess the performance of low-income students (see Samuda, 1975). Her system uses the Wechsler Intelligence Scale for Children—Revised to evaluate what children have learned "about the dominant Anglo core culture" (Mercer, 1989, p. 296). However, the students' scores are standardized according to their age and sociocultural background. The latter is determined by asking parents 25 questions that previously have been found to differentiate low-income African-American and Mexican-American students from their middle- and upper-class counterparts (Mercer, 1989; Samuda, 1975). Norms, based on probability samples of students from California, are available for African-American, Anglo, and Latino students. Critics of this approach do not think that separate norms are necessary because they are costly, do not provide for comparisons with the general population, and fail to take into account the complexity of cultural identity (Mercer, 1989; Samuda, 1975).

*Content bias.* According to Tyler and White (1979), content bias occurs when test content and procedures reflect "the dominant culture's standards of language function and shared knowledge and behavior" (p. 3). It is most severe when test tasks, topics, and vocabulary reflect the culture of mainstream society to such an extent that it is difficult to do well on a formal test without being culturally assimilated (Padilla, 1979; Troike, 1984).

Samuda (1975) points out that the original developers of intelligence and aptitude tests were not particularly concerned about issues of test bias because they "believed that it was possible to assess intelligence independently of environment and that measures of IQ were true expressions of intellectual potential" (p. 63). However, the high correlation between students' socioeconomic status and their IQ test performance caused psychologists and educators to raise questions

about the role of environment in intelligence and whether the tests were measuring acquired knowledge and skills not related to intelligence.

In an attempt to develop more equitable ways of accounting for cultural differences, test developers initiated several reforms. In the earliest responses to this problem, psychologists tried to develop intelligence tests that were either "culture free" or "culture fair." Culture-free tests (such as Catell and Catell's Culture-Free Intelligence Test [later renamed IPAT Culture-Fair Intelligence Test], 1941–1963) used nonverbal tasks considered to be culture free to measure students' intelligence. Developers of culture-fair tests took a different tact, questioning whether nonverbal tasks could, indeed, be culture free (see Eells, 1951). Both Eells (1953) and Davis (1951) thought that a culture-fair test (such as their Davis-Eells Test of General Intelligence or Problem-Solving Activity [Davis & Eells, 1953]) ought to include items from material equally familiar to everyone taking the test. In addition, the formal features of the test, including item format, language, and symbols, should be common and equally motivating to all groups. Davis (1951), in opting for nonverbal problem-solving tasks, noted that "linguistic skills are not in themselves a crucial index of mental status" and that intelligence represents "a great range of systems of mental behavior" (p. 23).

Both types of tests were severely critiqued. Reviewers of the culture-fair tests questioned whether it was possible to select tasks that were equally familiar to all cultures (Samuda, 1975). Also, neither the culture-free nor the culture-fair tests attained their major goal of reducing the correlation between socioeconomic status and IQ performance (Samuda, 1975). Moreover, Williams (1974) reported that both types of tests exhibited lower predictive validity than the conventional tests they were trying to replace.

A third movement went in exactly the opposite direction. Instead of trying to ignore cultural differences, it acknowledged them openly and tried to capitalize on them. Most prominent was the development of a culturally specific test for African-American adolescents and adults: the Black Intelligence Test of Cultural Homogeneity (Williams, 1975). Developers of culturally specific tests contended that it was impossible to separate culture from intelligence and that students from lower socioeconomic groups were as intelligent as those from higher socioeconomic groups. The major problem, they observed, was that the tests used by the educational community were based on knowledge and skills unique to middle- and upper-class Anglo culture (see Samuda, 1975; Williams, 1975). According to Williams (1974), the rationale for the culture-specific intelligence test was that "if a child can learn certain familiar relationships in his own culture, he can master similar concepts in the school curriculum, so long as the curriculum is related to his background experiences" (p. 37).

The Black Intelligence Test of Cultural Homogeneity (Williams, 1975) did make a difference; in fact, it completely reversed the traditional pattern of Anglos outperforming African Americans. Critics of the culture-specific concept raised questions about the tests' generalizability, the feasibility of standardizing the

tests, and the difficulty of trying to define who belongs in a specific cultural group (Eells, 1951; Samuda, 1975).

Bias also stems from practices that implicitly define success according to values and criteria of mainstream society (e.g., canon, language skills, and strategies). When the cultural values reflected on a test are not a necessary part of the competency being assessed, content bias is clear. Perhaps one of the most blatant examples of content bias occurs when young children's oral language development is evaluated by presenting them with vocabulary and tasks that represent mainstream culture instead of their own culture. Although Labov (1969) and other linguists successfully disputed Bereiter and Englemann's (1966) contention that low-income children are verbally deprived, oral language assessment based on mainstream vocabulary, values, and the production of standard English still occurs (Bruce, Rubin, Starr, & Liebling, 1984; Stallman & Pearson, 1990).

Test developers have acknowledged that they must maintain constant vigilance to prevent content bias from finding its way into standardized reading tests. According to Johnston (1984b), developers of these tests have attempted to control for differences in students' knowledge of topics in several ways. First, they have ensured that students will encounter a variety of topics; the logic seems to be that, other things being equal, a wide range of topics will, on average, render all individuals equally advantaged or disadvantaged. Sensitive to the problem of passage-independent questions (questions that can be answered without reading the passage), developers have attempted to eliminate them through careful piloting. Finally, through latent trait theory based on population-level differences, they have tried to statistically control the influence of prior knowledge. Despite these efforts, knowledge differences persist. For example, García (1991) found that Spanish-speaking Latino students scored significantly lower than Anglo students on a test measuring knowledge of the topics of standardized test passages. When topical knowledge was controlled statistically, Latino-Anglo comprehension differences disappeared. Even more important, on passages for which Latino knowledge exceeded (e.g., a passage about the piñata custom) or equaled (e.g., history of the newspaper) Anglo knowledge, there were no differences in comprehension.

As we indicated earlier, some state reform movements have tried to control topical bias by measuring it as a part of the assessment process, presumably with an eye toward interpreting comprehension scores in light of topical knowledge scores (Pearson & Valencia, 1987; Wixson et al., 1987). These positive actions notwithstanding, there is no guarantee that such reform movements will help students from culturally and economically diverse backgrounds: Simply assessing prior knowledge, while it might help to explain performance, does not guarantee that all students will be given the opportunity to read personally or culturally familiar topics (García, 1991).

While tests cannot cover topics that privilege the experiences of all students, it could be argued that any topical or informational bias present in the early

grades should diminish in the later grades as students become socialized into the culture of school and gain greater control over the content of school subjects. Even under the tenuous assumption that cultural assimilation can or should occur, the lesson that young students may learn from early alienating experiences is that they are not allowed to enter the cultural conversation of the classroom (García, Pearson, & Jiménez, in press). The extent to which these early lessons influence participation in school and/or later academic development has not been widely studied. Nonetheless, the available studies of home-school sociocultural and sociolinguistic discontinuities suggest that differences in cultural knowledge and interactional styles do affect students' academic achievement (Au & Jordan, 1982; Delgado-Gaitan, 1987; Heath, 1981, 1982; Michaels, 1981; Philips, 1972). We also know that student placement in compensatory programs based on such tests can lead to differential educational experiences that are not necessarily beneficial (McGill-Franzen, 1987).

*Linguistic and cultural biases.* Several factors adversely affect the formal test performance of students from diverse linguistic and cultural backgrounds, among them "speededness" (the inability of students to complete all of the items included on a test as a result of prescribed time limitations) (Mestre, 1984; Rincón, 1980), test anxiety and testwiseness (García, 1991; Rincón, 1980; Tyler & White, 1979), and the differential interpretation of questions and foils (García, 1991) or even the testing event itself (Cicourel, 1974; Taylor, 1977). Second-language speakers of English may have difficulty with English vocabulary (Durán, 1983; García, 1991; Pennock-Román, 1990; Sanchez, 1934). In addition, bilingual educators have warned that it is difficult to determine the language in which a bilingual student should be tested (Seidner, 1982; Troike, 1982), just as it is almost impossible for a formal test to capture what bilingual students know in their two languages (García, 1992b).

Speededness has proven particularly problematic for bilingual students. Mestre (1984) found that speededness adversely influenced the mathematical test performance of Latino college engineering students, who stated that they were unable to complete the problem-solving test because they spent "so much time trying to understand what [was] being asked" (p. 19). Rincón (1980) found a similar disadvantage for Latino high school students on the School and College Ability Tests. García (1991) reported that the systematic, methodical approach used by Latino students throughout the earlier part of a reading test had to be abandoned in favor of a more hurried approach as they sensed time was running out. She noted that the speededness effect is consistent with bilingual research that has demonstrated that bilinguals (a) take longer to process either of the two languages than monolinguals, (b) read more slowly in their second language, and (c) develop receptive competencies (decoding or comprehension) in their second language more rapidly than productive competencies (encoding or writing and speaking) (Chamot, 1980; Eaton, 1980; Mägiste, 1979). Because time limits are set during pilots that use predominantly monolingual samples, many bilingual students are

likely to be deprived of the time they need to complete the examination at a pace comparable to that of monolingual students.

Unfamiliar English vocabulary causes difficulty, particularly on multiple-choice tests (Durán, 1983; García, 1991; Pennock-Román, 1990; Sanchez, 1934). The problem is especially acute when knowledge of uncommon English vocabulary is essential for understanding a passage, an item, or test instructions. Even reasonably fluent second-language students may misinterpret English vocabulary as a result of polysemy, nuance, or connotation (Clarke, 1979; Cziko, 1978; Perkins, 1983). The polysemy problem is exacerbated by the policy of paraphrasing: Item writers typically paraphrase text language when it appears in item stems and foils. The logic behind this practice is compelling; without paraphrasing, test takers might be able to compare foils with the text in a mindless attempt to "grab" at answers. García (1991) found that this practice hinders the test performance of second-language students who may know a common English word (e.g., dog) but not less familiar synonyms (canine or mongrel), which frequently are found in the foils. Thus, second-language students may know the correct answer to a question but not be able to find it in the set of foils when the English word they are looking for has been replaced by another, often less common synonym or when a phrase has been substantially paraphrased (e.g., *natural environment* for *free state*) (García, 1991).

Determining whether bilingual students should be tested in English or their native language is difficult (Figueroa, 1989; Geisinger, 1992; Seidner, 1982). Part of the problem is that formal language tests do not capture the variety of ways in which such students acquire and use their two languages (Cummins, 1984; García, 1992b; Savignon, 1983; Troike, 1982). Translating the test from one language to the other also does not solve the problem (Cabello, 1984; Olmeda, 1981). For example, Olmeda (1981) explains that

> direct translations do not yield technically equivalent forms because the domains sampled by the different language versions may have little overlap, . . . the translated items may exhibit psychometric properties substantially different from those of the original English items . . . [and] the interpretation of scores remains difficult . . . because the test content remains culture-bound. (p. 1083)

Determining language dominance is important because the verbal abilities of Spanish-speaking bilingual students tend to be underestimated on English verbal aptitude and reading tests (García, 1991; Pennock-Román, 1992). For example, in comparing College Board data for bilingual Puerto Rican and Mexican-American students, Pennock-Román (1988) found that the mean Scholastic Aptitude Test (SAT) verbal scores for students who learned English first in the home were 50 to 80 points higher than for those who learned Spanish first. On average, they also had an advantage of 20 to 30 points on quantitative SAT scores.

There are several plausible explanations for the insensitivity of English language tests in assessing bilingual students' content and strategic knowledge. First,

a test in one language cannot document the students' knowledge across both languages (García, 1992b). For example, a bilingual student who has learned about U.S. history in English but biology in Spanish may not be able to demonstrate equivalent knowledge of either subject in the other language. Second, bilingual students (including Latinos, Navajo, Samoans, and Arabs) frequently demonstrate greater understanding of an English text when they are allowed to use their native language in the assessment task (Chamot, 1980; Eaton, 1980; García, 1991). This finding works in both directions; for example, Lee (1986) found that English-speaking undergraduates who were learning Spanish also were able to demonstrate longer and more accurate recall of Spanish text when they were allowed to respond in English.

## Problems Arising From Test Use

Some would claim that the problems of bias in test development pale in comparison with the problems of bias that arise from the way tests are used to make decisions about students, teachers, schools, and districts. In the domain of test use, issues arise related to validity and using test results.

### *Consequential Validity*

Recently, measurement scholars (e.g., Linn, Baker, & Dunbar, 1991; Messick, 1989) have begun to include discussions of consequential validity in treatments of test characteristics. Consequential validity entails evaluating the effect of the test on the lives of students, both in and out of school. Many educators oppose formal tests, particularly for students from diverse cultural, linguistic, and economic backgrounds, because the very consequences of their use can be detrimental. This is especially true when the tests are used for gatekeeping purposes, such as placement in special programs and particularly schools or universities and entrance into the workplace (especially into highly desired and economically rewarding employment slots).

*Consequential validity for students of color.* To say that tests will not be used to impede the progress of students of color is to deny the history of their use. Historically, test scores were used to keep African-American and Latino students in segregated schools (Chachkin, 1989). More recently, excessive reliance on test scores for placement purposes has sent disproportionate numbers of minority students into special education programs and low tracks in middle and high school (Chachkin, 1989; García et al., 1989; Rebell, 1989). On the other side of the scale is at least one example in which high performance by a minority group on a placement test has led to the creation of supplemental predictive indices. Asian-American students have earned very high quantitative scores on the SAT, especially the nonverbal section, leading to their overrepresentation in admissions pools for elite public universities. Suddenly new admissions criteria (e.g., writing samples) have been added, some would argue, in order to prevent their overrepresentation (Tsang, 1989).

The extent to which predictive equations based on population-level performance data are relevant for students from diverse cultural and linguistic backgrounds continues to be controversial. In terms of making predictive decisions for students, such as who should be accepted into a particular college or who should be allowed into a special program, the key choice is between a single prediction equation for all students and separate equations for different populations. Geisinger (1992) notes that few researchers who have compared regression lines across mainstream and diverse cultural and linguistic groups in the United States have discovered bias. And, in fact, reviews of such studies on Latino students by both Durán (1983) and Pennock-Román (1990) indicate that while correlations between SAT scores and freshman college grades sometimes were lower for Latino students than for Anglo students (Durán, 1983), there were no significant differences in regression equations. In some cases, the Latino students' first-year college performance was slightly overpredicted (Pennock-Román, 1990).

On the other hand, Pennock-Román (1992) points out that if speededness and language background are key factors that affect "the precision with which abilities are estimated" in Latino students and other linguistic minority groups, "then the accuracy of performance will be lower for these groups in comparison" with other Anglo students (pp. 113–114). Durán (1983) also argues that predictions based on regression equations are not always useful for Latino students because they fail to take into account factors (e.g., socioeconomic status, school adjustment prior to college, migration-immigration history, and language background) that may influence both the students' performance on the tests and their performance in college. However, simply augmenting the prediction equations with relevant background variables does not always lead to more understandable findings. For example, Pennock-Román (1992) reports that efforts to account for immigrant status in conducting correlational analyses of Latino students' test scores have been inconclusive because they could not capture "the complex relationship between immigration history" and Spanish language maintenance that characterizes certain regions of the United States (p. 109).

*Consequential validity in the workplace.* Ironically, assessment in the workplace has prompted more action related to consequential validity than has assessment in school settings. Gifford (1989) described one of the most celebrated cases related to predictive validity. As a result of an error in the norming procedures on the Armed Services Vocational Aptitude Battery (ASVAB), 25% of the recruits during the 1976–1980 period were accepted into the armed forces without the requisite test performance. Because the norming procedures inflated their low test scores, many of them also were accepted into specialized training programs that required "high" test scores. Interestingly, their overall failure rate was not significantly higher than that of the "qualified" enlistees.

The ASVAB example points out another problem inherent in using regression analysis to determine test bias: The test may not evaluate what is needed to

perform well in a particular context. For example, the "unqualified" enlistees might have performed better on the job than on the ASVAB as a result of differences in "incentives, sanctions, motivations, and pressures" (Gifford, 1989, p. 25). Gifford speculates that success in the military requires individuals who can work collaboratively and who have relatively good social skills, neither of which is measured on the ASVAB. Even more important, motivation to succeed in the military may be much higher than motivation to do well on a test. Somehow a paper-and-pencil task seems inadequate to the task of determining whether individuals are motivated to follow orders and work together in order to survive. Also, a new criterion, one that had never been a part of the placement test, was added to job performance ratings: race relations. This last criterion, coupled with a serendipitously misnormed test, opened the door for African-American recruits, who then took advantage of the opportunity to demonstrate that they could succeed on the outcomes valued within the program.

Corporations have adopted several interesting practices to avoid violating government regulations related to discrimination in employment testing. For example, several significant court rulings have concluded that tests cannot be used when they adversely affect minorities and are not "sufficiently job related" (Patterson, 1989). To avoid litigation, test companies frequently examine scores for employment screening and/or promotion on the basis of a normal distribution of test takers within relevant groups (race, sex, or ethnicity) instead of a normal distribution for the entire sample (Schwartz, 1989). Under the assumption that real intelligence and achievement (as opposed to measured intelligence and achievement) are normally distributed in all groups, such a practice can compensate for the otherwise discriminatory effect of a biased test. To date, education appears unwilling to make such an assumption in using tests to screen students for admissions or entry into special programs in spite of the availability of cross-cultural tests (e.g., Mercer's, 1979, SOMPA test).

Another litigation-avoidance practice used by testing companies is to lower the originally established cutoff scores on cognitive ability tests so that fewer minorities are ineligible. While this practice is cheaper than developing norms for different groups, it calls into question the validity of a test for all groups (Patterson, 1989).

Fear of litigation has forced some test companies to abide by what has come to be called the "golden rule" (named after the insurance company in the lawsuit, not the biblical allegory): Items on which minorities do not do well are excluded on a first cut and included later only as a "last resort" (see Chachkin, 1989; Rebell, 1989). This practice, because it eliminates items, can compromise the original curricular design of the test. The psychometric community has reacted strongly to the Golden Rule ruling and practice (Rebell, 1989).

It is interesting, and somewhat ironic, to note that the arena of employment assessment rather than educational assessment has led to more widespread use of innovative practices to compensate for the predictive bias of employment tests.

Just why, as a society, we have been more willing to question the validity of employment tests than that of educational tests, even though they rely on the same development process and are produced by the same companies, is not clear; a few scholars have cited the higher stakes associated with employment tests as well as the greater voice accorded to adults relative to children.

### *Misusing Test Results*

*Blatant misuse.* An example from Georgia illustrates a classic case of test misuse. In its zeal to jump on the outcomes-based education bandwagon, a district used grade-equivalent scores on the California Achievement Test to determine students' eligibility for a high school diploma (they had to reach a grade-equivalent score of 9.0), not taking into account other, more traditional, criteria of accomplishment such as grades and course patterns (Chachkin, 1989). This use of a test clearly violates standards published by the testing community (American Psychological Association, American Educational Research Association, and National Council on Measurement in Education, 1985). The school district failed its students by violating two important assessment principles: (a) The district relied on a single measure to make a decision that had major influence on the lives of individual students, and (b) it completely misused norm-referenced scores. Under the assumption that students in the district conformed to the national population, some sizable percentage of them were doomed to fail the graduation criterion. Whether they were performing well or poorly on the skills and knowledge taught in the school could not be determined because there is no good way of establishing the implicit curriculum standard represented by a grade-equivalent score of 9.0.

*English-only assessment.* Bilingual educators warn that we still do not know the optimal stage of second-language development at which to begin testing second-language students only in English (Figueroa, 1989). Figueroa points out that the consistent discrepancy between low verbal IQ scores and high performance IQ scores of all bilingual populations, including Asian Americans, underscores the need to regard their English language achievement scores with great caution; to place too much stock in such scores is to open the door to faulty inferences, misunderstandings, and decisions with negative consequences. Yet, policymakers blithely mandate English-only assessment. In Illinois, for example, a state mandate requires that third-grade bilingual students be tested in English on statewide tests within 1 year of entry into a bilingual education program. The very act of testing bilingual students in English before they have had an opportunity to acquire competence undermines the purpose of bilingual education. Teachers should have the curricular opportunity to nurture cognitive development and subject-matter knowledge in students' native language as they acquire independence in English. If this mandate prevails, teachers will be driven to adopt the practice of early, and inevitably premature, transfer from native-language

instruction to all-English instruction in order to improve scores on the mandated test.

*Remedial placement.* The courts have intervened to offset the adverse impact of using test scores to place students of color in remedial programs. For example, in *Larry P. v. Riles, Superintendent of Public Instruction for the State of California* (as cited in Chachkin, 1989), the judge decreed that IQ tests (even if individually administered) were not to be used in the placement of African-American students in special education programs. The judge concluded that an overreliance on IQ scores for placement purposes had resulted in a disproportionate number of African-American students being placed in classes for the educable mentally retarded. Although experts for the defendants stated that the IQ tests were not biased but reflected differences in socioeconomic status and culture, the court noted that the tests were normed on White, middle-class sample groups and had not been validated for use with African-American children (Chachkin, 1989).

Similarly, the disproportionate placement of Latino students in special education programs resulted in a 1974 out-of-court settlement in California (*Diana v. California State Board of Education,* 1974, as cited in Figueroa, 1989). The terms of the settlement stipulated that bilingual students had to be tested in their own language as well as in English and that nonverbal IQ test scores could be substituted for full-scale IQ test scores in assessing Latino students with limited English proficiency (Figueroa, 1989). The first portion of this settlement is now part of the Education for All Handicapped Children Act of 1975 (Public Law 94–142), which states that "such materials [tests] or procedures shall be provided and administered in the child's native language . . . unless it clearly is not feasible to do so" (section 612(5)(c)).

Although the courts have actively constrained the use of test scores to place minority students in remedial programs, they have not actively constrained the use of the same or similar tests to keep minority students from being placed in gifted programs or college-bound high school tracks. The rulings for gifted placement are at odds with those in job placement, in which the courts have ruled that separate prediction equations and/or lower cut scores must be used to counteract employment discrimination. Chachkin (1989) points out that, in the educational field, the courts have based their decisions on test relevance (curricular validity) and reliability rather than predictive or consequential validity. The courts generally have ruled that tests used for diplomas must reflect the knowledge and skills that the students have been taught. According to Chachkin, these curricular validity analyses have, for the most part, operated at the surface level; for example, they have not involved item analyses to determine the extent, depth, and detail of the match between the test and the curriculum.

*Disproportionate curricular influence.* Tests have assumed increasing influence in shaping school curricula (Shepard, 1989). For example, Smith (1991) found that teachers are very sensitive to the publication of test scores. They are willing to alter their curriculum to avoid low scores on a test they do not believe in,

even though the practices they engage in to raise test scores result in personal feelings of "dissonance and alienation" and "guilt" about the harm they feel they are inflicting on children. The irony is that teaching to the high-stakes test renders the test invalid (Haladyna, Nolan, & Haas, 1991) through a phenomenon labeled test score pollution. Test score pollution occurs when a score on a test rises or falls without a concomitant change in the underlying construct that is supposedly measured by the test. The validity of most multiple-choice tests rests on the assumption that no one ever teaches to them directly; tests that serve as perfectly reasonable "barometers" of achievement for a certain construct crumble when they are required to serve as a blueprint for a curriculum.

The problem of undue curricular influence is more severe for low-income students. Surveys of test use in schools reveal that teachers who teach low-income students tend to be held more accountable (or at least they feel that they are more accountable) to tests (Center for the Study of Testing, Evaluation, and Educational Policy, 1992; Dorr-Bremme & Herman, 1986; Rothman, 1992). In an updated survey of standardized test use in schools, Herman and Golan (n.d.) found that teachers in classrooms with a majority of Chapter 1 students reported

> more emphasis on testing, less school attention to broader instructional renewal, more adjustments made to instructional planning to incorporate aspects of the test, more classroom time spent on test preparation activities, and less classroom time spent on non-tested subjects and skills. (p. 2)

Since most of the tests used in these programs focus on discrete skills, low-income students tend to receive a fragmented, skills-based curriculum.

The curricular influence of standardized tests on the instruction of low-income students is further compounded by the role that such tests play in Chapter 1, a federal compensatory program that provides low-income students with additional instruction in reading and/or mathematics. Although proposals for reform are under consideration, federal eligibility and evaluation guidelines require districts to use nationally normed test scores to help determine students' eligibility (their academic performance must be below grade-level criteria) and to demonstrate performance gains (U.S. General Accounting Office, 1993). The diagnostic-prescriptive model that tends to underlie many of the Chapter 1 reading programs also results in the additional use of formal tests (Madden & Slavin, 1987; Slavin, Karweit, & Madden, 1992), usually from the criterion-referenced tradition. In addition to the standardized tests used in the fall and spring semesters to demonstrate performance gains, many of the programs administer a variety of formal reading measures throughout the year in order to identify specific areas of difficulty on which students need to work. Madden and Slavin complain that the consequence of the diagnostic-prescriptive model is twofold: Instructional time is lost to testing, and instruction, when it does occur, is tied to the tasks on the formal reading measures. As a result, few students receive instruction that helps them to improve their comprehension of connected text (Allington, Stuetzel, Shake, & Lamarche, 1986).

When "high stakes" are tied to the use of standardized tests, the curricular influence of such tests in low-income settings is exacerbated. A survey of 2,200 mathematics and science teachers, augmented by intensive visits to six urban sites, revealed that teachers in low-income settings were the most likely to teach to the test (Center for the Study of Testing, Evaluation, and Educational Policy, 1992; Rothman, 1992). One of the teachers, a fifth-grade teacher in an inner-city school, explained that she had been using the mathematics "curriculum guide to identify objectives in order to teach to the test" because a certain percentage of students in each district school had to attain a cutoff score on the standardized test or the district would be taken over by the state (Center for the Study of Testing, Evaluation, and Educational Policy, 1992, Vol. 1, p. 1). The consequence was that she had "little time to bring in things—connect things" (p. 1). Instead, she followed the textbook 95% of the time because it fit the curricular guide and the items covered on the state-required standardized test. Sadly, a careful analysis of this test, several other standardized tests, and a sample of textbook tests revealed little attention to the types of conceptual knowledge and problem-solving abilities advocated by reformers (Rothman, 1992).

*Avoiding assessment.* In recent years, as the high-stakes assessment craze has gathered momentum and schools and districts have become concerned about their comparative performance on mandated tests, a new testing abuse has arisen. One of the few ways of excluding students from mandated assessment programs is to document their participation in particular special education programs. In this regard, it is interesting to note that, over the past decade, students classified as learning disabled have increased by 1.8 million, while those in Chapter 1 have decreased by 1.5 million (Allington & McGill-Franzen, 1989). Speculations about the possible reasons for this shift in categorization have included the increased funding available for special education programs and the possibility of excluding students who are not likely to do well on accountability assessments (e.g., Allington & McGill-Franzen, 1989; McGill-Franzen & Allington, 1993). Whatever the motivations, the net effect is to increase the likelihood that low-performing students will be placed in special programs.

## ALTERNATIVE ASSESSMENT

It is not a simple task to find a single term to refer to assessment practices that do not fit into the category of formal assessment. What shall we call these efforts that do not adhere to the traditional criteria of standardization, efficiency, cost-effectiveness, objectivity, and machine scorability? Probably the most commonly used label is informal assessment (García & Pearson, 1991b), but the connotation of something that lacks form (and, by implication, quality) makes it less than desirable; more important, it fails to acknowledge the rigor that can characterize these curriculum-embedded assessments. Other labels have arisen recently as educators have struggled to distinguish their efforts from the formal

assessment tradition of standardized assessments. These labels include the following:

- performance assessment: a term borrowed from the arts and athletics to emphasize on-line judgments of process/action rather than product
- alternative assessment: a description intended to capture the sense of rejecting the conventions of formal assessment
- authentic assessment: a term intended to connote the relationship of an assessment task to application in everyday situations and to distinguish this enterprise from "contrived" formal assessments
- portfolio assessment: another term appropriated from the arts to capture the sense of student involvement in the assessment process and the concept of assessment as a collection of artifacts that provide evidence about the range, depth, and trajectory of students' performance
- situated (or contextualized) assessment: a label used to emphasize the notion that assessment is particular rather than general and that it exists in and is influenced by a set of contexts (physical, political, social, and cultural)
- dynamic assessment: one of the oldest alternatives, coined to capture the sense that assessment constantly changes in character and purpose as new information unfolds
- curriculum-embedded assessment: a term that emphasizes the importance of a close match between assessment and instruction
- assessment by exhibition: a description from Sizer's (1992) work used to emphasize both portfolios and performance as key features of a good assessment system

Irrespective of the label, the hope, and the expectation, is that the instructional future will be brighter for students, especially students of cultural, linguistic, and economic diversity, once these new tools have replaced the current crop of commercially available, standardized, multiple-choice tests.

Like their formal counterparts, these alternative approaches come with a spotted history. In fact, concerns about the bias and nonobjectivity of informal assessment fed, at least in part, the development and use of formal assessment (Madaus & Tan, 1993; D. P. Resnick, 1981). Madaus and Tan (1993) remind us that quantitative scoring of students' "oral disputations" and written examinations occurred as early as the 19th century in an attempt to counter the subjectivity and bias that was thought to taint the qualitative judgments of examiners. D. P. Resnick (1981) points out that similar concerns about blatant ethnic bias in interviews and written examinations prompted the development of culturally and racially blind standardized tests within the U.S. civil service system in the early part of this century.

We have chosen to concentrate our review on two of the most popular of these new alternatives, authentic classroom assessment and performance assessment. However, we will deal with aspects of many of the other approaches in our list

as natural overlap between the categories arises. For example, portfolios are commonly discussed as an integral part of the authentic classroom assessment tradition (Goodman, Goodman, & Hood, 1989), and they are an equally important component of the performance assessment enterprise of reform efforts such as the New Standards Project (Simmons & Resnick, 1993).

The distinctive characteristics of authentic classroom assessments are that they are situated in the classroom, designed by the teacher, and used to evaluate student performance within the classroom curriculum context (see Calfee & Hiebert, 1991). Defining performance assessment is trickier. Performance assessments may or may not be designed and evaluated by the teacher; they can be externally imposed, as they are in some states such as Vermont (Koretz et al., 1992) and Kentucky (Kentucky Department of Education, 1992). They may or may not be situated within classroom curriculum contexts; in fact, those that are externally imposed are likely to be disconnected from at least some classroom curricula. The common feature to all that we include in this category is a focus on performance either directly (in the case of live observation and judgment of performance) or indirectly (in the case of the products of performance as represented, for example, by a portfolio).

As we delve into an analysis of how these new assessments relate to issues of diversity, it is important to note how new and ill studied these phenomena are. For example, few, if any, researchers have specifically investigated the cultural bias of authentic classroom assessment or performance assessment. While little is known about the impact of these assessments on students of cultural, linguistic, and economic diversity, we can gain insights about their potential advantages and disadvantages by analyzing them closely. We can also infer potential advantages and disadvantages by examining research efforts that have studied social contexts similar to those within which these new assessments operate. Again, few, if any, researchers have evaluated the specific impact of performance assessments in which students collaborate to produce a final product, but we can extrapolate from research that has focused on collaborative endeavors within cultural traditions (Au & Jordan, 1982; Delgado-Gaitan, 1987; Philips, 1972).

## Authentic Classroom Assessment

### *Aims and Tools*

Used most commonly to distance activities from the contrived assessments associated with "doing" school, the term is frequently associated with the whole language evaluation tradition (Cambourne & Turbill, 1990; Goodman et al., 1989). While authentic assessment, in terms of evaluating how well students complete real-life tasks, could and does occur outside of the classroom, we have chosen to focus on its implementation within the classroom context. In this type of authentic assessment, the goal is to gather evidence about how students are approaching, processing, and completing "real-life" tasks in a particular domain.

Authentic classroom assessment can include a range of activities: anecdotal records (Bird, 1989; Geneshi, 1985), notes from teacher observation (Dalrymple, 1989; Hood, 1989), teacher-student conferencing (Atwell, 1987; Routman, 1988), and portfolios of student work (Atwell, 1987; García & Pearson, 1991b; Routman, 1988; Tierney, Carter, & Desai, 1991). In the field of literacy, authentic classroom assessment might be extended to include recordings of oral reading (Routman, 1988), student think-aloud protocols (García, 1992b; Tierney, Readance, & Dishner, 1990), story retellings (Morrow, 1989; Muth, 1989; Tierney et al., 1990), logs of students' voluntary reading (Atwell, 1987; Routman, 1988), student journals (Atwell, 1987; Routman, 1988), writing portfolios (Tierney et al., 1991), and research projects (Goodman et al., 1989). In the field of science, it might include laboratory reports, science journals or logbooks, projects, and oral presentations (Mitchell, 1992).

Teachers who use authentic classroom assessment tend to focus on documenting the growth of individual students over time instead of comparing their performance with that of other students or groups. They generally record their findings in a narrative or descriptive format and share them with students and parents (Calfee & Perfumo, 1993). Because of the range of activities that can be included and because of the focus on individual progress, authentic classroom assessment has the potential to provide teachers with multiple lenses on student performance (García & Pearson, 1991b; Valencia & Place, in press).

Portfolios, or the collection of student work in a "folio," represent one of the most commonly used activities in authentic classroom assessment (Calfee & Perfumo, 1993; Harp, 1991; Mitchell, 1992). Although portfolio assessment has a longer tradition within writing and art than other curricula, it is now being extended to most school subjects (Mitchell, 1992). Valencia (1990; Valencia & Place, in press) distinguishes among four types of portfolios, three of which could characterize the use of portfolios for authentic classroom assessment. The first type, the showcase portfolio, highlights students' responsibility for selecting and appraising their best work. The second, the documentation portfolio, is used (either by the teacher or the student) to provide evidence of student progress over time. The third, the process portfolio, displays ongoing work on a larger project, frequently with annotations by the student concerning what each entry illustrates about the process. Only the fourth type, the evaluation portfolio, falls outside the purview of authentic classroom assessment. In the evaluation portfolio, contents are usually specified and scored, often by external examiners (see the section on performance-based assessment). This classification scheme, although useful for understanding the relationships among purpose, audience, and format, may give the impression that the types should always be kept separate. To the contrary, it is possible, perhaps even desirable, for a portfolio project within a classroom or a school to combine the first three types (Valencia & Place, in press).

As we have indicated, the whole language movement has helped to promote authentic classroom assessment (Goodman et al., 1989). Consistent with the political and epistemological tenets of whole language (Harp, 1991), students often help to choose the types of tasks they complete and the criteria by which their work is to be evaluated. They collaborate with the teacher in evaluating their own progress and accomplishments by setting personal goals; identifying tasks, activities, or projects that will be used in their evaluations; and conducting their own self-evaluations (Atwell, 1987; Routman, 1988). Cambourne and Turbill (1990) note that the type of assessment found in whole language classrooms is consistent with "responsive evaluation" (Stake, 1976), a tradition that assumes that engaging the "human as the instrument" is a practice that is equally valid as, if not more valid than, the more "objective" practices of formal assessment. Just as responsive evaluation implicates evaluators as participant-observers, so authentic classroom assessment requires teachers to observe the participants (the students), respond to their performance, and plan instruction accordingly.

Assessment within this responsive evaluation tradition is both continuous and individualized, and, as Harp (1991) explains, it bears a different relationship to instruction than does formal assessment. In traditional classrooms, instruction is program oriented, and the curricular sequence is likely to be determined by a scope and sequence of skills and/or content. In such a system, assessment (sometimes teacher-made but more frequently commercially produced tests) is used to determine how well students have mastered the instructional program. By contrast, in reform-oriented classrooms (often based on whole language philosophy or some other instantiation of a constructivist learning paradigm), instructional activities are based on the teacher's ongoing evaluation of student performance, interests, and dispositions.

In addition to progress over time, the responsive evaluation perspective provides teachers the opportunity to document the type of aid students are given to assist their development as they work both independently and as part of a group. In this sense, responsive evaluation, especially as it has been adapted to assessment within whole language classrooms, shares with the dynamic assessment tradition a commitment to socially mediated models of assessment and instruction (Feuerstein, 1979).

One final point should be made about authentic classroom assessment systems. Their inherent link to school and classroom curricula is both a blessing and a curse. When effectively implemented as part of a well-thought-out and individually oriented curriculum, authentic classroom assessment can provide teachers with the types of diagnostic information necessary to serve the individual needs of students in their class. On the other hand, if the classroom curriculum is weak, fragmented, or misguided, the resulting assessments used to evaluate student progress will not provide good diagnostic information or serve the academic needs of the students.

## *Insights From Research on Cultural, Linguistic, and Economic Diversity*

Few researchers have studied the impact of authentic classroom assessment on students from diverse cultural, linguistic, and economic backgrounds. However, the broader body of research on the relationship between culture and schooling suggests that these more situated approaches to assessment may allow teachers to accommodate cultural variations in ways that formal assessments are not likely to achieve. Just as surely, they pose some problems that, if not addressed, will lead to serious difficulties in their use.

### *Advantages*

*Variety.* Unlike formal tests, authentic assessment does not inhibit the teacher's use of a variety of methods to assess and diagnose the learning of students whose styles of learning and thinking may not fit the standardized testing paradigm (Darling-Hammond & Goodwin, 1993). Kochman (1989) points out that many African Americans and Anglos differ in interaction patterns during work, play, and school. Perhaps one of the most obvious differences is the way in which the two groups voice self-expression, with Anglos preferring a "minimalist style" that emphasizes economy and efficiency (get to the point), as well as modesty (understatement, self-effacement, and restraint), and African Americans preferring "stylistic self-expression" that emphasizes what Kochman terms "inventive exaggeration" and "dramatic self-conscious flair" (pp. 268–269). Kochman illustrates this point by referring to Koogler's (1980) description of two classroom activities that involve the same African-American and Anglo children but different teachers. In the first classroom activity, the children are directed by an Anglo teacher who wants them to participate in a planned dance activity. In the second classroom activity, they are directed by an African-American teacher, who planned a music activity but allowed the children in the class to adapt it to their own stylistic expression. The planned dance activity in the first class was not realized because of the teacher's determination to stick to the original plan. Instead of allowing the children, particularly the African-American children, to adapt the lesson to reflect their own stylistic expression and their own pairing and natural leader preferences, the teacher insisted that her plan be followed and interpreted the students' attempts at adaptation as insubordination. In contrast, the music activity planned by the African-American teacher was realized, perhaps because the teacher was flexible and accepted the evolving modifications that the students, many of them African-American, made as the activity unfolded.

*Cultural resources.* The research on home-school discontinuities (Delgado-Gaitan, 1987; Erickson & Mohatt, 1982; Heath, 1982, 1983) suggests another advantage for authentic assessment. If, as a profession, we think it is important to allow and even encourage teachers to build on the cultural resources of their students, we need, at the very least, assessments that do not penalize teachers and students for cultural adaptations. Even more preferable would be assessments

that openly reflect cultural values. One of the most celebrated cultural adaptation success stories is the reading program of the Kamehameha Early Education Program (KEEP) in Honolulu, Hawaii (Au, 1981, 1993a; Au & Jordan, 1982; Au & Mason, 1981). Au and Jordan (1982) report that teachers in the program make several accommodations in an attempt to enhance the Hawaiian children's level of participation. First, they establish authority through the means operating within the culture (by earning it through interactions rather than assuming it because of their position). Second, they encourage students to use "talk story," a Hawaiian interactional style that involves mutual participation and co-narration of a story by both the teacher and the students. Third, the KEEP reading program integrates informal learning practices characteristic of the Hawaiian community with formal learning practices characteristic of the school, so that both the teacher and students are comfortable with what is being taught and how it is being learned.

Erickson and Mohatt's (1982) study of two successful teachers of Native American children on the Odawa Indian Reserve in northern Ontario represents another example of how teachers can adapt to students' styles of thinking and learning. In this case, one of the teachers was from the Indian reserve, while the other one was Anglo. Although the Native American teacher's interactional styles ("tempo of teaching and overall directiveness") were congruent with those found in the Native American community, not everything she did reflected Native American culture. Erickson and Mohatt point out that she tended to bridge mainstream and Native American culture through "mixed" cultural forms that allowed her to use the standard curriculum and standard "teacherly" ways while simultaneously accommodating to "Odawa principles of communicative etiquette" (p. 168). The Anglo teacher also used mixed forms, although his interactional style was more mainstream, especially at the beginning of the school year. What is most interesting is that his style of teaching changed so that, by the end of the year, he was using some of the same methods of teaching as the Native American teacher. For example, he had the students sit together in table groups instead of individually in rows; he tended to call on the table groups more than individual students; and he decreased the amount of time spent on whole-group lessons, spending more time with students one on one or interacting with small groups of students. Both of these examples, because of their emphasis on the importance of the group rather than individuals, point to authentic assessment as the medium most likely to honor students' cultural resources.

*Flexibility.* The flexibility inherent in the authentic assessment tradition should provide advantages for linguistically diverse students. García (1992b) points out that authentic assessment can be used to document what bilingual students know and can do in both languages, something that is difficult if not impossible to capture on formal, especially English-only, assessment measures. For example, preschool and kindergarten bilingual children, especially those who are acquiring English in one setting and their native language in another, may know some vocabulary concepts in one language but not in the other. Their total knowledge

across the two settings will be underestimated on a formal test because the test will be written in only one of the two languages. If teachers allow bilingual students to use both languages in the classroom, they can document the initial appearance of the concept in each language. Bilingual students frequently will demonstrate greater learning if they are allowed to choose their preferred language of response: their native language, English, or both (Chamot, 1980; García, 1991; Lee, 1986).

Perhaps even more important in designing instruction, flexible approaches to assessment allow teachers to differentiate between difficulties caused by lack of English and difficulties caused by lack of knowledge or expertise. As we reported earlier, García (1991) discovered that Latino students frequently answered comprehension questions incorrectly on a written English test not because they did not understand what they had read but because they could not find a familiar English paraphrase of the vocabulary from the selection they had just read. When the test items were translated into Spanish, they readily answered the questions. Using dynamic assessment techniques, teachers could document the problems that students face, indicate the type of scaffolding provided to help them perform successfully, record how they respond to the help, and note when they no longer face those difficulties.

Authentic assessment holds great promise in providing valid information about emerging knowledge and development. A serious problem with formal assessment is that a low score tends to obscure what a student may have actually accomplished. For example, in an innovative study of vocabulary learning, Stallman (1991) found that while all students, even low-performing ones, learned new vocabulary in the context of reading, traditional assessments were insensitive to the partial knowledge that they were acquiring. Only when she used a vocabulary test that measured partial knowledge was she able to document growth for all students. Similar types of findings have been reported in mathematics. In a case study of children's difficulties with mathematics, Ginsburg and Allardice (1984), in one of the few research efforts that includes substantial representation of low-income and African-American students, found that assessments using formal written formats characteristic of schooling (test items, word problems, and the like, what they termed System 3 knowledge) often obscure the less formal mathematical knowledge that students possess within what they termed System 1 (intuitive senses of mathematical concepts, such as number, equality, and inequality) and System 2 (counting to solve problems). Because the skills of System 3 are assessed in formal tests, teachers who used these tests tended to underestimate their students' knowledge of mathematics. As a result, the students received inappropriate and watered-down instruction. Ginsburg and Allardice point out that teachers need to know how children conceptualize mathematics and the ways in which they use informal skills, error strategies, and invented procedures (in other words, the very skills and strategies that are not normally included on formal tests).

*Adaptation.* Perhaps the greatest potential advantage of authentic assessment is that it can be tailored to document the issues that individual classroom teachers regard as important for their students. For example, in Spanish-English bilingual classrooms, teachers will want to know what literacy tasks a child can complete in English, in Spanish, or in both languages (García, 1992b). They will want to know the extent to which their students interpret material and vocabulary based on their cultural and linguistic experiences or on mainstream experiences (García, 1991). Similarly, they will want to know the extent to which bilingual students are capable of using their knowledge of native-language reading to help in their second-language reading (Downing, 1984; Jiménez, 1992; Jiménez, García, & Pearson, 1991). Teachers working with dialect-speaking African-American youths on improving their writing also might want to evaluate these students' use of dialect features apart from their ability to develop a persuasive essay (García & Pearson, 1991b). It is difficult to imagine formal assessments that could or would attempt to gather such information.

### Potential Difficulties

*Knowledge requirements.* Classroom-based assessment for students of cultural, linguistic, and economic diversity requires teachers who are knowledgeable not only about the academic domains being assessed but about students' cultures and languages (García & Pearson, 1991b). Although teachers probably know more about their students than outside evaluators, teacher bias is a potential problem. Second-language experts, in particular, have been concerned about assessments that rely excessively on teacher judgment to determine the language dominance or proficiency of second-language speakers of English (August & García, 1988). For example, not all of the teachers who work with these students know their home languages. Those unfamiliar with bilingual development may not know what to expect of students of different age and experience levels. Moll, Estrada, Diaz, and Lopes (1980), in a study of Spanish-English bilingual students' reading, recount an incident in which Spanish-proficient readers were placed in a low-English reading group because the mainstream teacher, who did not know Spanish and was not familiar with second-language literacy research, misinterpreted their less-than-fluent English pronunciation as symptomatic of serious reading problems in English. Instead of providing the students with reading instruction that built on their reading expertise in Spanish, the teacher provided them with low-level decoding instruction. African-American dialect-speaking students have received similar instructional emphases; well-meaning teachers have tried to help by correcting dialect-motivated, meaning-preserving errors instead of helping them acquire comprehension strategies (Cazden, 1988; Collins, 1982; Cunningham, 1976–1977).

*Discourse conventions.* Bias can also result from our lack of awareness about the ways in which our own cultural values influence judgments about students (García, 1992a; Hymes, 1972), particularly those whose cultures are often margin-

alized by school practices. Before we use classroom activities for assessment purposes, we need to know the extent to which students are familiar with them (García & Pearson, 1991b). For example, even a practice as seemingly innocuous and universal as storybook reading can be culturally problematic. Teachers who wish to learn about students' oral language development by noting the ways in which they use language during storybook reading (e.g., by periodically tape-recording or videotaping storybook reading sessions or by recording observations in a notebook) will need to know how students react and interact in these situations. Heath's (1983) comparative study of literacy events, including storybook sharing, in three communities (African-American working class, Anglo working class, and Anglo middle class) is relevant to this issue. Only the middle-class preschool children freely engaged in school-like storybook interactions. The Anglo working-class families used storybooks in interaction patterns that were quite different from what was expected in school. Although the African-American children were immersed in a rich oral environment, storybook reading and parent-child interactions involving storybooks were not a dominant characteristic of these households. This example suggests that teachers who try to use storybook interaction patterns as an oral language assessment tool would have to account for children's prior experiences with this activity. Children unaccustomed to the school version of the activity might mask a high level of oral language development by hesitating to participate in an activity that seems strange.

Mainstream school culture has promoted a widespread discourse pattern for classroom discussions. Interactions tend to follow a pattern in which the teacher *initiates* an interaction, students *respond,* and the teacher *evaluates.* Teachers use this pattern for a variety of classroom discourse functions, including assessing student learning (Cazden, 1988; Mehan, 1979). Inferences that teachers draw from such interactions assume that students are familiar with and recognize the discourse function of the pattern. Sociolinguistic evidence does not support the validity of such an assumption. For example, Heath (1982) found that cultural differences between teachers and students accounted for African-American students' reticence to participate in classroom story discussions as well as teachers' misunderstandings about their story comprehension. The African-American students, who expected questions that had "real" answers (e.g., they were unknown to the questioner), were surprised by questions for which the questioner already knew the answer (e.g., What is the boy doing in the picture?). As the following transcript indicates, they were reticent to respond to school-like questions.

| | |
|---|---|
| Teacher: | What is the story about? |
| Children: | (silence) |
| Teacher: | Uh . . . Let's see . . . Who is it the story talks about? |
| Children: | (silence) (Heath, 1982, p. 105) |

As the continuation of the transcript suggests, the students' silence was better explained by their curiosity about the interchange than their lack of comprehension.

Teacher: Who is the main character: Um . . . What kind of story is it?
Child: Ain't nobody can talk about things being about theirselves! (Heath, 1982, p. 105)

*Social conventions for work and play.* Teachers' expectations about the context and the method of task completion might give them a misleading impression of students' capabilities. On the basis of an ethnographic study of Mexican-American children, Delgado-Gaitan (1987) noted a conflict between the working contexts of their culture (sibling and peer groups to accomplish tasks assigned by parents) and their school (the teacher expected them to accomplish tasks by working individually and competitively). When the students tried to work in groups or help each other to complete tasks, they were accused of cheating and reprimanded accordingly. Similar findings have been reported for Native American (Erickson & Mohatt, 1982; Philips, 1972), Hawaiian (Au & Jordan, 1982; Boggs, 1972, 1978), and African-American students (Kochman, 1989). Cultural accommodation to context and method does assist learning (e.g., Au & Mason, 1981). In one of the more interesting recent developments, college calculus teachers have found substantial performance gains for African-American, Latino, and rural Anglo students when they are encouraged to work in cooperative study groups to solve difficult problems (Treisman, 1992).

## Performance Assessment

### *Aims and Characteristics*

With the exception of a long tradition of performance assessment in art, music, and athletic competition, one of the earliest modern-day uses of performance assessment as a regular part of the curriculum has been in the writing field. For the last 20 years, writing educators and scholars have rejected indirect multiple-choice assessments (which typically require students to spot errors in spelling, punctuation, grammar, and usage) in favor of direct writing assessments in which students write for an extended period, usually an hour, in response to a prompt.

Recently, as a part of the alternative assessment movement, performance assessment has permeated the entire school curriculum (Mitchell, 1992). A variety of states, including Arizona, California, Maryland, Vermont, Kentucky, and New York (Mitchell, 1992); some national groups, such as the New Standards Project (Simmons & Resnick, 1993) and the Coalition of Essential Schools (Sizer, 1992); and a few schools, such as Walden III High School in Racine, Wisconsin (Mabry, 1992), and Central Park East Secondary School in New York (Mitchell, 1992), have pursued performance assessments in several subject areas (e.g., reading, writing, mathematics, and science). Maryland (Kapinus, Collier, & Kruglanski, in press) even seems to be moving in the direction of an interdisciplinary performance assessment.

Performance assessments, as they are currently being articulated and implemented, possess several distinctive characteristics. First, they represent or closely simulate performance in real-world settings. Second, they are inherently entangled

with instruction. As Shavelson, Baxter, and Pine (1992) suggest, "a good assessment makes a good teaching activity, and a good teaching activity makes a good assessment" (p. 22). Third, they are grounded in the essence of the discipline (Mitchell, 1992). Thus, a good mathematics task will reflect the knowledge, skill, and dispositions that mathematicians believe are at the core of mathematical thinking; a good science task will engage students in genuine scientific inquiry; and a good writing task will allow students to use the tools that real writers use. Fourth, scoring goes beyond a quantitative summary of a student's competence to encompass mastery of process and dispositions; scoring will "capture not just the right answer, but also, the reasonableness of the procedure used to carry out the task or solve the problem" (Shavelson et al., 1992, p. 22).

In mathematics, performance assessment might confront students with a problem requiring the use of several mathematical operations in addition to both inductive and deductive reasoning. For example, in one Maryland eighth-grade task (Mitchell, 1992), students are asked to negotiate the various steps involved in designing a restaurant. They take on the roles of designer, developer, market researcher, financier, and builder. In the process, which extends over several days, they work with the teacher in small groups and individually. This activity culminates in a package that includes a market research questionnaire, data displays from the questionnaire, a scale model of the restaurant, written cost estimates, and a summary paragraph (written to the local zoning board) explaining the decisions they have made. The task is scored holistically for each of five dimensions: communication, reasoning, problem solving, connections, and technology. The score is based on the teacher's evaluation of the product and observations of the process.

In California, a typical language arts task (Weiss, in press), administered over a period of 3–5 days, might require students to read and respond to two full-length, intact stories (or a story and a thematically related poem); meet with a small group of peers to brainstorm ideas for a writing project; write a first draft; participate in a peer conference; and write a final draft. Holistic scores, one each for reading and writing, would probably be assigned by applying rubrics developed by teachers from around the state.

Performance assessment borrows assumptions and practices from both formal assessment and authentic classroom assessment. It borrows from formal assessment the goal of providing external indices of student progress that can be used to make judgments about individual competence and program effectiveness. While it emphasizes close ties to district, school, and classroom curricula, it also recognizes the importance of examination tools that are external to the environment (and individuals) under evaluation. This distance is particularly important when examinations are used for purposes of "certifying" competence, mastery, or quality (Simmons & Resnick, 1993). Thus, it would make sense to talk of a districtwide or even a statewide performance assessment, just as we now freely discuss districtwide or statewide standardized tests.

From authentic classroom assessment, performance assessment borrows a bias toward providing real-world, authentic tasks that can be incorporated into the curriculum. For example, one criterion used by the New Standards Project for evaluating stand-alone performance assessments is that teachers should want to use them as instructional units even if they are under no obligation to use them as tests. Performance assessment also rejects the assumption of decomposability (L. B. Resnick & Resnick, 1992): the idea that the way to ensure complete assessment of a domain is to decompose it into more basic components and assess each component thoroughly. Instead, students' knowledge and ability to perform "authentic" tasks are assessed by evaluating how they can integrate information and skills to perform the task.

Although performance assessment shares with authentic classroom assessment an emphasis on human judgment, it addresses the inherent subjectivity of judgment with different tools. In a classroom, a teacher deals with subjectivity by collecting many samples of performance across time and contexts. In performance assessment, a quasi-psychometric set of guidelines for confronting subjectivity has evolved. To guard against subjectivity and bias in performance-based decisions, scorers undergo systemic training. Most scoring systems supply scorers with rubrics that spell out the standards or criteria for various levels of work. The levels vary considerably between settings. For example, in the KEEP program (Au, in press), criteria are spelled out for below, at, and above grade-level work for each elementary grade. In the New Standards Project (Simmons & Resnick, 1993), the levels are defined in terms of what a student (with the help of a teacher) would have to do to meet the standard of excellence set for all students (4 = meets the standard, 3 = needs revision, 2 = needs further instruction, and 1 = too little evidence to draw a conclusion). To anchor the rubrics in performance, several illustrative samples of student work at each score point (usually called anchors or benchmarks) are provided to scorers. Finally, scorers are trained in the use of the rubric and the anchors for several hours, and sometimes days, before they score "live" samples of student work.

Teachers who have been involved in various stages of this process (developing rubrics, choosing anchors, or undergoing training for scoring) have reported a great sense of professionalization (Mitchell, 1992). A by-product of building the tools needed to conduct performance assessment (including portfolio assessment) is the opportunity for teachers to engage in professional development and reflection (Darling-Hammond & Ancess, in press). This suggests that if performance assessment is to influence instruction, it is important to involve the teachers who will be implementing the assessments at every stage in the process: task development, rubric development, anchor selection, and scoring.

The process of developing rubrics can make assessment very public. In the New Standards Project, for example, the rubrics are shared in advance with students and are available for inspection throughout task completion. The exam-

ples below represent the "student" version of the writing and reading rubrics used in the 1993 New Standards Project (1993) pilot examination.

Writing: Before you begin to write, we think it is important for you to know what the scorers will look for when they read your work and assign a score to it. An outstanding paper will:

- help the reader understand how the mural expresses what your class has learned about Harriet Tubman.
- tell the reader your thoughts and feelings.
- be well-organized and easy to follow.
- show that you have chosen words carefully to express what you want to say.
- show that you use mechanics (spelling, punctuation, etc.) to make your meaning clear to readers.

Reading

My responses should show that . . .

- I understand the idea of the whole text as well as the important parts,
- I can connect the ideas in these readings to other texts, my own ideas, and my own experiences,
- I can evaluate the way the author writes,
- I can agree or disagree with the author's ideas, and
- I can reflect on my reading and writing to develop new ideas of my own.

As we noted earlier, the distinction between performance assessment and authentic classroom assessment is less a matter of what students do than how educators evaluate student work and use the results. Within a classroom, a teacher is more likely to be interested in monitoring individual student progress over time. Those who examine the same activity with an external lens are more likely to be interested in certifying individual student mastery (Did the student meet the standard?) or evaluating the effectiveness of a classroom, school, or district program (What proportion of the students met the standard?).

Portfolios have gained popularity because of their potential to provide teachers with an "insider's" perspective. In terms of Valencia's four types of portfolios (Valencia & Place, in press), the first three (showcase, process, and documentation) promote this insider's view. Increasingly, however, educators are considering the potential of portfolios for purposes of external evaluation, as exemplified by Valencia's fourth category: the evaluation portfolio. To our knowledge, at least within public education, the longest standing example of using portfolios for certification involves the Rites of Passage Examination process at Racine's Walden III High School (Mabry, 1992). Since the early 1970s, seniors desiring to graduate from Walden III have done so not by accumulating Carnegie units but by defending a portfolio consisting of a number (up to 14) of optional and required entries to a committee of teachers, citizens, and peers. In 1992, the first class of seniors from Central Park East Secondary School in New York (see Darling-Hammond & Ancess, in press; Mitchell, 1992) received diplomas on the basis of portfolio presentation. Both of these sites are of interest to our review because their populations are culturally and ethnically diverse.

Performance-based assessment is still in its developmental stages. Although task or construct validity (Does the task measure what it is supposed to measure?)

does not appear to be a major problem, reliability (Will the same performance result in the same score, regardless of who scores the assessment or when the student takes it?) is still a concern for many (Mitchell, 1992). More serious than the issue of reliability is the matter of generalizability, or what Shavelson and his colleagues (1992) have called intertask reliability: How confident can we be that this sample or these samples of performance truly represent an individual's level of mastery over a particular domain of inquiry? The data from Shavelson and his colleagues' foray into performance assessment in science are sobering. Students did not achieve consistent scores across performance tasks. This problem is not new, however; it has been with us as long as ETS has been administering advanced placement examinations. Low task generalizability presents grave equity concerns for educators who would like to implement a performance assessment system in which students either choose or are assigned different tasks.

As was the case with authentic classroom assessment, research about performance assessment, especially concerning diverse populations, is very sparse. Once again, we will turn to research on the academic performance and instruction of culturally, linguistically, and economically diverse students to assess the potential, both positive and negative, that performance assessment bears for them.

### *Insights From Research on Cultural, Linguistic, and Economic Diversity*

In terms of cultural and linguistic diversity, the potential of performance assessment varies according to the features that it borrows from formal and authentic classroom assessment. For example, the fact that performance assessment tasks or systems are typically developed by individuals outside the context of an individual classroom, although frequently with the input of practicing teachers, carries both positive and negative implications. On the positive side, it reduces the knowledge burden that individual teachers would otherwise have to bear and permits a situation in which knowledge about academic domains and students' cultures and languages can be distributed across teachers. The danger stems from the reality that there is nothing inherent in the tasks or the existing processes that will guarantee the involvement of task developers or teachers who are knowledgeable about language and cultural factors. Because performance assessment is so new and is being developed by a variety of individuals and groups, it will almost inevitably be operationalized in many ways. In the sections to follow, we point out the possible advantages and disadvantages that different versions of performance assessment may offer students from diverse cultural, linguistic, and economic backgrounds.

#### *Advantages*

*Avoiding skill decomposition.* The bias toward holistic tasks and holistic approaches to judging performance decreases the likelihood that performance assessment will promote skill decomposition. Too often, in the name of reducing the complexity of an overwhelming curriculum, low-performing students are

provided instruction that decomposes complex tasks and requires mastery of one component of a task before students can begin to approach the next. For example, in the field of reading, it is not unusual for teachers to focus on low-level decoding tasks in their instruction of low-achieving readers. Low-achieving readers often do not have the opportunity to read real texts or to write because they have not demonstrated complete mastery of individual components, especially word-level processes, of reading and writing (García & Pearson, 1991a; García et al., in press). As we noted in our discussion on formal assessment, studies of teachers' use of standardized tests have indicated that teachers of low-income students tend to provide instruction that matches such tests because they are more concerned about these students' academic performance within the curricular framework operative in the classroom (Dorr-Bremme & Herman, 1986; Rothman, 1992). Performance assessment tries to resolve this problem by providing all students with complex tasks that require the authentic, integrated application of knowledge and strategies.

*Dynamic approaches to assessment.* Despite a bias toward evaluating products (finished tasks or portfolios), performance assessment also can be designed to provide students with the scaffolding needed to perform successfully. In this regard, it is similar to dynamic assessment (Feuerstein, 1979), in which teachers are encouraged to provide increasing amounts of scaffolding to determine which tasks students can complete independently and which they can complete with varying levels of assistance. In a sense, dynamic assessment holds task completion constant—by assuming that any student can complete a given task as long as he or she receives expert guidance—and varies the amount of social guidance (i.e., scaffolding) that students might need. If dynamic assessment becomes a key component of performance assessment, it obviously would be beneficial to all students. This dynamic feature should be especially helpful to students who are learning English as a second language. Within the philosophical parameters of dynamic assessment, teachers would be able to provide students with background knowledge essential to text comprehension, translate obscure English vocabulary that might block an otherwise transparent linguistic translation, or provide other forms of assistance that bilingual students might need in order to comprehend and complete tasks in English (see García, 1991, 1992b).

*Client participation.* Performance assessment also allows the participation of students, teachers, and even parents in the assessment process. In portfolio traditions, for example, students often decide which pieces will be placed in the portfolio. Parents and students are often involved in determining the significance of pieces or performances, in examining and evaluating progress over time, and in deciding on the consequences of the work. In the work described by Murphy and Smith (1991), for example, students were asked not only to select entries for the portfolio but to provide annotations explaining their choices as well as a reflection on their journey from the beginning to the end of the year.

Other things being equal, active participation should benefit students from diverse cultural, linguistic, and economic backgrounds, especially if they are allowed to share their rationale for particular entries and to express their point of view freely, without fear of reprisal or harsh criticism. For example, an Anglo student from Appalachia who speaks a dialect of standard English might choose to include a sample of writing in which she wrote a strong, persuasive essay for publication in a student newspaper even though it was sprinkled with dialect features. If such an entry had to stand on its own without benefit of explication, the student would certainly be putting herself at risk. But with the benefit of an explanatory annotation, she might emphasize the conceptual and rhetorical power of the writing while acknowledging its variance with, and perhaps even admitting a personal need to increase her mastery of, standard English. When students and teachers are involved in and assume ownership of the assessment process, they can focus more honestly and forthrightly on strengths and weaknesses. The greatest benefit might be providing dialect-speaking students with a way of balancing their need to learn the "power code," standard English, with their personal need to honor their own dialect and language identity (see Delpit, 1988; Gee, 1990).

*Intercontextuality.* Performance assessment also allows students to be tested in a range of settings. For example, there are opportunities for students to work alone, in pairs, and in groups. This type of flexibility is consistent with the literature on cultural diversity, which has suggested cultural preferences for different ways of participating in classroom activities (Au, 1981, 1993a; Delgado-Gaitan, 1987; Erickson & Mohatt, 1982; Philips, 1972). If teachers are allowed to vary settings in accordance with the cultural preferences of their students, perhaps even varying the "cultural comfort" of the context, they can evaluate fairly the impact of context for particular students. If, on the other hand, only a single participation structure is allowed in a classroom, even if that single structure is based on the currently popular movements of cooperative learning and collaboration, some students will be disadvantaged.

*Assessment as public discourse.* Developers of performance assessment talk about making the criteria for success public. Currently, assessments vary on the dimension of "publicness." Some are very private, shrouded in secrecy, hidden in closets, and shrink-wrapped to discourage tampering, foul play, or cheating. Gatekeeping tests, such as the Scholastic Aptitude Test, the American College Test, or the Graduate Record Examination, anchor the most "secret" end of the continuum. Performance assessments, as they are currently being piloted, fall at the public end of the continuum. In the New Standards Project, as illustrated earlier, the general criteria for different levels of performance (a sort of generic rubric consisting of statements about dimensions of quality) are made known to teachers and students prior to the examination. Making criteria public should benefit low-performing students, many from culturally, linguistically, and eco-

nomically diverse backgrounds, by allowing them access to standards for success that are frequently kept hidden.

### *Potential Difficulties*

*Canon issues.* It remains to be seen whether performance assessment will reflect diverse points of view and knowledge (see Gordon & Bhattacharyya, 1992; Greene, 1993). So far, no one has discussed whether performance assessment will privilege one or another canon (Hirsch, 1987) or can be extended to include literature and topics that address multicultural issues relevant to those students whose cultural participation in the canon has been marginalized. Greene's (1993) question hits the target on this issue: "To what extent are 'multiple voices silenced over the years' now part of the 'ongoing conversation'?" (p. 13). If proponents of performance assessment try to achieve equity by establishing uniform assessments at the school, district, or state level, it is likely that students from diverse cultural, linguistic, and economic backgrounds will be deprived of "the opportunity to see themselves [and] their history from their own cultural perspective as well as see the world around them from multiple perspectives" (Greene, 1993, p. 13).

Experts in multicultural education point out how difficult it is for mainstream educators to identify topics that are culturally relevant to minority students (Banks & Banks, 1993; Hernandez, 1989; Sleeter & Grant, 1988). For example, in selecting heroes for inclusion in a curriculum, mainstream educators may choose individuals who they believe made a significant contribution to mainstream American society, including members of minority groups. Such a list might include Martin Luther King and Tecumseh, along with Eleanor Roosevelt and John Kennedy. Rarely, however, would they identify heroes and heroines who made a significant contribution from the perspective of a particular ethnic or cultural group. Even worse, when mainstream educators tell the stories of minorities, they sometimes do so from a mainstream perspective rather than a particular cultural perspective. For example, a published account, written for a juvenile audience, of Phillis Wheatley, a famous African-American poet, notes that she was lucky as a slave to be owned by a master who was prosperous and treated her relatively well. The tone of this account betrays a biased mainstream perspective.

*Minority participation in development.* Even the assurance that minority educators have been involved in selecting and developing topics, tasks, and rubrics cannot guarantee representation and relevance. For example, Mitchell (1992) reports that minority participation was a critical problem in the development of the California Assessment Program. Panels of teachers were intended to reflect the proportional distribution of distinct ethnic groups in the student population. However, the percentage of minority teachers employed in California is so far below the percentage of students that the goal could not be achieved.

The selection process itself is critical. In her research on culturally relevant teaching, Ladson-Billings (1992) selected successful teachers of African-Ameri-

can students by using an approach Foster (1990) termed "community nomination." In this approach, parents from the community identify those teachers whom they consider to have been successful in educating their children. Ladson-Billings points out that the teachers selected by the African-American parents encouraged students "to choose academic excellence" at the same time that they "allowed them to maintain a positive identification with their own heritage and background" (p. 382). In doing the latter, the teachers helped their students to see the contradictions and inequalities in our society that affected their performance and participation. Ladson-Billings (1992) explains that these teachers empowered their students "intellectually, socially, emotionally, and politically by using cultural referents to impart knowledge, skills, and attitudes" (p. 382).

*Overreliance on literacy.* Many of the performance assessment tasks being developed, even those in mathematics and science, rely heavily on students' ability to read and write standard English (NCTM, 1989; Simmons & Resnick, 1993). In fact, those who support this reliance point out that such tasks promote curriculum integration and allow us to evaluate students' dispositions to use their subject-matter knowledge and skills to solve problems. The cost of this integrated approach to assessment is the confounding of literacy skills with subject-matter skills and knowledge. This confound affects our ability to interpret the subject-matter performance of students, especially those who may not be fluent in standard English. As we noted earlier in reviewing formal assessment, reliance on reading and writing to show competency in mathematics may actually disadvantage some groups of students who currently are doing quite well in mathematics, such as Asian-American students (Tsang, 1989), as well as provide misleading profiles of other students (Ginsburg & Allardice, 1984). Gardner (1985), in supporting his multiple intelligences theory, carries this argument about overreliance a step further, arguing that most assessments reward mathematical as well as verbal capacity to the exclusion of other human capacities.

*Outside examiners.* When performance assessment requires the use of outside examiners, as has been suggested for teacher certification (National Board for Professional Teaching Standards, 1990), issues of cultural awareness, trust, and possible misinterpretations become paramount. In fact, the standards developed by the assessment profession (American Psychological Association, American Educational Research Association, and National Council on Measurement in Education, 1985) acknowledge this point directly:

> In educational, clinical, and counseling applications, test administrators and users should not attempt to evaluate test takers whose special characteristics—ages, handicapping conditions, or linguistic, generational, or cultural backgrounds—are outside the range of their academic training or supervised experience. (Standard 6.10, p. 43)

Labov's (1969) work demonstrates that establishing trust requires more than cultural understanding. He found that even an African-American researcher was

unable to elicit substantial samples of the natural language of African-American children until he switched from a formal to an informal assessment context. Children who uttered only guarded, monosyllabic responses in the formal setting talked almost nonstop in the informal setting (which the researcher signaled by interviewing pairs of good friends, bringing in potato chips, positioning himself at their level, and introducing topics that were taboo in schools).

*Adaptability.* Whether performance assessment proves to be a useful tool for students of cultural, linguistic, and economic diversity depends on its capacity to provide information that is both culturally and individually relevant to their success in school. What may be important for understanding and interpreting the progress of an African-American dialect-speaking child may be different from what is important for a bilingual child or a monolingual Anglo child. The teacher of the dialect-speaking African-American student needs to know whether the child is learning to distinguish between language contexts in which dialect is appropriate, and perhaps even preferable (e.g., in a speech, in the dialogue of a short story, or in a poem), and language contexts in which the use of dialect might result in discrimination (a job interview) or misinterpretation of competence (an academic essay). The teacher of the bilingual student needs to know, among other things, whether concepts that appear unknown in one language are actually known in the second, whether the student can take advantage of his or her bilingualism to transfer knowledge from one language to the second, and whether the student demonstrates greater subject-matter competence in one or another of the languages.

## Reflections on Alternative Assessment

As we noted earlier, it is difficult to assess the impact of alternative assessment on the schooling and academic performance of students from diverse cultural, linguistic, and economic backgrounds because research and development are at such a primitive stage. Two major questions need to be answered as our work unfolds. First, to what extent will teachers and administrators in low-income schools embrace these new assessments? Second, to what extent will their use change the type of instruction now offered to low-income students? An examination of the current literature on the implementation of literacy portfolios (both in the authentic classroom tradition and in the performance assessment tradition) heightens our concerns about the answers to these two questions.

Findings from studies of classroom portfolio implementation (Au, in press; Dewitz, Carr, Palm, & Spencer, 1992; Flood, Lapp, & Monken, 1992; Johnston, Guice, Baker, Malone, & Michelson, 1993; Valencia & Place, in press) suggest that teachers' use of portfolios depends on the type of instruction they customarily use, the fit between portfolios and their teaching paradigm, and who initiates the use of portfolios. Calfee and Perfumo's (1993) survey of best-case portfolio practices in the language arts indicated that these practices tend to take place in schools in which school-based decision making is valued, teachers ascribe to a

whole language philosophy, and cooperative learning is well established. Reviews of instructional practices in low-income schools (Cazden, 1988; Dorr-Bremme & Herman, 1986; García et al., in press; Herman & Golan, n.d.) suggest that the instructional practices described by Calfee and Perfumo are not characteristic of instruction in such schools.

When teachers' customary instruction does not fit the instructional paradigm that underlies portfolio assessment (with its emphasis on process as well as product, authentic tasks, and teacher and student reflection), conflicts in use occur, with teachers rarely using portfolios in an integrated manner for both instruction and evaluation (Dewitz et al., 1992; Flood et al., 1992; Johnston et al., 1993). This finding appears to hold across schools with varying demographic compositions. For example, Au (in press) traced the use of portfolios by teachers of Hawaiian students in the KEEP program. In this setting, the teachers were learning about portfolio assessment at the same time that they were being guided (by the KEEP staff) toward a "whole literacy approach." Au reported that the most serious problem underlying the use of portfolios by the teachers was not logistical (although there were logistical problems) but conceptual. Most of the teachers were uneasy about leaving the criterion-referenced tests and scope and sequence of skills found in the commercial materials. They were not comfortable using professional judgment to make instructional decisions based on the portfolio assessment measures. Rueda and García (1992) reported similar findings for teachers of Latino language minority students. Although some of the teachers used aspects of authentic assessment in their classrooms, there were discrepancies between their beliefs about instruction, assessment, and bilingual students and the theoretical assumptions that underlie many of the new assessment initiatives. Finally, Johnston et al. (1993) studied the assessment practices of teachers who were using literature-based reading programs, in which, at least from a philosophical perspective, alternative assessment would be advocated. They found a tendency for some teachers to ignore the potential information offered by alternative assessments and to attribute the low performance of less successful students to their home life or earlier schooling. Within this very complex political and curricular context, the tendency toward child-based attribution was more severe in situations in which there was greater pressure for accountability (despite the new curriculum and assessment practices, teachers in some districts were still accountable to district- and state-administered standardized tests) and lower levels of knowledge about the new curriculum.

We found only one study that has coupled analyses of the implementation of alternative assessment with data on student performance. Au (1993b), in a follow-up to her earlier work (Au, in press), examined the relationship between portfolio implementation and the literacy performance of native Hawaiian elementary students. In the initial year of implementation, Au (in press) found low levels of performance on the benchmarks within the KEEP portfolio system. KEEP staff members were concerned that the teachers might not be using the data from

the portfolio system to inform instruction. In the next year, they implemented a portfolio checklist in the classrooms of several highly committed volunteer teachers in order to evaluate the degree to which the portfolios were actually being used. When supervisors used the checklist to monitor teachers' use of the system, the percentage of students who achieved at or above grade level on their portfolio benchmarks rose dramatically, at least in comparison with students in lower compliance classrooms. Au's findings suggest that teachers who are committed to the use of alternative assessment can use it to help improve their students' performance when they receive the necessary support and guidance.

Although few researchers have systematically investigated the effects of alternative assessment on student performance, Darling-Hammond and her colleagues (Darling-Hammond & Ancess, in press; Darling-Hammond, Ancess, & Falk, in press; Darling-Hammond & Ascher, 1990) have documented the positive impact of portfolios and other forms of authentic assessment in several urban school settings, most of which enroll high proportions of ethnically and linguistically diverse students. Foremost among these is Central Park East Secondary School. After examining the work at Central Park East, Darling-Hammond and Ancess (in press) concluded that the school's graduation-by-exhibition system of portfolio assessment "activates continual inquiry into the goals and standards of the school in a manner made more compelling because it focuses on the actual work of students" (p. 12). Teachers at Central Park East report that the portfolio system has shifted their focus from an emphasis on factual content, which was promoted by the multiple-choice Regents examination, to the structure of the discipline, the interactions between disciplines, and the use of knowledge to solve problems. Darling-Hammond and Ancess (in press) reported similar shifts in focus for performance assessment systems at Hodgson Vocational-Technical High School in Delaware, the Urban Academy in New York City (an alternative high school for alienated or previously unsuccessful students), and New York's International High School (with a 100% immigrant population). Consistent with Mitchell's (1992) report of a shift in professional development, elementary teachers at Brooklyn's P.S. 261 who have begun using the Primary Language Record (a London-developed observation-assessment system for documenting early language and literacy growth) have reported that the multiple perspectives required in the system, along with a clear focus on individual student performance, have transformed their staff development from a transmission to an inquiry model of learning and knowledge construction (Darling-Hammond, Ancess, & Falk, in press).

The most encouraging development in teacher response to alternative assessment is the impact of these new approaches on teacher reflection and professionalization. An example from Central Park East ably demonstrates this change. When teachers at this school began the process of examining student work in the meticulous and thoughtful manner demanded in portfolio assessment, they began to raise questions central to the schooling of low-income and minority students:

How do we "encourage historically discouraged learners while simultaneously upholding a standard of excellence? Which support structures for student achievement need to be adjusted and how? How can teachers in all grades refine their teaching to support student success?" (Darling-Hammond & Access, in press). An interesting, and open, question is whether a standardized test has ever encouraged this sort of reflection.

## CONCLUDING REMARKS

In general, educational assessment has not been friendly to students from diverse cultural, linguistic, and economic backgrounds. The documentation on this point is clear for formal tests. And some of the problems associated with authentic classroom assessment and performance assessment raise questions about whether they will prove to be any friendlier than formal assessment. Whether they do depends on their capacity to accommodate and even privilege cultural, linguistic, and individual differences in the process of gathering information to make a wide range of educational decisions. We conclude with a set of questions to guide thinking for the future on both research and policy fronts. We recognize that our questions are more dilemmas than they are problems. That does not make them less worthy of our attention; it only complicates our sense of professional efficacy.

*Can assessment prove to be helpful to students of cultural, linguistic, and economic diversity?* If assessment is to fuel the entire reform process and improve the opportunities provided to students traditionally not well served by schools, standards for curriculum and performance must be accompanied by standards that guarantee students the opportunity to learn. In a review of the at-risk literature, García and her colleagues (García et al., in press) found that most of the students who drop out of school are poor or attend schools with a disproportionate number of poor students. Investment in assessment without investment in resources will not improve the instructional opportunities offered to students of poverty. In fact, one without the other will do nothing but exacerbate current inequalities. Wolf, LeMahieu, and Eresh (1992) point out that fair and equitable use of performance assessment requires that three different types of standards be made public: content standards, or what the students should know; performance standards, or how well the students should know the content; and delivery standards, or what must be provided to ensure that all students have access to the knowledge and opportunity to learn required to meet the content and performance standards.

*Can assessments, even those that claim neutrality or universality, privilege anything other than mainstream culture?* Metaphors of equality, such as a level playing field or a common yardstick, are common in discussions of equity issues. Yet, the one-size-fits-all approach is likely to perpetuate the differences in academic performance that we find in indices such as dropout rates (National Center for Education Statistics, 1988), SAT scores (Pennock-Román, 1990; Durán, 1983), and NAEP standards attainment (Applebee et al., 1987; Mullis &

Jenkins, 1990). Put differently, the level playing field approach establishes one kind of equity (students perform the same task under the same conditions) while allowing another kind of equity (the opportunity to perform familiar tasks in familiar contexts) to vary dramatically. Ironically, in the few documented instances in which assessment tools that recognize, acknowledge, and value cultural and linguistic diversity (García, 1991; Goodman et al., 1989; Mercer, 1979) have been used, a very different picture of students' capabilities has been produced. If we want to establish the second type of equity, an equity in which all students have the opportunity to put their "best foot forward," choice—choice of passages to read, questions to answer, prompts to write to, projects to complete, or even sociolinguistic contexts in which to work—may become our primary tool. Potentially, authentic classroom assessment and performance assessment both can be open and flexible, allowing for diverse ways of solving problems and accomplishing tasks. Whether this will actually occur is uncertain.

*Can we learn to use new reference points to evaluate performance?* Formal assessments traditionally have evaluated individual students' work by comparing it with that of others. As a result, only a small percentage of students excel, the large majority are average, and the rest are below average. The end result is that educational personnel are given very little diagnostic information about student performance. In addition, everyone who participates in the assessment must be judged by the same criteria. If assessment is to be useful to educational personnel, not only for evaluation purposes but for instructional purposes, then it needs to focus on what students can do and not just on how well they do relative to other students.

Few would question the assertion that norms have dominated our thinking in regard to reference points that we use for attaching significance to observations. We are, as a profession and as a nation, seemingly compelled to compare the performance of an individual, a school, a district, and even a state or our nation with the performance of some identifiable reference group. Given the history of assessment in this country, the nature of the testing industry, and the difficulty we experience in trying to escape normative thinking, we must ask whether reform based on assessments that are referenced to standards rather than norms will ever work. Will we ever be able to change our basic assessment question from "How well did Jorge do in comparison with the rest of the school?" to "What can he or, for that matter, everyone in the school do well?"

*Can we provide the professional development needed to ensure that assessment tools are used fairly and appropriately?* Our review of formal assessment suggests that even the best intentioned of us violate basic standards of test interpretation and use; these violations have had a particularly pernicious effect on students of diversity. To those who suggest alternative assessment as a solution to the problem, we would emphasize the simple fact that the knowledge requirements of alternative assessments are even greater than those of formal assessment. Scoring a performance of student work, like scoring a performance in athletic and artistic arenas,

requires that experts examine the performance in order to attach a value to it. Human judgment, with all its attendant problems of bias and subjectivity, becomes an inherent part of the assessment process when scores are referenced to standards rather than norms. Alternative assessments, with such a premium placed on teacher judgment, make sense only under the assumption that high levels of professional knowledge—about subject matter, language, culture, and assessment—are widely distributed in the profession. Thus, the implications for professional development are very serious: Whenever and wherever that assumption is shaky, substantial investments in staff development will be necessary.

*Can we come to terms with the social nature of alternative assessments?* If assessment is to reward students for a variety of problem-solving behaviors, then it also needs to reflect the social nature of learning and work. However, the social nature of most performance assessments creates a dilemma for educators: The work that particular students submit for review and evaluation is not solely their own. How, then, are we to make judgments about who deserves to gain access to certain programs and who does not? On the other hand, looked at from the lens of the world of work, it may not matter that students have received help; in fact, it may provide a more valid prediction. If we are to believe the rhetoric of the modern industry and total quality management, individuals may never be required to complete tasks completely on their own. Their work milieu, like the performance assessment context, may be inherently social, in which case knowing how they can perform with the assistance of others actually provides a predictive advantage. The solution to this problem may be to make certain that we are always apprised of the context of any assessment. What, if any, assistance was provided? What resources were available? Was the atmosphere so alien, so decontextualized, as to invalidate any conclusions one might want to draw about a student? Even better, we might adopt a practice of securing performance judgments across multiple assessment contexts: alone; in pairs; in groups; in stiff, formal contexts; in relaxed environments. We might even learn something about the optimal assessment contexts for different types of students. Put differently, we might summarize this dilemma by asking, To what extent is scaffolding in the form of teacher instruction and peer interaction and cooperative learning a legitimate and expected part of the assessment process?

*Can assessments that focus on outcomes ever provide teachers with instructionally useful information?* Standardized tests have been criticized for years for failing to provide teachers with information that they can use to plan instruction for individuals or classes of students. In fact, this criticism has led some testing companies to attempt to provide subscale scores so that teachers have a profile of student performance rather than a single outcome. The truth of the matter, however, is that single norm-referenced scores tend to dominate reporting practices in our schools.

A critical question concerning authentic classroom assessment and performance assessment is whether they will provide information that is any more useful in

planning instruction for groups or individuals. We are not sure. In performance assessment, for example, there is an inherent tension between holistic scoring and dimensional or featural scoring. Holistic scoring has the virtue of avoiding the decomposition fallacy: the mistaken idea that by breaking an integrated performance into component processes, each can be evaluated independently. Short of implying that teachers should provide even more opportunities for students to perform cognitively challenging tasks, holistic scores on performance tasks give teachers little direction about how to provide helpful instruction for those who do not achieve high standards. Will teaching to a "good" product be any better than teaching to a bad one?

Teachers need much more fine-grained information in order to plan instruction sensibly. Dimensional scoring systems, in which a writing teacher, for example, might have scores for voice, audience awareness, style, organization, and mechanics, have the virtue of providing more fine-grained information. But the cost of that information is the possibility that it will prompt teachers to provide excessively narrow, decontextualized, and unintegrated instruction on each dimension. Ironically, for students of poverty, this is exactly the type of instruction that the reform movements are trying to eradicate.

*Can a single assessment, or even a system of assessments, meet the needs of multiple audiences and constituencies?* Several educators have suggested that multiple assessment systems, in which different types of assessment are used for different purposes (Chapman, 1993; Mitchell, 1992), might be the best resolution for some of the daunting problems we face. The logic seems to be that almost any assessment can prove useful, as long as we do not try to adapt it to a purpose for which it was never intended (Calfee & Hiebert, 1991). Mitchell proposes a multi-indicator system that "includes a variety of evidence about the progress of individual students and the quality of educational programs" (p. 19) to help educators satisfy different purposes and constituencies: student selection and sorting, accountability to political authorities (school boards and legislators), program monitoring, and instructional improvement/curriculum reform. Her system includes assessments from each of the main categories (formal assessment, authentic classroom assessment, and performance assessment) we have discussed in this chapter.

In an earlier paper (García & Pearson, 1991b), we championed this approach ourselves. We even proposed specific types of assessment for different levels of aggregation and informational needs: (a) NAEP-like trend data as a barometer of the national and state scenes; (b) responsive evaluational approaches, including site visits by outside experts and portfolios of student and teacher work for school and district program evaluations; and (c) authentic classroom assessment for the information needed by teachers, students, and parents. In truth, however, our experience in conducting this review—reading tale after tale of the adverse effects of assessment, particularly formal assessment, on students of cultural, linguistic, and economic diversity—coupled with our personal experiences in the politics

of state assessment, leads us to wonder whether there is any role for formal assessment in the highly politicized context of our current educational system. Nothing inherent in formal tests, particularly recently developed tests with strong grounding in current learning theory, renders them inappropriate for certain functions, such as wide-scale trend monitoring. However, we fear that as long as externally imposed assessments are a "part" of any system, formal assessments will be politically privileged over all other forms of assessment within that system.

*Will we ever be able to address these dilemmas?* These questions raise dilemmas rather than problems. Any solution we find is likely to raise even more problems. So the best we can hope for is to decide which problems we are willing to live with in the process of solving those we believe are intolerable.

The issue of privilege brings us to this most central of dilemmas, one that we must all, both the testers and the tested, come to terms with: At every level of analysis, assessment is a political act. Assessments tell people how they should value themselves and others. They open doors for some and close them for others. The very act of giving an assessment is a demonstration of power: One individual tells the other what to read, how to respond, how much time to take. One insinuates a sense of greater power because of greater knowledge (i.e., possession of the correct answers). The political dilemma is a problem for all students, but it is particularly acute for students from diverse cultural, linguistic, and economic backgrounds whose cultures, languages, and identities have been at best ignored and at worst betrayed in the assessment process.

The brightest ray of hope emanating from our recent candidates for assessment reform is their public disposition. If assessment becomes a completely open process in all of its phases from conception to development to interpretation, then at least the hidden biases will become more visible and at best everyone will have a clearer sense of what counts in our schools.

## REFERENCES

Allington, R. L., & McGill-Franzen, A. (1989). School response to reading failure: Instruction for Chapter 1 and special education students in grades two, four and eight. *Elementary School Journal, 89,* 530–542.

Allington, R. L., Stuetzel, H., Shake, M., & Lamarche, S. (1986). What is remedial reading? A descriptive study. *Reading Research and Instruction, 26*(1), 15–30.

American Psychological Association, American Educational Research Association, and National Council on Measurement in Education. (1985). *The standards for educational and psychological testing.* Washington, DC: American Psychological Association.

Applebee, A. N., Langer, J. A., & Mullis, I. V. S. (1987). *The nation's report card: Learning to be literate in America: Reading.* Princeton, NJ: Educational Testing Service.

Atwell, N. (1987). *In the middle: Writing, reading, and learning with adolescents.* Portsmouth, NH: Heinemann.

Au, K. H. (1981). Participation structures in a reading lesson with Hawaiian children: Analysis of a culturally appropriate instructional event. *Anthropology and Education Quarterly, 11,* 91–115.

Au, K. H. (1993a). *Literacy instruction in multicultural settings.* Fort Worth, TX: Harcourt Brace Jovanich.

Au, K. H. (1993b, July). *Portfolio implementation in the Kamehameha Elementary Education Program: A progress report.* Paper presented to the Board of the Standards Project for English Language Arts, Snowbird, UT.

Au, K. H. (in press). Portfolio assessment: Experiences at the Kamehameha Elementary Education Program. In E. Hiebert, P. Afflerbach, & S. Valencia (Eds.), *Authentic reading assessment: Practices and possibilities.* Newark, DE: International Reading Association.

Au, K. H., & Jordan, C. (1982). Teaching reading to Hawaiian children: Finding a culturally appropriate solution. In H. T. Trueba, G. P. Guthrie, & K. H. Au (Eds.), *Culture in the bilingual classroom: Studies in classroom ethnography* (pp. 153–177). Rowley, MA: Newbury House.

Au, K. H., & Mason, J. (1981). Social organization factors in learning to read: The balance of rights hypothesis. *Reading Research Quarterly, 17,* 115–152.

August, D., & García, E. E. (1988). *Language minority education in the United States: Research, policy, and practice.* Springfield, IL: Charles C Thomas.

Banks, J., & Banks, C. M. (Eds.). (1993). *Multicultural education: Issues and perspectives* (2nd ed.). Boston: Allyn & Bacon.

Bereiter, C., & Englemann, S. (1966). *Teaching disadvantaged children in the preschool.* Englewood Cliffs, NJ: Prentice-Hall.

Bird, L. B. (1989). The art of teaching: Evaluation and revision. In K. S. Goodman, Y. M. Goodman, & W. J. Hood (Eds.), *The whole language evaluation book* (pp. 15–24). Portsmouth, NH: Heinemann.

Boggs, S. T. (1972). The meaning of questions and narratives to Hawaiian children. In C. B. Cazden, V. P. John, & D. Hymes (Eds.), *Functions of language in the classroom* (pp. 299–327). New York: Teachers College Press.

Boggs, S. T. (1978). The development of verbal disputing in part-Hawaiian children. *Language in Society, 7,* 325–344.

Brandt, R. (1992). Overview: A fresh focus for curriculum. *Educational Leadership, 49*(8), 7.

Bruce, B., Rubin, A., Starr, K., & Liebling, C. (1984). Sociocultural differences in oral vocabulary and reading material. In W. S. Hall, W. E. Nagy, & R. Linn (Eds.), *Spoken words: Effects of situation and social group on oral word use and frequency* (pp. 466–480). Hillsdale, NJ: Erlbaum.

Cabello, B. (1984). Cultural interface in reading comprehension: An alternative explanation. *Bilingual Review, 2,* 12–20.

Calfee, R., & Hiebert, E. (1991). Classroom assessment of reading. In R. Barr, M. L. Kamil, P. Mosenthal, & P. D. Pearson (Eds.), *The handbook of reading research* (Vol. 2, pp. 281–309). New York: Longman.

Calfee, R., & Perfumo, P. (1993). Student portfolios: Opportunities for revolution in assessment. *Journal of Reading, 36,* 532–537.

Cambourne, B., & Turbill, J. (1990). Assessment in whole-language classrooms: Theory into practice. *Elementary School Journal, 90,* 337–349.

Catell, R. B., & Catell, A. K. S. (1949–1963). *IPAT Culture Fair Intelligence Test: Scales I, II, III.* Champaign, IL: Institute for Personality and Ability Testing.

Cazden, C. B. (1988). *Classroom discourse: The language of teaching and learning.* Portsmouth, NH: Heinemann.

Center for the Study of Testing, Evaluation, and Educational Policy. (1992). *The influence of testing on teaching math and science in Grades 4–12* (Vols. 1–5). Boston: Boston College, Center for the Study of Testing, Evaluation, and Educational Policy.

Chachkin, N. J. (1989). Testing in elementary and secondary schools: Can miscue be avoided? In B. Gifford (Ed.), *Test policy and the politics of opportunity allocation: The workplace and the law* (pp. 163–187). Boston: Kluwer Academic Publishers.

Chamot, A. U. (1980, November). Recent research on second-language reading. *National Association of Bilingual Education (NABE) Forum,* pp. 3–4.

Chapman, C. (1993, June). *A statewide partnership to develop a multiple assessment system.* Paper presented at the Large Scale Assessment Conference, Albuquerque, NM.

Cicourel, A. (1974). Some basic theoretical issues in the assessment of the child's performance in testing and classroom settings. In A. Cicourel, K. H. Jennings, S. H. M. Jennings, K. C. W. Leiter, R. Mackay, H. Mehan, & D. Roth (Eds.), *Language use and school performance* (pp. 300–351). New York: Academic Press.

Clarke, M. A. (1979). The short-circuit hypothesis of ESL reading—Or when language competence interferes with reading performance. *Modern Language Journal, 64,* 203–209.

Cole, N. S., & Moss, P. A. (1989). Bias in test use. In R. Linn (Ed.), *Educational measurement* (3rd ed., pp. 201–219). New York: American Council on Education.

Collins, J. (1982). Discourse style, classroom interaction and differential treatment. *Journal of Reading Behavior, 14,* 429–437.

Cummins, J. (1984). *Bilingualism and special education: Issues in assessment and pedagogy.* Clevedon, England: Multilingual Matters.

Cunningham, P. M. (1976–1977). Teachers' correction responses to Black-dialect miscues which are non-meaning changing. *Reading Research Quarterly, 12,* 637–653.

Cziko, G. A. (1978). Differences in first- and second language reading: The use of syntactic, semantic, and discourse constraints. *Canadian Modern Language Review, 34,* 473–489.

Dalrymple, K. S. (1989). Well, what about his skills? Evaluation of whole language in middle school. In K. S. Goodman, Y. M. Goodman, & W. J. Hood (Eds.), *The whole language evaluation book* (pp. 111–130). Portsmouth, NH: Heinemann.

Darling-Hammond, L., & Ancess, J. (in press). Authentic assessment and school development. In J. B. Baron & D. P. Wolf (Eds.), *The ninety-third yearbook of the National Society for the Study of Education.* Chicago: National Society for the Study of Education.

Darling-Hammond, L., Ancess, J., & Falk, B. (in press). *Authentic assessment in action.* New York: National Center for Restructuring Education, Schools, and Teaching, Teachers College, Columbia University.

Darling-Hammond, L., & Ascher, C. (1990). *Accountability in big city schools.* New York: National Center for Restructuring Education, Schools, and Teaching and Institute for Urban and Minority Education, Teachers College, Columbia University.

Darling-Hammond, L., & Goodwin, L. (1993). Progress toward professionalism in teaching. In G. Kawelti (Ed.), *Challenges and achievements of American education* (pp. 19–52). Alexandria, VA: Association for Supervision and Curriculum Development.

Davis, A. (1951). What are some of the basic issues in the relation of intelligence tests to cultural background? In K. Eells, A. Davis, R. J. Havighurst, V. E. Herrick, & R. W. Tyler (Eds.), *Intelligence and cultural differences: A study of cultural learning and problem-solving* (pp. 22–28). Chicago: University of Chicago Press.

Davis, A., & Eells, K. (1953). *Davis-Eells Test of General Intelligence or Problem-Solving Ability, Grades 1–2, 3–6.* Yonkers-on-the-Hudson, NY: World Book.

Delgado-Gaitan, C. (1987). Traditions and transitions in the learning process of Mexican children: An ethnographic view. In G. Spindler & L. Spindler (Eds.), *Interpretive ethnography of education: At home and abroad* (pp. 333–359). Hillsdale, NJ: Erlbaum.

Delpit, L. D. (1988). The silenced dialogue: Power and pedagogy in educating other people's children. *Harvard Educational Review, 58,* 280–298.

Dewitz, P., Carr, E. M., Palm, K. N., & Spencer, M. (1992). The validity and utility of portfolio assessment. In C. K. Kinzer & D. J. Leu (Eds.), *The forty-first yearbook of the National Reading Conference: Literacy research, theory, and practice: Views from many perspectives* (pp. 153–160). Chicago: National Reading Conference.

Dorr-Bremme, D. W., & Herman, J. L. (1986). *Assessing student achievement: A profile of classroom practices.* Los Angeles: University of California, Center for the Study of Evaluation.

Downing, J. (1984). A source of cognitive confusion for beginning readers: Learning in a second language. *The Reading Teacher, 37,* 366–370.

Durán, R. P. (1983). *Hispanics' education and background: Predictors of college achievement.* New York: College Entrance Examination Board.

Eaton, A. J. (1980). A psycholinguistic analysis of the oral reading miscues of selected field-dependent and field-independent native Spanish-speaking, Mexican-American, first-grade children. In *Outstanding dissertations in bilingual education* (pp. 71–86). Rosslyn, VA: National Clearinghouse for Bilingual Resources.

Edelsky, C., & Harman, S. (1988). One more critique of reading tests—With two differences. *English Education, 20,* 157–171.

Education for All Handicapped Children Act (Public Law 94–142), 20 U.S.C. § 1401 (1975).

Educational Testing Service. (1988). *A summary of data collected from Graduate Record Examinations test takers during 1986–1987, Data Summary Report #12.* Newark, NJ: Author.

Eells, K. (1951). What is the problem? In K. Eells, A. Davis, R. J. Havighurst, V. E. Herrick, & R. W. Tyler (Eds.), *Intelligence and cultural differences: A study of cultural learning and problem-solving* (pp. 3–9). Chicago: University of Chicago Press.

Eells, K. (1953). Some implications for school practice of the Chicago studies of cultural bias in intelligence tests. *Harvard Educational Review, 23,* 284–297.

Erickson, F., & Mohatt, G. (1982). Cultural organization of participation structures in two classrooms of Indian students. In G. Spindler (Ed.), *Doing the ethnography of schooling: Educational anthropology in action* (pp. 132–174). New York: Holt, Rinehart & Winston.

Feuerstein, R. (1979). *The dynamic assessment of retarded performers: The learning potential assessment device, theory, instruments, and techniques.* Glenview, IL: Scott, Foresman.

Figueroa, R. A. (1989). Psychological testing of linguistic-minority students: Knowledge gaps and regulations. *Exceptional Children, 56,* 145–152.

Flood, J., Lapp, D., & Monken, S. (1992). Portfolio assessment: Teachers' beliefs and practices. In C. K. Kinzer & D. J. Leu (Eds.), *The forty-first yearbook of the National Reading Conference: Literacy research, theory, and practice: Views from many perspectives* (pp. 119–127). Chicago: National Reading Conference.

Foster, M. (1990). The politics of race: Through the eyes of African-American teachers. *Journal of Education, 172,* 123–141.

García, G. E. (1991). Factors influencing the English reading test performance of Spanish-speaking Hispanic students. *Reading Research Quarterly, 26,* 371–392.

García, G. E. (1992a). Ethnograpy and classroom communication: Taking an emic perspective. *Topics in Language Disorders, 12*(3), 54–66.

García, G. E. (1992b). *The literacy assessment of second-language learners* (Tech. Rep. No. 559). Urbana-Champaign: University of Illinois, Center for the Study of Reading.

García, G. E., & Pearson, P. D. (1991a). Modifying reading instruction to maximize its effectiveness for all students. In M. S. Knapp & P. M. Shields (Eds.), *Better schooling for the children of poverty: Alternatives to conventional wisdom* (pp. 31–59). Berkeley, CA: McCutchan.

García, G. E., & Pearson, P. D. (1991b). The role of assessment in a diverse society. In E. Hiebert (Ed.), *Literacy for a diverse society: Perspectives, practices, and policies* (pp. 253–278). New York: Teachers College Press.

García, G. E., Pearson, P. D., & Jiménez, R. (in press). *The at-risk dilemma: A synthesis of reading research.* Urbana-Champaign: University of Illinois, Center for the Study of Reading.

García, G. E., Stephens, D. L., Koenke, K. R., Pearson, P. D., Harris, V. J., & Jiménez, R. T. (1989). *A study of classroom practices related to the reading of low-achieving students: Phase one* (Study 2.2.3.5). Urbana: University of Illinois, Reading Research and Education Center.

Gardner, H. (1985). *The mind's new science.* New York: Basic Books.

Gee, J. P. (1990). *Social linguistics and literacies: Ideologies in discourses.* Bristol, PA: Falmer.

Geisinger, K. F. (Ed.). (1992). Fairness and psychometric issues. In K. F. Geisinger (Ed.), *Psychological testing of Hispanics* (pp. 17–42). Washington, DC: American Psychological Association.

Geneshi, C. (1985). Observing communicative performance in young children. In A. Jagger & M. T. Smith-Burke (Eds.), *Observing the language learner* (pp. 131–142). Newark, DE: International Reading Association.

Gifford, B. R. (1989). The allocation of opportunities and the politics of testing: A policy analytic perspective. In B. Gifford (Ed.), *Test policy and the politics of opportunity allocation: The workplace and the law* (pp. 3–32). Boston: Kluwer Academic Publishers.

Ginsburg, H., & Allardice, B. (1984). Children's difficulties with school mathematics. In B. Rogoff & J. Lave (Eds.), *Everyday cognition: Its development in social context* (pp. 194–219). Cambridge, MA: Harvard University Press.

Goodman, K. S., Goodman, Y. M., & Hood, W. J. (1989). *The whole language evaluation book.* Portsmouth, NH: Heinemann.

Gordon, E. W., & Bhattacharyya, M. (1992). Human diversity, cultural hegemony, and the integrity of the academic canon. *Journal of Negro Education, 61,* 405–418.

Greene, M. (1993). The passions of pluralism: Multiculturalism and the expanding community. *Educational Researcher, 22*(1), 13–18.

Haladyna, T. M., Nolan, S. B., & Haas, N. S. (1991). Raising standardized achievement test scores and the origins of test score pollution. *Educational Researcher, 20*(5), 2–7.

Harp, W. (1991). Principles of assessment and evaluation in whole language classrooms. In W. Harp (Ed.), *Assessment and evaluation in whole language programs* (pp. 35–50). Norwood, MA: Christopher Gordon.

Heath, S. B. (1981). What no bedtime story means: Narrative skills at home and school. *Language in Society, 11,* 49–76.

Heath, S. B. (1982). Questioning at school and at home: A comparative study. In G. Spindler (Ed.), *Doing the ethnography of schooling: Educational anthropology in action* (pp. 102–131). New York: Holt, Rinehart & Winston.

Heath, S. B. (1983). *Ways with words: Language, life, and work in communities and classrooms.* Cambridge, England: Cambridge University Press.

Herman, J. L., & Golan, S. (n.d.). *The effects of standardized testing on teaching in schools.* Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing, UCLA Graduate School of Education.

Hernandez, H. (1989). *Multicultural education: A teacher's guide to content and process.* Columbus, OH: Merrill.

Hiebert, E. H., & Corley, R. (1993). *A comparison of students' reading on standardized and performance assessments in a high-stakes testing context.* Unpublished manuscript, University of Colorado.

Hirsch, E. D. (1987). *Cultural literacy: What every American needs to know.* Boston: Houghton Mifflin.

Hood, W. J. (1989). If the teacher comes over, pretend it's a telescope. In K. S. Goodman, Y. M. Goodman, & W. J. Hood (Eds.), *The whole language evaluation book* (pp. 27–42). Portsmouth, NH: Heinemann.

Hymes, D. (1972). Introduction. In C. B. Cazden, V. P. John, & D. Hymes (Eds.), *Functions of language in the classroom* (pp. xi-lvii). New York: Teachers College Press.

Jiménez, R. T. (1992). *Opportunities and obstacles in bilingual reading.* Unpublished doctoral dissertation, University of Illinois, Urbana-Champaign.

Jiménez, R. T., García, G. E., & Pearson, P. D. (1991, December). *The strategic reading processes of bilingual Hispanic children who are good readers.* Paper presented at the National Reading Conference, Miami, FL.

Johnston, P. (1984a). Prior knowledge and reading comprehension test bias. *Reading Research Quarterly, 19,* 219–239.

Johnston, P. (1984b). *Reading comprehension assessment: A cognitive basis.* Newark, DE: International Reading Association.

Johnston, P., Guice, S., Baker, K., Malone, J., & Michelson, N. (1993, April). *Assessment of teaching and learning in "literature-based" classrooms.* Paper presented at the annual meeting of the American Educational Research Association, Atlanta.

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

Kapinus, B. A., Collier, G. V., & Kruglanski, H. (in press). Maryland School Performance Assessment Program: A new view of assessment. In S. W. Valencia, E. H. Hiebert, & P. Afflerbach (Eds.), *Authentic reading assessment: Practices and possibilities.* Newark, DE: International Reading Association.

Karier, C. J. (1973a). Business values and the educational state. In C. J. Karier, P. Violas, & J. Spring (Eds.), *Roots of crisis: American education in the twentieth century* (pp. 6–29). Chicago: Rand McNally.

Karier, C. J. (1973b). Testing for order and control in the corporate liberal state. *Educational Theory, 22,* 159–180.

Kentucky Department of Education. (1992, August). Teaching and assessing reading. *Education News,* pp. S8, S10, S12.

Kochman, T. (1989). Black and White cultural styles in pluralistic perspective. In B. Gifford (Ed.), *Test policy and test performance: Education, language and culture* (pp. 259–296). Boston: Kluwer Academic Publishers.

Koogler, C. C. (1980). Behavioral style differences and crisis in an integrated kindergarten classroom. *Contemporary Education, 51,* 126–130.

Koretz, D., Stecher, B., & Deibert, E. (1992). *The Vermont Portfolio Assessment Program: Interim report on implementation and impact, 1991–92 school year* (Tech. Rep. No. 350). Los Angeles: UCLA, Center for the Study of Evaluation.

Koretz, D., Stecher, B., & Deibert, E. (1993). *The reliability of scores from the 1992 Vermont Portfolio Assessment Program* (Tech. Rep. No. 355). Los Angeles: UCLA, Center for the Study of Evaluation.

Labov, W. (1969). The logic of nonstandard English. In R. D. Abrahams & R. C. Troike (Eds.), *Language and cultural diversity in American education* (pp. 225–261). Englewood Cliffs, NJ: Prentice-Hall.

Ladson-Billings, G. (1992). Liberatory consequences of literacy: A case of culturally relevant instruction for African-American students. *Journal of Negro Education, 61,* 378–391.

Langer, J. A. (1987). The construction of meaning and the assessment of comprehension: An analysis of reader performance on standardized test items. In R. O. Freedle & R.

P. Durán (Eds.), *Cognitive and linguistic analyses of test performance* (pp. 225–244). Norwood, NJ: Ablex.

Lee, J. F. (1986). On the use of the recall task to measure L2 reading comprehension. *Studies in Second Language Acquisition, 8,* 201–211.

Linn, R. L. (1983). Predictive bias as an artifact of selection procedures. In H. Wainer & S. Messick (Eds.), *Principles of modern psychological measurement: A festschrift for Frederic M. Lord* (pp. 27–40). Hillsdale, NJ: Erlbaum.

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

Mabry, L. (1992, December). Twenty years of alternative assessment at a Wisconsin high school. *The School Administrator,* pp. 12–13.

Madaus, G. F., & Tan, A. G. A. (1993). The growth of assessment. In G. Cawelti (Ed.), *Challenges and achievements of American education: The 1993 ASCD yearbook* (pp. 53–79). Alexandria, VA: Association for Supervision and Curriculum Development.

Madden, N. A., & Slavin, R. E. (1987, April). *Effective pull-out programs for students at risk.* Paper presented at the annual meeting of the American Educational Research Association, Washington, DC.

Mägiste, E. (1979). The competing language systems of the multilingual: A developmental study of decoding and encoding processes. *Journal of Verbal Learning and Verbal Behavior, 18,* 79–89.

McGill-Franzen, A. (1987). Failure to learn to read: Formulating a policy problem. *Reading Research Quarterly, 22,* 475–490.

McGill-Franzen, A., & Allington, R. L. (1993). Flunk 'em or get them classified: The contamination of primary grade accountability data. *Educational Researcher, 22*(1), 19–22.

Mehan, H. (1979). *Learning lessons.* Cambridge, MA: Harvard University Press.

Mercer, J. (1979). *SOMPA: Technical and conceptual manual.* New York: Psychological Corporation.

Mercer, J. (1989). Alternative paradigms for assessment in a pluralistic society. In J. A. Banks & C. A. M. Banks (Eds.), *Multicultural education: Issues and perspectives* (pp. 289–304). Boston: Allyn & Bacon.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: American Council on Education.

Mestre, J. P. (1984). The problem with problems: Hispanic students and math. *Bilingual Journal, 32,* 15–19.

Michaels, S. (1981). Sharing time: Children's narrative styles and differential access to literacy. *Language in Society, 10,* 423–442.

Mitchell, R. (1992). *Testing for learning: How new approaches to evaluation can improve American schools.* New York: Free Press.

Moll, L. C., Estrada, E., Diaz, E., & Lopes, L. M. (1980). The organization of bilingual lessons: Implications for schooling. *Quarterly Newsletter of the Laboratory of Comparative Human Cognition, 2,* 53–58.

Moore, E. G. J. (1989). Ethnic group differences in the Armed Services Vocational Aptitude Battery (ASVAB) performance of American youth: Implications for career prospects. In B. Gifford (Ed.), *Test policy and test performance: Education, language and culture* (pp. 183–229). Boston: Kluwer Academic Publishers.

Morrow, L. M. (1989). Using story retelling to develop comprehension. In K. D. Muth (Ed.), *Children's comprehension of text: Research into practice* (pp. 37–58). Newark, DE: International Reading Association.

Mullis, I. V. S., & Jenkins, L. B. (1990). *The reading report card, 1971–88: Trends from the nation's report card.* Princeton, NJ: National Assessment of Educational Progress, Educational Testing Service.

Murphy, S., & Smith, M. A. (1991). *Writing portfolios: A bridge from teaching to assessment.* Markham, Ontario, Canada: Pippin.

Muth, D. K. (1989). *Children's comprehension of text: Research into practice.* Newark, DE: International Reading Association.

National Board for Professional Teaching Standards. (1990). *Initial policies and perspectives of the National Board for Professional Teaching Standards.* Detroit, MI: Author.

National Center for Education Statistics. (1988). *Education indicators.* Washington, DC: U.S. Department of Education, Office of Educational Research and Improvement.

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

New Standards Project. (1993). *Harriet Tubman (4th grade pilot assessment task).* Pittsburgh, PA: University of Pittsburgh, Learning Research and Development Center.

Oakland, T., & Matuszek, P. (1977). Using tests in nondiscriminatory assessment. In T. Oakland (Ed.), *Psychological and educational assessment of minority children* (pp. 52–69). New York: Brunner/Mazel.

O'Connor, M. C. (1989). Aspects of differential performance by minorities on standardized tests: Linguistic and sociocultural factors. In B. Gifford (Ed.), *Test policy and test performance: Education, language and culture* (pp. 129–189). Boston: Kluwer Academic Publishers.

Olmeda, E. L. (1981). Testing linguistic minorities. *American Psychologist, 36,* 1078–1085.

Padilla, A. (1979). Critical factors in the testing of Hispanics: A review and some suggestions for the future. In National Institute of Education, *Testing, teaching, and learning: Report of a conference on research in testing* (pp. 219–244). Washington, DC: U.S. Government Printing Office.

Paris, B. G., Lawton, T. A., & Turner, J. C. (1992). Reforming achievement testing to promote student learning. In C. Collins & J. M. Mangieria (Eds.), *Teaching thinking: An agenda for the 21st century* (pp. 223–241). Hillsdale, NJ: Erlbaum.

Patterson, P. O. (1989). Employment testing and Title VII of the Civil Rights Act of 1964. In B. Gifford (Ed.), *Test policy and the politics of opportunity allocation: The workplace and the law* (pp. 83–120). Boston: Kluwer Academic Publishers.

Pearson, P. D., & Valencia, S. (1987). Assessment, accountability, and professional prerogative. In J. E. Readence & R. S. Baldwin (Eds.), *Research in literacy: Merging perspectives: Thirty-sixth yearbook of the National Reading Conference* (pp. 3–16). Rochester, NY: National Reading Conference.

Pennock-Román, M. (1988). *The status of research on selective admission tests and Hispanic students in post-secondary education* (ETS Research Report No. RR-88-36). Princeton, NJ: Educational Testing Service.

Pennock-Román, M. (1990). *Test validity and language background: A study of Hispanic American students at six universities.* New York: College Entrance Examination Board.

Pennock-Román, M. (1992). Interpreting test performance in selective admissions for Hispanic students. In K. F. Geisinger (Ed.), *Psychological testing of Hispanics* (pp. 99–135). Washington, DC: American Psychological Association.

Perkins, K. (1983). Semantic constructivity in ESL reading and composition. *TESOL Quarterly, 17,* 19–27.

Philips, S. (1972). Participant structures and communicative competence: Warm Springs children in community and classroom. In C. B. Cazden, V. P. John, & D. Hymes (Eds.), *Functions of language in the classroom* (pp. 370–394). New York: Teachers College Press.

Presseisen, B. Z., Smey-Richman, B., & Beyer, F. S. (1992). *Cognitive development through radical change: Restructuring classroom environments for students at risk* (Tech. Rep.). Philadelphia: Research for Better Schools.

Rebell, M. A. (1989). Testing, public policy, and the courts. In B. Gifford (Ed.), *Test policy and the politics of opportunity allocation: The workplace and the law* (pp. 135–162). Boston: Kluwer Academic Publishers.

Resnick, D. P. (1981). Testing in America: A supportive environment. *Phi Delta Kappan, 62*(9), 625–628.

Resnick, L. B. (1989). *Tests as standards of achievement in schools.* Paper presented at the Educational Testing Service Conference, The Uses of Standardized Tests in American Education, New York.

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 Academic Publishers.

Rincón, E. (1980). Test speededness, text anxiety, and test performance: A comparison of Mexican American and Anglo American high school juniors (Doctoral dissertation, University of Texas at Austin, 1979). *Dissertation Abstracts International, 40,* 5772A.

Rodriguez, O. (1992). Introduction to technical and societal issues in the psychological testing of Hispanics. In K. F. Geisinger (Ed.), *Psychological testing of Hispanics* (pp. 11–15). Washington, DC: American Psychological Association.

Rogoff, B., & Lave, J. L. (1984). *Everyday cognition: Its development in social context.* Cambridge, MA: Harvard University Press.

Rosenblatt, L. (1985). Viewpoints: Transaction versus interaction—A terminological rescue operation. *Research in the Teaching of English, 19,* 96–107.

Rothman, R. (1992, October 21). Study confirms 'fears' regarding commercial tests. *Education Week,* pp. 1, 13.

Routman, R. (1988). *Transitions: From literature to literacy.* Portsmouth, NH: Heinemann.

Royer, J. M., & Cunningham, D. J. (1981). On the theory and measurement of reading comprehension. *Contemporary Educational Psychology, 6,* 187–216.

Rueda, R., & García, E. (1992). *A comparative study of teachers' beliefs about reading assessment with Latino language minority students.* Santa Cruz: National Center for Cultural Diversity and Second Language Learning, University of California at Santa Cruz.

Samuda, R. J. (1975). *Psychological testing of American minorities: Issues and consequences.* New York: Harper & Row.

Sanchez, G. (1934). The implications of a basal vocabulary to the measurement of the abilities of bilingual children. *Journal of Social Psychology, 5,* 395–402.

Savignon, S. J. (1983). *Communicative competence: Theory and classroom practice.* Reading, MA: Addison-Wesley.

Schwartz, D. J. (1989). Non-discriminatory use of personnel tests: Conference remarks. In B. Gifford (Ed.), *Test policy and the politics of opportunity allocation: The workplace and the law* (pp. 121–126). Boston: Kluwer Academic Publishers.

Seidner, S. S. (1982). *Issues of language assessment: Foundations and research* (Proceedings of the First Annual Language Assessment Institute, June 17–20, 1981). Springfield: Illinois State Board of Education.

Shavelson, R. J., Baxter, G. P., & Pine, J. (1992). Performance assessments: Political rhetoric and measurement reality. *Educational Researcher, 21*(4), 22–27.

Shepard, L. (1989). Why we need better tests. *Educational Leadership, 46*(7), 4–9.

Simmons, W., & Resnick, L. (1993). Assessment as the catalyst of school reform. *Educational Leadership, 50*(5), 11–15.

Sizer, T. (1992). *Horace's school: Redesigning the American high school.* Boston: Houghton Mifflin.

Slavin, R. E., Karweit, N. L., & Madden, N. A. (Eds.). (1992). *Effective programs for students at-risk.* Boston: Allyn & Bacon.

Sleeter, C., & Grant, C. (1988). *Making choices for multicultural education: Five approaches to race, class, and gender.* New York: Macmillan.

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

Stake, R. (1976). A theoretical statement of responsive evaluation. *Studies in Educational Evaluation, 2,* 19–22.

Stallman, A. C. (1991). *Learning vocabulary from context: Effects of focusing attention on individual words during reading.* Unpublished doctoral dissertation, University of Illinois, Urbana-Champaign.

Stallman, A. C., & Pearson, P. D. (1990). Formal measures of early literacy. In L. M. Morrow & J. K. Smith (Eds.), *Assessment for instruction in early literacy* (pp. 7–44). Englewood Cliffs, NJ: Prentice-Hall.

Taylor, O. (1977). Sociolinguistic dimension in standardized testing. In M. Saville-Troike (Ed.), *Georgetown University Roundtable on Language and Linguistics* (pp. 257–266). Washington, DC: Georgetown University Press.

Tierney, R. J., Carter, M. A., & Desai, L. E. (1991). *Portfolio assessment in the reading-writing classroom.* Norwood, MA: Christopher Gordon.

Tierney, R. J., Readance, J. E., & Dishner, E. K. (1990). *Reading strategies and practices: A compendium* (3rd ed.). Boston: Allyn & Bacon.

Treisman, U. (1992). Studying students studying calculus: A look at the lives of minority mathematics students in college. *College Mathematics Journal, 23,* 362–372.

Troike, R. C. (1984). SCALP: Social and cultural aspects of language proficiency. In C. Rivera (Ed.), *Language proficiency and academic achievement* (pp. 44–54). Avon, England: Multilingual Matters.

Troike, R. C. (1982). Zeno's paradox and language assessment. In S. S. Seidner (Ed.), *Issues of language assessment: Foundations and research* (Proceedings of the First Annual Language Assessment Institute, June 17–20, 1981, pp. 3–9). Springfield: Illinois State Board of Education.

Tsang, C. L. (1989). Informal assessment of Asian Americans: A cultural and linguistic mismatch? In B. Gifford (Ed.), *Test policy and test performance: Education, language and culture* (pp. 231–254). Boston: Kluwer Academic Publishers.

Tyler, R. W., & White, S. H. (1979). Chairmen's report. In National Institute of Education, *Testing, teaching, and learning: Report of a conference on research in testing* (pp. 3–32). Washington, DC: U.S. Government Printing Office.

U.S. General Accounting Office. (1993). *Student testing: Current extent and expenditures, with cost estimates for national examination.* Washington, DC: U.S. General Accounting Office, Program Evaluations and Methodology Division.

Valencia, S. W. (1990). A portfolio approach to classroom reading assessment: The whys, whats, and hows. *Reading Teacher, 43,* 338–340.

Valencia, S. W., Hiebert, E., & Kapinus, B. (1992). National Assessment of Educational Progress: What do we know and what lies ahead? *Reading Teacher, 45,* 730–734.

Valencia, S., & Pearson, P. D. (1987). Reading assessment: Time for a change. *Reading Teacher, 40,* 726–732.

Valencia, S. W., & Place, N. (in press). Literacy portfolios for teaching, learning, and accountability: The Bellevue Literacy Assessment Project. In E. Hiebert, P. Afflerbach, & S. Valencia (Eds.), *Authentic reading assessment: Practices and possibilities.* Newark, DE: International Reading Association.

Weiss, B. (in press). California's new English-language arts assessment. In S. W. Valencia, E. H. Hiebert, & P. Afflerbach (Eds.), *Authentic reading assessment: Practices and possibilities.* Newark, DE: International Reading Association.

Wiggins, G. (1989). Teaching to the (authentic) test. *Educational Leadership, 46*(7), 41–47.

Wiggins, G. (1992). Creating tests worth taking. *Educational Leadership, 49*(8), 26–33.

Williams, R. L. (1974). The silent mugging of the Black community. *Psychology Today, 7,* 32, 34, 37, 38, 41, 101.

Williams, R. L. (1975). The BITCH-100: A culture-specific test. *Journal of Afro-American Issues, 3,* 103–116.

Wixson, K. K., Peeters, C. W., Weber, E. M., & Roeber, E. D. (1987). New directions in statewide reading assessment. *Reading Teacher, 40,* 749–754.

Wolf, D. P., LeMahieu, P. G., & Eresh, J. (1992). Good measure: Assessment as a tool for educational reform. *Educational Leadership, 49*(8), 8–13.

Manuscript Received April 2, 1993
Revision Received August 10, 1993
Accepted August 12, 1993