CHAPTER EIGHT

Assessment and Evaluation

CHAPTER OUTLINE

One of our reviewers proposed that we start this chapter with DON'T SKIP THIS CHAPTER EVEN IF YOU HATE TESTING! Assessment plays a most complicated role in the education of English-language learners. Assessment is also a key and sensitive issue for underachieving culturally and linguistically diverse students who are proficient in English. Teachers' and administrators' jobs can be suddenly questioned when new state standards require certain levels of student performance on state-mandated tests. Or a beloved student placed inappropriately in a dead-end program makes a teacher agonize that all the hard work was to no avail. When test results are misused or misunderstood, our first reaction is to angrily protect our students and demand that no testing be conducted. Yet assessment can also serve to assist our students, to assure us that all groups of students are receiving an equal educational opportunity, to identify areas of great need, and to secure additional funding. We educators—teachers and administrators—must understand assessment, keep informed of the constantly changing policy framework, and use assessment well to inform our practice.

Assessment issues continue to be a hot topic in policy debates at federal and state levels, as well as at the local school district level among school board members, administrators, teachers, and parents. Sometimes the policy dialogue incorporates assessment issues for culturally and linguistically diverse students. But too often bilingual/ESL educators are left out of the decision making. To be included, we must build up our expertise on assessment and leap into the debate.

Assessment is a key piece of the school reform movement. It drives the decisions that schools make about the types of programs that are implemented, the levels of resources and support for teachers and students, and the administrative strategies that make a difference in the sociocultural context of schooling. We use the term *assessment* to refer to the measurement of educational progress over time, including the progress of individual students, groups of students, and ultimately the effectiveness of school programs, to enable all students "to use their minds well" (U.S. Department of Education, 1991, p. 3). A test given on one day represents one measure, or one snapshot, of a student's performance at that moment in time. But *assessment* is a more comprehensive term that incorporates many different ways of measuring student growth and school program effectiveness. Assessment is most appropriate when it takes into account multiple and varied measures across time (Wolf, Bixby, Glenn, & Gardner, 1991).

This chapter addresses why and how we assess students of linguistically and culturally diverse heritages, as well as why and how we evaluate school programs that serve these students. Part I provides a policy framework, including an overview of the purposes and potential abuses in assessment practices, and a historical review of U.S. federal government requirements in assessment practices for language minority students. Part II addresses individual student assessment for placement in a program, ongoing classroom assessment, and completion of a school program or exit from a special program. Part III examines how we define student success and program success, as well as the types of assessment tools available to us with their relative strengths and weaknesses, concluding with an analysis of the 1990s standards movement with its "high-stakes" tests and the implications for language minority students.

THE POLICY FRAMEWORK

Purposes of Assessment in Bilingual/ESL Programs

Why do we assess student performance? What tests and other measures do we use? Who demands that we keep track of students' progress? School personnel frequently have to make quick decisions in response to legal mandates or administrative pressures, without the benefit of an evaluation specialist or the time or financial resources to carry out a comprehensive student assessment plan. Teachers complain bitterly that many tests mandated by the school system or the state seem to be inappropriate because they include items with cultural bias or vocabulary or tasks the students have not yet experienced, do not measure what they say they measure, are normed on white middle-class children, do not reflect the cognitive or learning styles of the students, involve a culturally inappropriate testing situation, are not given in the child's primary language, and many other teacher concerns. All of these may be legitimate or not-so-legitimate complaints, depending on the circumstances and purposes for which a test is being used.

Results from formal assessment measures have been widely misused, leading to inappropriate program placement, grade retention, social isolation, and tracking, with devastating consequences for many groups of students (Oakes, 1985; Shepard, 1991a; Wheelock, 1992). On the other hand, school systems that informally allow language minority students to be exempted from taking the mandated tests have brought upon themselves serious legal problems regarding equity and access issues. How can we resolve these assessment dilemmas?

U.S. schools continue to use testing to make life-affecting decisions for us through placement tests, reading and math tests, achievement tests across the curriculum, tests to determine eligibility for special programs, minimum competency tests to graduate from high school, admissions tests for university study, and tests to select who may enter various professions. This is the present reality of our overly test-conscious society. Major reflection and research through the authentic assessment movement of the past two decades has led to more varied measures required by states and local school districts, including performance and portfolio assessment as part of the ongoing assessment process.

But it is clear that we must not eliminate assessment practices required at the district and state level, as many teachers would like to do, because that leads to the underachievement of our students. When there is no comparison of language minority students' performance to that of others outside our school and school system, then as our students leave school, they will find gatekeepers, in the form of formal tests, that keep them from having access to an equal educational opportunity and ultimately to choices for their professional lives as adults. Our role, then, is to understand assessment, to value and benefit from its uses, for our own professional growth as well as that of our students, and to become expert in the development of authentic assessment measures integrated with daily classroom activities. Rather than begrudging the tests used, we can join the dialogue to develop and choose a wide range of assessment measures that help us guide our students to long-term academic success.

Overall, there are five major purposes for assessment in bilingual/ESL settings: entry and placement in a school program, ongoing assessment within that program, completion of a school program or exit from a special program, evaluation of a school program, and accountability, "to guarantee that students attain expected educational goals or standards, including testing for high school graduation" (O'Malley & Valdez Pierce, 1996, p. 3). We shall examine assessment practices for language minority students within each of these five purposes, as well as the policy influences that sometimes play a role in the assessment decisions made by each school district.

Federal Policy: Office for Civil Rights

At the federal level, a major force for implementation of school policies regarding the assessment of language minority students has been the Office for Civil Rights (OCR) within the U.S. Department of Education. In 1970, the OCR issued the first memorandum regarding national origin minority groups with limited English-language skills. (For the complete text of this memorandum, see Chapter Two.) This memorandum established basic federal policy regarding the rights of language minority students to an education, and the Supreme Court judges used it to inform their decision in *Lau* v. *Nichols*. The major points of the 1970 OCR memorandum focused on school districts taking "affirmative steps" to assist students with English-language proficiency development as well as access to the standard curriculum, and monitoring inappropriate placement of students into "educational dead-end or permanent tracks." The recommended OCR identification and assessment procedure for language minority students in general, as well as for students of limited English proficiency in particular, grew out of the guarantees of basic rights provided by Title VI of the Civil Rights Act, in which

> No person in the United States shall, on the grounds of race, color, or national origin, be excluded from participation in, be denied the benefits of, or be subjected to discrimination under any program or activity receiving federal financial assistance. (P.L. 88–352, Section 601, July 2, 1964)

as well as the Equal Educational Opportunities Act of 1974, which states that

> No state shall deny equal educational opportunity to an individual on account of his or her race, color, sex, or national origin, by . . . the failure of an educational agency to take appropriate action to overcome language barriers that impede equal participation by its students in its instructional programs. [20 USC §1703(f)]

This interpretation of all students' rights of access to educational programs, guaranteed by the Civil Rights Act, led to the specific focus on "overcoming language barriers" that "impede equal participation" in instructional programs (Equal Educational Opportunities Act) and thus had a lasting impact on OCR compliance guidelines for assessment of language minority students, which were enforced from the mid-1970s through 1980 and beyond. The legal force of the 1974 *Lau* v. *Nichols* decision reinforced this focus on the assessment of English language proficiency in federal government enforcement guidelines.

It was not until the 1981 *Castañeda* v. *Pickard* decision that this federal court extended the legal interpretation beyond that of English proficiency development to the broader definition originally implied in the 1970 OCR memorandum that students also had the right of access to the standard school curriculum. In this 1981 court decision, schools were asked to assess school program effectiveness across the full curriculum of math, science, social studies, and language arts. Many subsequent court decisions applied the three-part "Castañeda test," asking whether a school program was based on sound educational theory, implemented well, and effective across the curriculum. Following a policy of conducting few OCR compliance reviews under Republican presidents from 1981 through 1992, it was not until the OCR vigilance increased, beginning in 1993, that school districts have been held accountable for providing language minority students access to the full curriculum. This is an important issue that we will continue to expand upon in the remainder of this chapter.

As a result of the initial federal policy emphasis upon English language proficiency as the main "obstacle" to students' access to an equal educational opportunity, during the 1970s many school systems focused only on English proficiency measures to assess placement in and exit from special programs. State legislation followed suit, and many state laws were developed during the 1970s using that focus. Even worse, because the initial federal legislation used the term *non-English-speaking* and *limited-English-speaking* to identify students who needed help, language assessment instruments that were developed in the early 1970s focused only on *oral* language proficiency, ignoring literacy development. During this period, many students were placed in special programs only until their spoken English reached a certain level, at which time they were exited from special services and left to sink or swim in the mainstream.

This policy legacy is still with us today. Policy battles often get sidetracked, focusing on oral English language development as the main goal. When more thoughtful policy makers recognize that reading and writing English is an equally important goal, they often forget about the importance of providing students with uninterrupted access to the full curriculum. Yet access to the full curriculum was the intent of the original 1970 OCR memorandum, and since *Castañeda* v. *Pickard* (1981), it has remained the legal interpretation of school district requirements in each federal court decision since 1981. Current OCR enforcement guidelines emphasize language minority students' access to the full curriculum.

Overall, OCR policies from 1970 to the present regarding language minority students have focused on monitoring local school district assessment practices for the following purposes: (1) identification of language minority students, including those not yet proficient in English as well as those achieving below grade level; (2) language proficiency assessment for each individual student in both the English language and the student's primary language, to determine appropriate educational services; (3) ongoing assessment of student achievement; and (4) if needed, exit criteria when students are placed in a special program separate from the mainstream. When OCR staff visit a chosen school district to conduct a compliance review, initial priorities focus on evaluation of the identification and assessment system for language minority students that is in place. (Other categories for OCR compliance review are listed in Chapter Two.)

Identification

Identification of the target population (all students whose primary home language is other than English) is typically done by means of a home language survey, conducted when each new student first registers at school. The home language survey must be provided in the primary language of the parents, either in written questionnaire form, orally by parent interview, or by student interview, if the student is of middle or high school age. Common questions asked in a home language survey focus on identifying the first language of the student, the student's preferred language use, and family members' use of languages other than English at home. Examples of home language surveys can be found in Linse (1993, pp. 38–39).

L_1 and L_2 Proficiency Assessment

The second step to determining appropriate school program placement is to conduct language proficiency assessment, both in English and in the student's primary language. Historically, many school districts adopted the 1975 *Lau Remedies* classification categories as follows:

A. Monolingual speaker of the language other than English (speaks the language other than English exclusively).
B. Predominantly speaks the language other than English (speaks mostly the language other than English, but speaks some English).
C. Bilingual (speaks both the language other than English and English with equal ease).
D. Predominantly speaks English (speaks mostly English, but speaks some of the language other than English).
E. Monolingual speaker of English (speaks English exclusively). (U.S. Office for Civil Rights, 1975)

Those students who were classified as belonging to categories A and B were entitled to receive some form of ESL assistance. ESL assistance was to be combined with primary language instruction, wherever feasible, in each school district that had at least 20 students of the same primary language background.

Students who were classified as belonging to categories C, D, and E were also entitled to special services such as bilingual schooling if they were achieving below grade level on standardized tests that measured their academic performance across the curriculum. Thus the third step in assessment according to OCR guidelines is to determine students' academic achievement levels through tests that measure their performance in math, science, social studies, and language arts in both English and in their primary language. This third step has not been rigorously enforced until the 1990s, but is increasingly being examined by school districts under pressure from the community.

These general OCR guidelines are still in effect, but are interpreted to allow school districts considerable flexibility as to the specifics of implementation. Individual school districts may have legal consent decrees to follow that mandate criteria for assessment in cases where language minority parents have challenged the school system in the courts. Or state legislative mandates may specify assessment practices for identification, placement, and exit from special programs. With the increasing emphasis upon creative approaches to whole-school and systemwide school reform under current Title VII funding, the old OCR guidelines, interpreted narrowly, can lead to a remedial, compensatory perspective that is not in keeping with the reform movement nationwide. Yet the OCR guidelines do provide an important incentive for school districts to address the needs of language minority students by identifying these students, providing ongoing assessment of their academic achievement, and creating more meaningful and challenging instructional programs to meet all students' needs. When school districts do not respond to the OCR challenge, they are subject to losing all federal funding.

INDIVIDUAL STUDENT ASSESSMENT

Given the policy framework outlined above, with potential pressures from federal and state policies and local court decisions, each school district also makes many assessment policy decisions at the local level. Teachers and administrators are equally important participants in these assessment decisions. Assessment of individual students is an ongoing activity in schools that is an integral part of day-to-day instruction. How do we decide which students should be placed in ESL and bilingual classrooms? What considerations guide teachers' ongoing assessment in a bilingual, ESL, or grade-level class? How can the appropriateness of a student's current placement be assessed? What assessment practices are best for determining exit from a special program? When might special education placement for language minority students be appropriate? What do "intelligence tests" measure and do they help us with assessment decisions? Exploration of local school practices and developments in authentic assessment place the answers to these questions in an ever-changing perspective. We shall begin examination of individual student assessment decisions for language minority students with a brief historical look at older language assessment practices, to contrast them with current authentic language assessment practices. Language assessment is an important part of placement decisions for English-language learners.

Placement Decisions

Language proficiency measures are used with language minority students for many purposes in schools: for placement in special programs, for determining placement level in second-language classes (beginning, intermediate, advanced), for language dominance assessment to determine students' L_1 and L_2 language use for academic purposes, for ongoing assessment of linguistic and cognitive development as students move through school, and for exit from a special program or reclassification. All federal and state guidelines require English-language proficiency testing as one measure to determine language minority students' eligibility for special services. Some states also require L_1 proficiency testing. (For a summary of state assessment policies for language minority students, see O'Malley & Valdez Pierce, 1994; Rivera & Vincent, in press).

Following identification of all students whose primary language is other than English through a home language survey, the next step is assessment to determine who should be in what type of instructional services. One part of this assessment process is language proficiency testing in L_1 and L_2. A common practice in the 1970s and 1980s was to assess each student's language dominance through a test such as the Bilingual Syntax Measure (Burt, Dulay, & Hernández-Chávez, 1975) or the Language Assessment Scales (Duncan & De Avila, 1990, 1994). The concept of language dominance came from the legal interpretations following the *Lau* v. *Nichols* decision that required schools to determine whether the student's stronger language was L_1 or L_2. Students with low proficiency in L_2 (English) were provided bilingual instruction. In recent years, with the reform movement, federal policies have deemphasized language dominance testing, which focuses on prescriptive delivery of services for students to meet their language needs and distracts school personnel from the larger and more important agenda of providing students access to the full academic curriculum. Because many state policies still use language proficiency assessment for this purpose, we will provide a brief discussion here of the limits to the historical use of language dominance testing.

The question of which language is dominant when a student enters a program is really a question of language proficiency. The appropriate method of assessment of language dominance is therefore a test of proficiency in L_1 compared with a similar test of proficiency in L_2. Inappropriate practices in the past have included using students' surname, ethnicity, or parents' self-report as indicators of students' proficiency in L_1 and L_2. The opposite extreme—a language assessment measure that takes into account all domains of language (phonology, morphology, syntax, lexicon, semantics, pragmatics, paralinguistics, and discourse) plus oral and written dimensions of language (listening, speaking, reading, and writing), plus different levels of use of the language depending upon the age of the student and the student's level of cognitive development, plus all the potential contexts for language use—is impractical, time consuming, and too costly ever to carry out.

Instead, to assess language proficiency, we use indirect measures by sampling a student's language use at a given moment in time. From performance on tasks that integrate several linguistic domains across oral and written dimensions of language, a student's relative level of overall proficiency can be inferred, at least for that given moment in time. Although most teachers are not satisfied with a single measure, the initial test can serve placement purposes and can be followed by more extensive ongoing assessment as the student continues schooling. Indirect language proficiency measures might include free production tests (in which a familiar picture or topic is presented and the student is asked to talk or write about it), word association tests, cloze tests, story retelling, or peer interaction.

Very few standardized instruments for assessing proficiency in other languages are designed for the native speaker of that language, to assess academic language use. In general, bilingual assessment specialists in each school district develop indirect assessment measures that ask students to perform certain age-appropriate tasks orally in L_1 and other tasks that include reading and writing in L_1. Testing experts have found that translating tests available in English directly into other languages is not an acceptable practice. Translating alone violates the statistical support for a test (for example, item analysis and other statistical procedures for establishing validity and reliability must be reestablished for the translation), and often a straight translation will be culturally and linguistically irrelevant (Erickson & Omark, 1981; Lowe & Stansfield, 1988). Age-appropriate L_1 tests designed by local bilingual assessment experts can be practical, authentic, interactive, and directly relevant to the purpose for which the test is being used when designed with knowledge of measurement concepts and with clarity of purpose of the test (academic use of the language) (Bachman & Palmer, 1996).

The same points apply to L_2 proficiency testing in English. Many standardized English proficiency measures can be used for placement purposes. But if the test does not measure academic uses of the language, it may not serve the purpose intended. Nevertheless, an important reason to use a standardized proficiency measure upon entry is for purposes of program evaluation. It is crucial to know where students are in L_2 proficiency, as measured by a standardized instrument, when they enter the school program. Then groups of students at the same general L_2 proficiency level are followed across time, to measure progress in L_2 development using the same standardized instrument. Typically, on an instrument that uses a rating scale from 1 to 5, where 1 is novice level and 5 is advanced level, students acquiring ESL can reach the highest ratings on the test, levels 4 and 5, in two to three years. This measure, while important, is not to be confused with developmental (linked to cognitive) progress in language acquisition, which the school tests measure. We will examine this issue further in later sections of this chapter, "Assessment for Exit or Reclassification" and "Evaluation of Student Progress."

In general, in choosing language proficiency tests for placement decisions and program evaluation purposes, schools are advised to avoid older language assessment instruments developed in the 1950s, 1960s, and 1970s, most of which were based on theoretical views that had a scanty research base and are no longer accepted by most second-language acquisition theorists and researchers. Since that time, the field of second-language acquisition has exploded, and models of current language testing reflect the current research knowledge base. In brief, let us examine the changes that have occurred in language testing over the past three decades.

Choosing Language Assessment Measures

Older Discrete-Point Tests

The first attempts at standardized language assessment came with the development of discrete-point testing. A product of audiolingual materials and methods of teaching based on structural linguistics and behavioral psychology, discrete-point testing is best represented by Lado's (1961) *Language Testing*. A discrete-point test focuses on measuring only one small unit of language for each test item. For example, one item might measure one vocabulary word or one phoneme or one morpheme or one rule of syntax, while testing only one language mode (listening, speaking, reading, or writing). The underlying assumption in dividing language into discrete units was that mastery of the parts leads to mastery of the complete language system. More current language theories strongly question this assumption. The main reason discrete-point tests were adopted as an improvement over more informal methods of assessment was that with such item types, language tests could be standardized and tested for validity and reliability to meet psychometric standards of measurement. On most standardized tests used by school districts and states today, the language arts curricular subtest measures mostly discrete points of language. In contrast, the reading subtest involves items that require problem solving across the curriculum.

Integrative Tests

While discrete-point language testing grew in popularity during the 1960s, there were voices of dissent. Carroll (1961) proposed that a more appropriate measure of language would be an integrative test that measures a student's ability to integrate several discrete points of language at the same time. Support for integrative testing came from linguists' theories of communicative competence, including the work of sociolinguists who saw language as not easily divisible into parts but as a complex, interrelated whole. In theoretical discussions of communicative competence, researchers have emphasized that language proficiency testing must combine knowledge about a language (grammatical competence, including knowledge of rules of phonology, morphology, syntax, semantics, and the lexicon) with knowledge of how to use that language in a real-life context (adding to the domains listed above pragmatics, paralinguistics, and discourse knowledge, including appropriateness of use of language varieties, domains, repertoires, roles, statuses, attitudes, settings, topics, channels, and so on) (Cummins & Swain, 1986; Rivera, 1983, 1984).

Pragmatic Tests

Oller (1979, 1992) took the movement towards integrative tests one step further by proposing that language tests should be pragmatic, based on the branch of linguistics called pragmatics, which examines language in social context. A pragmatic test both integrates language modes across several language domains and approximates natural language use. Examples of pragmatic tests include oral production, paraphrase recognition, question answering, oral interview, and composition. Pragmatic multiple choice items also can be developed for standardized tests, if they meet the criteria of integration and naturalness.

Oller's major contribution to the language testing field has been to point out the absurd, unnatural uses of language present in many language tests. Older tests often provide a sequence of items or sentences, each one of which has no relation to the previous one. Oller recommends thinking of a test as storytelling or intriguing drama, with the first sentence leading to the next one. After constructing an assessment task, the teacher should review it for naturalness, for meaningful use of language, for sociolinguistic appropriateness, and for authentic language use. Pragmatic tests also usually involve higher-order processing of information, challenging students to think, problem solve, and be more creative in language use rather than simply testing memorized language rules.

Authentic Language Assessment

As part of the reform movement of the 1990s, the field of language testing has now entered a new and promising phase that combines Oller's perspective on pragmatic language testing with authentic assessment across the curriculum. The movement towards authentic assessment is a reflection of the changing views of teaching and learning that have emerged in recent decades, in which

> all individuals are thought to learn by constructing information about the world and by using active and dynamic mental processes.... Students learn most

> effectively through integrative experiences in programs that reflect the interdependence of listening, speaking, reading, writing, thinking, direct experience, and purposeful student interaction. (O'Malley & Valdez Pierce, 1996, p. 10)

Thus authentic language tasks call for higher-order thinking that naturally integrates language domains and modes across the content areas. Authentic language assessment also connects to the real world, both in and outside of school.

Many of the teaching strategies outlined in Chapter Four are intimately connected to authentic assessment. For example, natural listening, speaking, reading, and writing activities may be used as language samples to be evaluated by the teacher. Oral classroom activities, such as puppet dialogue, baking cookies, conducting a science experiment, or recreating the story of "leaving my home country" can be audiotaped or videotaped and rated using scoring rubrics, with criteria for each performance level, that the teacher has created. Likewise, written products produced by students, such as e-mail letters, stories, poems, dialogue journals, games, recipes, instructions for making things, class newspapers, math puzzles, conclusions from science experiments, or maps of the school community, can be assessed using appropriate rating scales or scoring rubrics designed by the teacher to evaluate student progress with consistency and fairness. Student portfolios provide a system for ongoing collection of student work, to be evaluated across time.

O'Malley and Valdez Pierce (1996) provide a comprehensive synthesis of research and practical teacher strategies for making authentic assessment an integral part of the teaching process for English language learners. They analyze in depth the use of oral interviews, story or text retelling, writing samples, projects and exhibitions, experiments and demonstrations, constructed-response items in which students respond in writing to open-ended questions, teacher observations, and portfolios. Their examples can be applied to language assessment for purposes of placement, monitoring of student progress, and completion or exit from a program.

Ongoing Classroom Assessment

After students have been assessed by bilingual assessment specialists for purposes of school program placement, the teacher is an integral partner in the assessment process. In the first weeks of placement, ongoing classroom assessment provides assurance that the initial assessment was fair and appropriate, since that was only the first of many varied measures to assess a given student's potential and academic progress. It is the responsibility of the classroom teacher to continue assessment across the curriculum, in oral language development, reading, writing, math, science, social studies, and learning strategies. In a bilingual program, the curricular areas taught through L_1 must be assessed in L_1 and the themes or subject areas taught through L_2 must be assessed in L_2. Care must be taken to separate the languages of instruction so that each language gets equal development across the curriculum, as the years progress.

L_1/L_2 Assessment of Reading and Writing

Sources to guide the teacher in authentic classroom assessment are plentiful, and it is beyond the scope of this short chapter to provide a detailed guide to performance and portfolio assessment. The greatest number of sources for the bilingual/ESL teacher are focused on the assessment of reading and writing. Among current sources on L_1/L_2 reading and writing assessment that provide both a theoretical, research-based perspective on authentic assessment combined with very practical, hands-on advice for bilingual/ESL teachers are the following:

Carrasquillo and Hedley (1993); Carson and Leki (1993); Escamilla, Andrade, Basurto, and Ruiz (1996); Farr and Tone (1994); Genishi (1993); Goodman (1990); Goodman, Bird, and Goodman (1992); Hamp-Lyons (1991); Hiebert (1991); Hudelson (1989); O'Malley and Valdez Pierce (1996); Pérez and Torres-Guzmán (1996); Rueda and García (1994); Spangenberg-Urbschat and Pritchard (1994); Valdez Pierce and O'Malley (1992); and Valencia, Hiebert, and Afflerbach (1994). *TESOL Journal,* a quarterly journal of teaching and classroom research in ESL/bilingual education, periodically publishes articles on alternative assessment.

New standards for the assessment of reading and writing that incorporate a multicultural perspective have been published jointly by the International Reading Association (IRA) and the National Council of Teachers of English (NCTE) (1994). These two large professional organizations have jointly concluded that in our postindustrial, information-driven society, we have moved from knowledge transmission to inquiry as a primary goal of schools:

> Within an inquiry framework, assessment is the exploration of how the educational environment and the participants in the educational community support education as a process of learning to become independent thinkers and problem solvers. . . . Inquiry values the ability to recognize problems and to generate multiple and diverse perspectives in trying to solve them. . . . An inquiry perspective . . . would value the question of how information from different sources can be used to solve a particular problem. It would value explorations of how teachers can promote critical thinking for all students. And it would raise the question of why our society privileges the knowledge and cultural heritage of some groups over others within current school settings. (IRA & NCTE, 1994, p. 6)

The IRA/NCTE standards define the nature of language, how language is learned, and the assessment of language, incorporating and celebrating the diversity present in all human uses of language. They then establish key principles for the assessment of reading and writing:

1. The interests of the student are paramount in assessment.
2. The primary purpose of assessment is to improve teaching and learning.
3. Assessment must reflect and allow for critical inquiry into curriculum and instruction.
4. Assessments must recognize and reflect the intellectually and socially complex nature of reading and writing and the important roles of school, home, and society in literacy development.
5. Assessment must be fair and equitable.
6. The consequences of an assessment procedure are the first, and most important, consideration in establishing the validity of the assessment.
7. The teacher is the most important agent of assessment.
8. The assessment process should involve multiple perspectives and sources of data.
9. Assessment must be based in the school community.
10. All members of the educational community—students, parents, teachers, administrators, policy makers, and the public—must have a voice in the development, interpretation, and reporting of assessment.
11. Parents must be involved as active, essential participants in the assessment process. (IRA & NCTE, 1994)

Because reading and writing is such an important means of experiencing the whole school curriculum, every content area teacher is assessing language with every task. And every task has multiple perspectives, from a critical thinking point of view. "Language is very much like a living organism. It cannot be put together from parts like a machine, and it is constantly changing. Like a living organism it exists only in interaction with others, in a social interdependence" (IRA & NCTE, 1994, p. 7). Assessment of L_1 and L_2 must reflect the living languages we humans use.

L_1/L_2 Content Assessment

Much less has been written about assessment approaches for culturally and linguistically diverse students across the curricular areas of mathematics, science, and social studies. Among the researchers guiding bilingual/ESL teachers in science education, Warren and Rosebery (1992) provide clues to meaningful assessment in science investigations. In inner-city bilingual classrooms, student-initiated science inquiry develops as

> Students pose questions, design research, use tools to make sense of the world, collect data, build and argue theories, and document and communicate their findings and interpretations in various ways. Students' inquiries stretch over long periods of time, not just weeks but in many cases months. They take unexpected turns. The context of students' scientific work is social rather than individual. Further, in a sense-making culture, teachers take on a variety of roles; they coach and model scientific practices, and they act as co-investigators. (Warren & Rosebery, 1992, p. 283)

Assessment in this context must focus on what it means to be scientifically literate. As teachers' assessment practices evolved out of the project, they examined "quality of the students' questions, their understanding of data, their critical-mindedness, and their initiative in defining questions or problems to explore" (Warren & Rosebery, 1992, p. 298). Spoken and written scientific discourse as well as student portfolios were used in the assessment process.

Uses of authentic assessment for English language learners across the curricular areas of ESL math, science, and social studies are discussed in some depth in O'Malley and Valdez Pierce (1996) which also provides guidelines for grade-level teachers working with students of mixed levels of proficiency in English. Secada (1992a) discusses curricular changes in mathematics reform and the need for bilingual/ESL teachers to incorporate a problem-solving, critical thinking perspective into both instructional and assessment practices in math.

Learning Strategy Assessment

Appropriate, ongoing assessment also must incorporate reflection on the learning processes that students are incorporating into their repertoires, in the context of interactive, discovery learning. Learning strategies were discussed in Chapters Two, Three, Six, and Seven. Authentic assessment provides a natural means of observing and analyzing learning strategy acquisition, including one of the key principles of authentic assessment—the participation of the student in self-assessment ("Hmmm—how did I get that answer?"). Teacher references that provide detailed ideas for bilingual/ESL teachers in learning strategy assessment are Chamot and O'Malley (1994); Oxford (1990); and a special issue on learning styles and strategies in *TESOL Journal* (Autumn 1996).

Assessment for Exit or Reclassification

State and federal policy guidelines or mandates often require some form of assessment to determine a student's readiness to join the educational mainstream when the student is placed in a separate bilingual or ESL program. Bilingual classes that are designed as enrichment programs, such as developmental bilingual education and two-way dual language or bilingual immersion programs, do not require exit criteria because they are considered mainstream or grade-level programs in which the standard school curriculum is taught through the medium of two languages, for as many grades as possible.

For programs in which exit or reclassification criteria are needed, many practices of the past two decades have inadequately measured students' readiness for grade-level classes. During the 1970s, because of the federal *Lau* guidelines' initial emphasis on oral proficiency assessment (due to the initial federal focus on "non- or limited-English-*speaking* students" in the 1968 Bilingual Education Act), state laws followed a similar pattern of placement and exit from special programs based only on students' *oral* proficiency in English. Early exit led to teachers in the grade-level (mainstream) classroom assuming something was wrong with the child, resulting in assessment for placement in special education. Overrepresentation of language minority students in special education classes, especially in the category of learning disabilities, is a legacy of these practices, and unfortunately this placement is still all too common (Baca & Cervantes, 1989; Cummins, 1984a; González, Brusca-Vega, & Yawkey, 1997; Ortiz, 1992).

When the 1978 reauthorization of the Bilingual Education Act expanded the definition of those eligible for special services to "limited-English-*proficient* students," the focus for entry and exit assessment changed to include all four language modes: listening, speaking, reading, and writing. During the 1980s, most state legislative mandates and guidelines gradually shifted the definition of those eligible for services to include assessment of both oral and written English proficiency. While this was an important step towards more appropriate assessment practices for exit and reclassification, this focus only on English language proficiency measures ignored students' readiness for the standard curriculum in mathematics, science, social studies, and language arts, taught through the medium of English. Over the decade of the 1980s, several states took the additional step of requiring academic achievement measures for reclassification. In a 1990-91 survey of states in the eastern half of the United States, standardized norm-referenced achievement tests in English (especially the reading and mathematics subtests) were mandated for exit from bilingual/ESL programs in Florida, Georgia, Illinois, Iowa, Michigan, Minnesota, New Jersey, and Wisconsin; and they were recommended in Louisiana, Maryland, Mississippi, South Carolina, and Tennessee (O'Malley & Valdez Pierce, 1994).

Use of the norm-referenced test has both a positive and a negative side. Since language minority students must take this test to examine their readiness for grade-level classes, this measure provides a comparison to the typical range of performance of native English speakers across the United States. Ultimately, our goal is for groups of students who were formerly English-language learners to reach the typical performance of native English speakers, as a group, on all tests across the curriculum. So this is an important measure to use, not only for exit decisions, but also to follow students' progress in grade-level classes. To be able to take the norm-referenced test, however, English-language learners must be far enough along in English proficiency development that they can understand the directions for the test, as well as what the tasks are, and they must have sufficient English vocabulary and content knowledge to be able to answer at least the questions of easy-to-intermediate level difficulty for all subject areas of the test. Typically, students initially starting with no proficiency in English take a minimum of two to four years to reach the level where the norm-referenced test is a more appropriate measure, but students in even the highest quality programs should not be expected, after only two to three years of exposure to English, to reach the 50th percentile or normal curve equivalent (NCE) on the norm-referenced test. Groups of students in high-quality bilingual programs typically reach the 50th NCE in L_2 after four to seven years of dual-language schooling. Groups of students who receive high-quality ESL content programs, with no L_1 support, during their first years of schooling in the United States, followed by academic work in grade-level classes entirely in English, can potentially reach the 50th NCE in L_2 after 7 to 10 years of all-English schooling, if they do not drop out of school, but many leave school in frustration (Thomas & Collier, 1997).

While a norm-referenced test can be a very important measure of long-term achievement, a norm-referenced test can be misused. First, this is an important *group* measure, but *individual* student progress should be examined using multiple measures (one of which can be the norm-referenced test), since one test represents only one point in time. Some states mandate cutoff scores (e.g., reaching the 27th percentile or the 40th percentile) on the standardized language proficiency and achievement tests, which represent arbitrary decisions based on group data rather than individual student expectations. Expecting every student to reach the 40th percentile does not take into account the normal distribution for any group, in which if the cutoff is the 40th percentile, 40 percent of the English learners in that group would not yet be expected to score at that level at that particular point in time. A rigid cutoff score also does not reflect the standard error of measurement, which helps to account for individual student variance in performance that is expected from day to day. Several states recommend that along with the norm-referenced test, teacher referral and grades be included in the exit decision. Others use criterion-referenced measures and a few use portfolio and performance assessment. We, along with other researchers and assessment experts, recommend that multiple measures, one of which should be a norm-referenced test, are most appropriate for exit decisions (De George, 1988; Del Vecchio, et al., 1994; LaCelle-Peterson & Rivera, 1994; O'Malley & Valdez Pierce, 1994; Saville-Troike, 1991; Troike, 1982). The true usefulness of the norm-referenced test for exiting decisions is not to provide a cutoff score but to provide information as to where, in the native-English-speaker distribution, the English-language learners belong, with each increasing year of exposure to the English language.

Second, norm-referenced tests should be used as only one of several measures, in order to vary item types used in the assessment process. Multiple measures provide greater variety in the types of assessment tasks. In contrast, norm-referenced tests typically rely on multiple-choice format because they have been developed for efficiency and cost-effectiveness when testing large numbers of students. Different assessment formats give students many different ways of demonstrating their problem-solving capabilities across the curriculum.

Third, if a cutoff score is used and the norm-referenced test is the only measure for exit decisions, students in segregated programs in which they do not receive access to the standard grade-level school curriculum often get further and further behind the typical performance of U.S. students across the country. The longer they remain in separate programs, the greater is the academic achievement gap with each passing school year. Thus, they lose ground to the native English speakers, resulting in national percentile scores that are lower with each passing year. This does not happen in a quality developmental or two-way bilingual program, but it can happen in bilingual and ESL classes that water down the curriculum or do not provide students access to all curricular subjects. When students are schooled initially in segregated, low-quality programs, and a norm-referenced cutoff score is used for exit, keeping them in the low-quality program is the equivalent of an educational dead-end or permanent track and denies students their basic rights of access to the standard school curriculum (Oakes, 1985; Pottinger, 1970; Wheelock, 1992). Assessment practices in bilingual/ESL classes should provide teachers ongoing feedback that they are preparing students for grade-level work across the curriculum, and as students exit from a special program, they must continue to be monitored across time, with the goal of eventual parity with native English speakers on all school tests. We will revisit this issue in the section below on evaluation of student progress.

Research on reclassification and exit from bilingual and ESL classes has shown that another potential problem with exit decisions occurs for the young child in grades K through three. As children who are initially assessed as non-English-proficient or limited-English-proficient move through the curriculum in the early grades, they all seem to be doing extremely well whatever type of special program they are placed in. While tests initially measure these students as being in the bottom 10 percent of all students taking the tests in English, they demonstrate on these tests that they can make greater progress in each school year than the typical native English speaker, and in these early grades they appear to be closing the gap. Many young children in bilingual and ESL classes reach anywhere from the 20th to the 40th normal curve equivalent (NCE) on the tests in English (L_2) in grades K through three, and most programs exit these students by the end of grade three, assuming that they will continue to close the gap and in one or two more years reach the 50th NCE or percentile. But as these same students move into the more demanding cognitive and academic work of grades 4 through 12, in grade-level classes taught all in L_2, they are not able to sustain their gains of the early childhood years. Thus, their percentile or NCE scores decrease relative to the native English speaker with each passing year. Only those who continue to receive strong academic support through L_1 at least from preschool and kindergarten through grade five are able to reach the 50th NCE or percentile and sustain the gains, remaining at the 50th percentile or above throughout the remainder of their schooling (Thomas & Collier, 1997).

Assessment for Bilingual Special Education

Today culturally and linguistically diverse students who need special educational services due to "autism, deaf-blindness, hearing impairments including deafness, mental retardation, multiple disabilities, orthopedic impairments, other health impairments, speech or language impairments, serious emotional disturbance, specific learning disabilities, traumatic brain injury, or visual impairments including blindness" are given the right to a free and appropriate public education, under the 1975 federal Public Law 94–142, The Education of All Handicapped Children Act, which was newly amended and renamed the Individuals with Disabilities Education Act (IDEA) in 1990 (González, Brusca-Vega, & Yawkey, 1997, p. 22). Assessment practices for placement in special education services are complex and beyond the scope of this book, but we will briefly discuss a few of the issues to consider in assessment decisions for students with special needs. The IDEA requires that in the case of a student who is referred for special testing, parental consent and notification must be provided in oral, manual, or written form in the parents' native language; that a multidisciplinary team with the appropriate cultural and linguistic background must be used in the assessment process; that the assessment must be administered in the student's native language; and that no single procedure may be used as a sole criterion in the evaluation process. A student who is assessed as requiring special services must have an Individualized Education Program designed by the multidisciplinary team, including a plan for ongoing student assessment, and placement in the least restrictive environment (González, Brusca-Vega, & Yawkey, 1997, pp. 44–46).

Even given the protection of the federal law and a number of court decisions that have set legal precedents for interpreting the rights of culturally and linguistically diverse students, assessment practices continue to lead to inappropriate placement. Ironically, because of court cases such as *Diana* v. *California State Board of Education* (1970), which brought to educators' attention the overrepresentation of culturally and linguistically diverse children in classes for the mentally retarded, and because of the federal requirements that assessment procedures not be racially, culturally, or linguistically discriminatory, students in need of bilingual special education services are underrepresented in some areas. One reason for this problem is the critical shortage of trained bilingual special education personnel to conduct appropriate assessments and to provide an instructional setting that is culturally and linguistically appropriate for the child (Ortiz & Yates, 1983). But bilingual/bicultural students are now overrepresented in classes for students with learning disabilities, and the label has disastrous consequences for students' self-esteem and academic achievement (Morris, 1997). For example,

> Wilkinson and Ortiz (1986) found that, after three years of special education placement, Hispanic students who were classified as learning disabled had actually lost ground. Their verbal and performance IQ scores were lower than they had been at initial entry into special education and their achievement scores were at essentially the same level as at entry. Neither regular education nor special education programs adequately served the academic needs of these language minority students. (Ortiz, 1992, p. 316)

To address the serious dilemma of the inappropriate placement of many culturally and linguistically diverse students into classes for the learning disabled, Ortiz (1992) suggests that even though prereferral intervention is helpful, this by itself is too simplistic in perspective. Prereferral strategies refer to helping the grade-level teacher modify instructional and classroom management patterns to accommodate diverse students' needs. Broader, deeper perspectives include Cummins' (1996b) model for radically changing power relations in schools so that students and teachers co-create a learning community in which all together value students' and teachers' languages and cultural backgrounds, incorporate global knowledge into the curriculum, create highly interactive learning environments, encourage community participation, promote critical literacy, and conduct assessment that empowers rather than disables students.

In any given student population, 5 to 10 percent may genuinely benefit from and need special education services. For students for whom English is not their first language who are assessed appropriately and found to be in need of special assistance, bilingual special education is a crucial service. Assessment decisions in bilingual special education can be examined in depth in many excellent sources, such as Baca and Cervantes (1989); Baca and De Valenzuela (1994); Carrasquillo and Baecher (1990); Cummins (1984a); García and Ortiz (1988); González, Brusca-Vega, and Yawkey (1997); Grossman (1995); Jones (1988); Ortiz (1992); Ortiz, Wilkinson, Robertson-Courtney, and Bergman (1990).

Bilingualism and Cognition: The Assessment of "Intelligence"

Studies that have examined the relationship between bilingualism and cognition were first conducted under the general rubric of the elusive concept called "intelligence." Intelligence testing has been conducted in the United States during most of the 20th century to examine an individual's ability to "perceive, organize, remember, and utilize symbolic information" (García, 1994, p. 150). Most of the U.S. studies through the early 1960s that compared bilingual and monolingual students on standardized intelligence tests concluded that bilinguals had "disadvantages" or "deficits" as a result of their bilingualism. The so-called negative cognitive effect is now generally referred to as an artifact of poorly designed studies. For example, many studies did not assess students' proficiency in L_1 and L_2, but assumed the students to be proficiently bilingual and administered the test in the bilinguals' second language (English), comparing their performance to that of native English speakers. Another common error in study design was to compare middle-class monolinguals to lower-SES bilinguals. A common bias of the earlier intelligence tests was their focus on mainstream U.S. culture (García, 1994).

Then in the early 1960s, Peal and Lambert (1962) conducted what is now considered a classic and well-designed study that more carefully controlled for student background variables:

> The picture that emerges of the French/English bilingual in Montreal is that of a youngster whose wider experiences in two cultures have given him advantages which a monolingual does not enjoy. Intellectually his experience with two language systems seems to have left him with a mental flexibility, a superiority in concept formation, and a more diversified set of mental abilities. . . In contrast, the monolingual appears to have a more unitary structure of intelligence which he must use for all types of intellectual tasks. (p. 6)

Since this study, hundreds of studies have found that bilinguals with sufficient proficiency in both L_1 and L_2 have cognitive advantages over monolinguals on measures of cognitive flexibility, linguistic and metalinguistic abilities, concept formation, divergent thinking, and creativity (Baker, 1993; Bialystok, 1991; Cummins & Swain, 1986; Díaz, 1983; Hakuta, 1986, 1990; Hamers & Blanc, 1989; Homel, Palij, & Aaronson, 1987).

IQ tests are just one of many types of tests that have been used to examine a student's cognitive abilities and potential. We know that IQ tests are predictors of academic potential because of high positive correlations of IQ scores with academic achievement tests. However, linguistic bias is present when IQ tests are administered by school psychologists in English to students who are not yet proficient in English. While students are often exited from ESL on the basis of their score on a language proficiency measure in L_2 showing that they have made good progress, they have not by any means reached full L_2 proficiency. Cummins (1984a) questions the practice of assuming L_2 proficiency and then blaming the student's low score on an IQ test on deficiencies in the student or his or her background. Since the court case of *Diana* v. *California* (1970), in which the plaintiffs were migrant, Spanish-speaking children who had been tested with an instrument to measure intelligence in English and placed in classes for the educable mentally retarded, the use of nonverbal measures of IQ has become standard practice in California (Figueroa, 1980). Several new assessment instruments have been developed during the past decade to more appropriately assess the culturally and linguistically diverse student when referred for possible placement in special education classes. (See the references in the previous section on bilingual special education for reviews of these instruments.) García (1994), summarizes U.S. school practices of the recent past:

> Intelligence tests and the concept of intelligence that underlies them have done and continue to do an educational disservice to culturally diverse students. The

> limitations of these educational "tools" should be recognized and educators should be extremely cautious when using them to generalize about students. (p. 156)

DEFINING ACADEMIC SUCCESS

How should we define language minority students' success in school? How should bilingual/ESL program success be defined? How can assessment guide us in evaluating our efforts? What assessment tools are available to us, and what are their relative strengths and weaknesses? How do we respond to the standards movement, with its "high-stakes" tests that can deny educational opportunity to English language learners, its accountability requirements for educators; and its citations of schools that are low achieving? We shall begin to address these questions using some of the findings from the Thomas and Collier (1997) "how long" research, which provides a new definition of student success for the field of bilingual/multicultural/ESL education.

Evaluation of Student Progress

Determining the short-term and long-term effectiveness of education programs for language minority students is one of the most important uses of assessment information. And student outcomes are a key criterion of program effectiveness. In this section, we will address the aspects of program evaluation that focus on school accountability when individual student progress is examined across time.

Teachers should be active participants in assessment for program evaluation purposes, as well as beneficiaries of the evaluative results of assessment activities, for several reasons. First, although teachers do not ordinarily focus on the evaluation of school programs, the classroom is the site of much of the testing and student assessment done for this purpose, and so teachers are critical to its success. Second, it is important that teachers provide useful formative input to program administrators regarding what instructional changes to the program are desirable. Third, teachers should be able to investigate the short-term impact of their efforts on students in this year's classroom, as well as the long-term achievement results of their work with students from previous years' classes.

Defining Student Success

Historically, most teachers have tended to define classroom success in terms of the degree to which language minority students make progress in mastering the school curriculum. This has been reinforced over the years by criterion-referenced testing, by teachers' classroom assessments, which monitor student progress from week to week, and by pretest and posttest comparisons of all types. In addition, the use of language proficiency tests to determine the degree of student progress has led to the perception that English learners only need to constantly increase their scores on such tests across time to be thought of as successful. In the aggregate, if most students are making progress by constantly increasing the group's average achievement, measured by a fixed set of items or tasks on several occasions, then the program to which these students belong is frequently deemed "successful."

After a typical period of two to three years, many of the English language learners in bilingual/ESL classes have achieved high scores on these static English language proficiency tests (a fixed set of items or tasks presented several times during the duration of the English learners' period of participation in an

ESL/bilingual program), and they are deemed ready for grade-level classes and exited from the ESL/bilingual program. Since these students are demonstrating high scores on an L_2 language proficiency measure, it is assumed that they will continue to make progress in the instructional mainstream, after leaving the "sheltered" conditions of the ESL/bilingual program. But there are several things wrong with this picture.

First, the definition of success, when defined as "making progress," is incomplete. It is quite possible that an English learner can make five months' progress in each school year, and arrive at the end of the school years with a larger achievement gap between the English learner and the typical native English speaker than existed at the beginning of the English learner's schooling! Why? **Because the typical native English speaker makes 10 months' progress in each 10 month school year,** and does so in all academic subjects, in his or her continuing, developmental acquisition of English and in the underlying cognitive development that goes on throughout the school years. The school tests measure that nonstop growth by constantly changing what they are measuring, with a new test each week, month, and year of school. In other words, an English learner must not only make progress, as measured by a static set of assessment items and tasks, but also make enough progress to keep up with the constantly advancing native English speaker in academic achievement, cognitive development, and linguistic development. In fact, in order to ever catch up to the native English speaker in these areas and to be able to compete with native English speakers in access to jobs and further education after high school, the English learner must actually outperform the average native English speaker across time in L_2 since a wide achievement gap (typically 25–30 NCEs) initially exists between the two groups when the English learner is first assessed on a school test in English (L_2). For example, when schooled only in English, the English learner who begins schooling three years behind the native English speaker in achievement as measured through tests in English, must make 15 years' progress (the normal 12 years plus 3 years of "catch up" achievement) during the 12 years of schooling, whereas the typical native English speaker is making 12 years' progress in 12 years of schooling.

A second problem with the "making progress" definition of school success for English learners is that it does not provide for the English learner's continuing cognitive development while he or she is learning English. If the English learner suffers "cognitive slowdown" while learning English because of school policies that deny students' continuing cognitive development through L_1, and the native English speaker continues normal developmental progress through Piaget's and other cognitive theorists' stages of cognitive growth, then a "cognitive gap" is created whose effects are cumulative. Lessened L_1 cognitive development will manifest itself in lower test scores during the more cognitively demanding years of middle and high school education.

A third problem is that program success defined by "making continuous progress" does not provide for the fact that the native English speaker is constantly acquiring the English language during the school years. At the beginning of schooling, native English speakers have acquired only a small portion of the complete adult system of the English language that they will gradually acquire both subconsciously and consciously throughout their school years. The acquisition of oral and written adult English is a long-term, developmental process that occurs throughout the school years, with native English speakers being instructed in their native language in a supportive sociocultural environment for learning. Thus, English learners also experience a long-term, developmental process for both L_1 and L_2, and they must accommodate to the reality that when instructed in *second* language, the sociocultural environment may not be as supportive as it is for native English speakers.

Thus, in order to arrive at parity with the native English speaker by the end of the school years in mastery of adult English, the academic subjects other than English, and cognitive development, the English learner requires L_1 and L_2 academic assessments that are sensitive to her or his growth academically, cognitively, and linguistically. Why is it so important that both teachers and administrators be sensitive to these needs for appropriate assessments and for their use in evaluating program success by the long-term-parity method? Because the English learner must be assisted so as not to fall behind academically as a result of watered-down instruction while learning English, must be stimulated cognitively to avoid the negative cognitive effects of an abrupt switch in instructional languages from L_1 to L_2, and must be continuously monitored in the long developmental linguistic process of acquiring English at the long-term level of native-speaker proficiency by end-of-high-school standards. The classroom teacher, administrator, and evaluator have responsibilities for collaboratively selecting and developing appropriate assessment strategies, for interpreting their results in instructionally useful ways, and for using assessment to assure equal educational opportunity for English learners and appropriate long-term achievement for native English speakers as well.

A New Definition of Program Success for English Learners

The concept of equal educational opportunity has been held by the courts to mean "similar instructional outcomes by the end of the school years." This means that the typical group scores (and the range of individual scores) for the two groups, English learners and native English speakers, should be equivalent by the end of the school years, assuming that both groups have experienced a full 12 years of schooling. Stated more formally, **using appropriate achievement measures (e.g., performance assessments, criterion-referenced tests, norm-referenced tests), the distributions of scores for English learners and for native English speakers should be indistinguishable after 12 years of school for both groups.** In short, a successful program for English learners, one that assures equal educational opportunity, should produce a situation by the end of the school years in which it is not possible to distinguish between the English learner group and the native English speaker group using an on-grade-level academic achievement test as administered in English. Within each group, some will still score at low, average, and high levels. However, the groups themselves should overlap each other in their range of scores, and the typical scores of the two groups should be similar.

This is a stringent definition of program success, and exceeds the instructional demands of only "making continuous progress." However, it is a definition of program success that leads potentially to instructional reform that will benefit both native English speakers and English learners. Also, this definition proceeds directly from the inherent intent of the U.S. Constitution to create "equal opportunity" for all U.S. citizens and residents, as well as the intent of court decisions that have extended this constitutional guarantee to educational outcomes as well. The use of appropriate assessment practices by teachers and schools is the primary mechanism by which the important goal of equal educational opportunity for all students is assured.

Assessment and Evaluation Implications

The first implication of the long-term-parity definition of program success is that both instruction and assessment, for both English learners and native English speakers, must be long term, not short term, and must not be restricted to the first few years of instruction, while the English learner is receiving assistance in an ESL/bilingual program. This means that assessment and program evaluation must be ongoing, not restricted to short-term, one-to-two-year time periods typical of program evaluations.

Second, assessment for English learners must not be restricted to measurement of the acquisition of English, but should reflect student growth in all academic subjects, in both L_1 and L_2, as well as growth in developmental cognitive stages throughout all the years of schooling. Teachers and administrators should monitor the progress of all students, language minority students and native English speakers, in math, science, social studies, language arts, and other school subjects. Teachers should assess and monitor students' cognitive development during the school years, and should provide instructional programs that facilitate high cognitive levels of problem solving that bear directly on students' success on all assessment measures.

Third, the comparison group for evaluations of programs for English learners should be the native-English-speaking group of the same age and school experience as the English learners. A program's success should be determined by the degree to which it contributes to the closing of the initial 25 to 30 NCE achievement gap, when tested in English, that exists between native English speakers and English learners even after several years of instruction. This achievement gap may not exist if the achievement of the English learners, as tested in their native language, is compared to the achievement of the English speakers, as tested in their native language. This is an important comparison to examine while English learners are in the process of acquiring the English language. L_1 academic and cognitive progress that is on grade level assures us that students are making the normal developmental progress needed that will bring them long-term success. At the end of schooling, the ultimate success of English learners in our schools is determined by the degree to which they acquire the full adult system of English and the full curriculum, as well as make cognitive progress comparable to that of English speakers.

How long is required for a bilingual or ESL program designed for English learners to close the initial achievement gap with native English speakers on standardized tests in English? We can estimate the length of time required by noting that many evaluators consider annual program gains of four to six NCEs, over and above the full year's progress made by typical native English speakers, to be a sign of a moderately strong instructional program. By comparison, gains of seven to nine NCEs are hallmarks of very strong to exemplary programs. Thus, a "good" program for English learners, in which English learners outgain native English speakers by five NCEs (equivalent to about one-fourth of a national standard deviation) per year, would require six years of sustained gains of five NCEs per year to completely close a 30-NCE initial achievement gap. An exemplary instructional program, in which typical English learners gained 7 to 8 NCEs per year, would require four years to close the 30 NCE gap. Thus, we can see from program evaluation considerations alone that outstanding English-learner instructional programs will require at least four to five years to allow English learners to reach educational parity with English speakers. More typically, programs that gain three NCEs per year may require as much as 10 years of instruction to close the initial achievement gap. However, students must not be segregated for 10 years. They need to experience the cognitive challenge of the grade-level classroom. Programs with the highest long-term success are the enrichment models of bilingual education, providing the cognitive challenge on grade level through both L_1 and L_2 (Thomas & Collier, 1997).

Program Evaluation

The types of program evaluations that could be conducted for a program designed to meet the needs of language minority students can be categorized in many ways. Evaluation approaches may be classified by philosophical, methodological, and paradigmatic differences upon which the evaluations are organized. Worthen, Sanders, and Fitzpatrick (1997) provide the following useful categories of evaluation approaches, based on the primary "orientation" of the evaluator and the major questions and organizers that underlie each approach:

- *Objectives-oriented approaches* determine how well goals and objectives have been met.
- *Management-oriented approaches* provide decision makers with useful information.
- *Consumer-oriented approaches* provide consumers with necessary information to enable informed choice among educational "products."
- *Expertise-oriented approaches* apply professional expertise to make evaluative judgments.
- *Adversary-oriented approaches* provide "pro and con" analyses of educational issues.
- *Participant-oriented approaches* provide evaluative data and interpretations emphasizing the needs and involvement of program stakeholders.

Within each of the above major categories are evaluation models that differ in emphasis on one or more of the characteristics within the category. The evaluation models and approaches most frequently used to evaluate programs for bilingual/ESL education are those in the objectives-oriented and management-oriented categories, primarily because those who fund or commission evaluations are most influenced by program objectives and by program management concerns. We would like to encourage the field to incorporate, from the above list, more varied approaches to bilingual/ESL program evaluation. These less-used approaches can offer much additional information, given that the phenomena examined in educational evaluations are quite complex and call for examination from multiple evaluative perspectives.

Whatever the approach employed in typical evaluations of instructional programs for language minority students, the primary goals of program evaluation are to provide information on and interpretation of the context in which the program operates, the resources required to operate the program, the degree to which the program operates as planned, and the types of instructional methods and strategies that the program employs, as well as the instructional outcomes of the program for the students who participate (Del Vecchio et al., 1994; Worthen, Sanders, & Fitzpatrick, 1997). Thus, in a typical evaluation of a program for language minority students, educators must collect and analyze data that address each of these evaluation goals. As educators whose primary duties directly affect the degree to which evaluation goals are addressed, teachers should have an important role to play in all areas of program evaluation. Ideally, teachers would actively participate in information collection, analysis, and interpretation along with administrators and program evaluators.

However, in practice, teachers are usually the primary information gatherers and information providers, but administrators, supervisory staff, and school boards are the primary consumers, analysts, and interpreters of the evaluative information. All too often, this leads to teachers who feel detached and disconnected from the judgments of program worth made by others. Also, teachers may lose the sense of "ownership" of the program and its evaluation that should be an important means of improving both the quality of instruction and the quality of the program evaluation. Thus, many teachers feel burdened by program evaluations rather than informed by them because teachers perceive that most evaluations primarily address the information and accountability needs of administrators and managers rather than the classroom-based needs of teachers. This is especially true in product evaluation, where information on the program's instructional impacts on student achievement, attitudes, attendance, and other outcomes is gathered primarily by teachers but used and interpreted mostly by administrators and others whose tasks include demonstrating "accountability" to a nervous and wary public.

The means of measuring the program's effect on student achievement is a data collection burden felt most keenly by teachers, and it attracts much criticism from them. In order to explore these concerns further, the major means of measuring the program's "bottom line" effects are summarized in the next section.

Major Types of Assessments Used in Program Evaluation

Although a well-designed evaluation examines a variety of program outcomes, typical evaluations of language minority student programs primarily emphasize student achievement outcomes. Among the major types of assessments used in collecting data on student achievement, we will briefly discuss norm-referenced testing, criterion-referenced testing, and performance assessment. These major categories of assessment are similar in that each has specified a measurement domain of skills, learnings, and capabilities that the assessment's designers have deemed appropriate for students to acquire. However, the domain of the typical norm-referenced test includes a cross-section of skills and learnings culled from major textbook series and curriculum guides that are generic, and perhaps national, in focus rather than specific to the curriculum of a particular school system. Criterion-referenced tests may focus on a local school district's curriculum or on a statewide curriculum, depending on whether a particular state determines curricular guidelines centrally or provides for local flexibility. Performance assessments usually reflect a local curriculum, but some states require student demonstration of state-required skills as well. All three types may also use greatly varied formats for item types. For example, it is a common misconception that norm-referenced tests have only multiple-choice formats, whereas norm-referenced, criterion-referenced, and performance assessments may all use multiple-choice formats as well as open-ended tasks or student-constructed answers. All three types can also be "standardized" tests, meaning that the conditions of test administration, the directions furnished to students, the items or tasks presented, and the scoring criteria are the *same* for all test takers.

These major types of student assessments, all used to measure student achievement outcomes in program evaluations, differ in the standard to which each compares student performance, the degree to which curricular areas are covered broadly or deeply, the degree to which each emphasizes the psychometric constructs of measurement validity versus reliability, and the degree to which student learning is measured directly ("authentically") through student demonstration of "real-world" skills and learnings acquired or indirectly through simulated or constructed tasks to which the student responds. They differ in other ways also, including the range of difficulty, variety, format, efficiency of scoring, type of response (constructed versus selected), and cognitive complexity of the items or tasks included as a part of the assessment. Although norm-referenced, criterion-referenced, and performance assessment instruments can be constructed to vary along most of the above dimensions, in practice each tends to appear in typical "packages" of characteristics from the above list. Each is discussed briefly below.

Norm-Referenced Tests

Norm-referenced tests (NRTs) compare a student's performance to the performance of other students in a nationally representative "norm group," made up of students from school systems all over the country. Most NRTs are designed to emphasize a broad but shallow coverage of a generic, pseudonational curriculum, as determined by analysis of the contents of major text series by the test developers. They include only a few test items on any particular curricular objective, but they include a wide variety of such curricular objectives. Because these tests are administered to thousands of students, typically they present a restricted range of item formats, often emphasizing more multiple-choice items that lend themselves to efficient test scoring, but including other item types as well that allow student-constructed answers. They include items of varying difficulty, from easy to very difficult, in order to assess the full range of student performance beyond minimum competencies. Since they are based on a generic curriculum, they may include items not emphasized in a local school district's curriculum, a source of much criticism by teachers who misunderstand these items. This inclusion of "national" items that may not appear in local curricula (or may appear in a later or earlier grade in the local district curriculum) is intentional and is justified by the fact that students do learn outside of school and are not limited to learning from their classroom experiences. Thus, the NRT addresses the question, "In comparison to students nationwide, when mastery of a generic curriculum is assessed, how does a local student (or a group of local students) compare?" Examples of commercial norm-referenced tests used in grades K through 12 are the Iowa Tests of Basic Skills (1993), the TerraNova (1996, including the newest edition of the Comprehensive Tests of Basic Skills), the California Achievement Tests (1992), and the Spanish Assessment of Basic Education (1991). Norm-referenced tests often used for admission to universities are the Test of English as a Foreign Language, the Scholastic Achievement Test, and the Graduate Record Examinations for admission to graduate school.

Typical scores for NRTs include percentiles and NCEs, which rank local students in a distribution of nationally representative students, leading to interpretations such as, "Zaida is at the 30th percentile in reading, indicating that 30 percent of all students scored lower than she did, and 70 percent scored higher." While NRT items are also reported in terms of mastery of groups of items that reflect curricular objectives, these reports are sometimes problematic because of the range of item difficulties in the test and because only a few items are devoted to each specific curricular objective. Thus, while NRTs provide much useful information to administrators and evaluators interested in assessing the performance of *groups of students* across schools and programs, many teachers find that other test types better address their needs for classroom diagnostic and achievement monitoring.

Criterion-Referenced Tests

Criterion-referenced tests (CRTs) compare a student's performance to a performance criterion or standard that has been specified prior to the assessment. Most CRTs are designed to focus on specific subareas or objectives of the curriculum and to provide enough items to allow for more complete coverage of this objective than the typical NRT. Also, CRT items tend to be less difficult than

NRT items, especially in the case of minimum competency tests. Typical CRT items are designed so that 70 to 90 percent of students will identify the correct answer. The classic CRT relies on forced-choice items, especially multiple choice, as do most NRTs. However, some current CRTs include tasks and items designed to provide authentic assessment features, so that teachers can assess student abilities to construct (rather than choose) correct answers.

The item domain of most CRTs includes items and tasks that are based on a local school system's curriculum, rather than a more generic curriculum. Some exceptions include CRTs that reflect the curricular goals and achievement standards that are tested in statewide assessments of student achievement. Student scores are typically reported in terms of number of items correctly identified from a group of items reflecting one curricular objective (e.g., "María got 7 out of 8 items correct in addition"). This score may be based on a predetermined number of correct items deemed necessary for mastery of the objective or mastery of the total item content of the test. Thus, the CRT typically addresses the question, "How does María's performance on items reflecting one curricular objective (or many objectives) compare to a predetermined standard or criterion for mastery?"

Teachers tend to favor criterion-referenced testing because the tests identify easy-to-teach curricular objectives. However, "teaching to the test," or emphasizing student learnings in areas covered by the test, can become an instructional practice that may seriously lower the cognitive level of instruction (Zehler, Hopstock, Fleischman, & Greniuk, 1994).

In addition, CRTs have other serious disadvantages and weaknesses. First, the exact standard or criterion for mastery is frequently arbitrary and difficult to support using psychometric or other theoretical underpinnings. Typically, it represents a compromise or consensus of widely varying opinions as to what constitutes mastery of a curricular area, rather than a substantial, defensible standard. Second, the items that measure mastery of each objective are, in effect, short (and thus unreliable) tests. Thus, it is quite common for students to demonstrate "mastery" of curricular material in one month that becomes "nonmastery" by the following month, if the same material is retested. In other words, the test-retest reliability of 8- to 10-item "tests" is quite low. Third, it is quite possible for students to demonstrate complete curricular mastery of items presented at low levels of difficulty, while their mastery of more cognitively complex material is quite low. This is most frequently seen when students score well on a local school system's CRT but score lower than expected when their performance is compared to that of similar students nationwide on an NRT. The basic problem is that it is politically difficult for school staff to set high levels for mastery when substantial numbers of their students fail to meet those levels and thus "master" the curriculum. The typical minimum competency approach to CRT test design provides little opportunity for advanced students to show what they can really do and provides little opportunity to identify students who have only superficially or temporarily mastered the curricular material.

Performance Assessment

A performance assessment is usually connected to one or more explicit instructional objectives that are not appropriately tested using a traditional test. Of interest is the student's ability to perform a task or to demonstrate the procedures used to complete the task. The teacher assigns the task to the student and then observes and rates the student's performance, based on previously identified evaluative criteria, instructionally significant procedures that the student must exhibit, or the student's ability to partially or completely demonstrate the desired behavior or skill. The key idea here is that the teacher is interested in knowing whether the student can produce a correct answer rather than recognize the correct answer from alternatives provided. Performance assessments can be designed to assess tasks of varying degrees of difficulty and cognitive complexity.

In the past, performance assessment has been used in the arts (e.g., piano performance) and in vocational-training situations (e.g., a master craftsman observing the work of an apprentice) to assess the student's ability to produce a product to high standards of skill. In recent years, this idea has been extended by educators to the measurement of higher-order cognitive skills (e.g., creating a historical time line or conducting a science experiment). Proponents argue that it provides for assessment of actual performance in a real-world situation, rather than the indirect assessment of a constructed performance, measured using a paper-and-pencil exercise. In addition, proponents assert that more complex instructional objectives can be assessed than is possible when using traditional NRTs or CRTs.

Teachers might construct scoring rubrics, rating scales, checklists of desired behaviors, or observation protocols to assess student performance as a part of their everyday information collection during classroom instruction. At some appropriate time, the teacher can summarize the data from these sources into a description of each student's progress toward mastery of specified instructional objectives. The teacher can then apply the information gleaned from these sources directly into instructional plans for each student, potentially increasing the degree of individualized instruction that is targeted toward each student's diagnosed needs.

Just as in the case of NRTs and CRTs, there are some potential drawbacks to this assessment approach. Although performance assessment is capable of more in-depth and real-world evaluation of student outcomes than the alternatives, it has weaknesses in the areas of objective scoring and labor intensiveness required for a high-quality assessment. In addition, teachers are initially attracted to this form of assessment by its obvious congruence with typical teaching tasks and may falsely perceive that it is easy to use and that valid, reliable results are "automatic" even with little preparation. However, unless teachers are thoroughly trained to produce comprehensive scoring guidelines and criteria (typically called rubrics) and are carefully and completely trained to apply these guidelines objectively and consistently, a performance assessment can be much less objective and less reliable than the alternatives (Thorndike, 1997). In other words, in spite of its potential advantages and intuitive appeal to teachers, much training, preparatory work, and time spent in assessment are necessary for this approach to be successful. In the absence of these factors, and in the typically time-pressed classroom, the results may be quite inferior to traditional alternatives and can take a lot more time. However, properly and fully implemented, this form of student assessment can yield extremely valuable instructional information that cannot be readily replicated using traditional alternatives.

Summary of Assessment Types

Each major type of assessment has strengths that provide information useful to teachers and administrators. Each type of assessment also has weaknesses that fail to address some important assessment goals. Given this situation, the tried-and-true advice is to use each type to glean the important information that each can provide. However, it is not necessary to administer all assessment types to all students on a regular basis. A good strategy is to use performance assessment and CRTs for continuous classroom monitoring of student achievement, and to administer NRTs periodically to provide the occasional "reality check" that compares local students' achievement levels to those of students and curricula beyond those of the local school system.

In particular, it is worth noting that NRTs administered in English are especially inappropriate for English learners during the first two to three years of instruction in English (L_2) because most of the characteristics of the national comparison group are distinctly dissimilar to those of the English learners at this point. However, as these students acquire proficiency in English and master the curriculum over time, a comparison of their long-term performance to that of native speakers of English becomes much more desirable. As students enter the high school years, one can make the case that using an NRT to assess their achievement levels relative to a more generic curriculum, in anticipation of post-high school education and jobs, becomes more desirable than assessment relative to a particular local school system's curriculum. NRTs are typically more predictive of future success in postsecondary education than classroom-based assessments. In summary, language minority students need experience with, and exposure to, all of the major forms of assessment for their long-term educational welfare. Teachers should look at the assessment records of language minority students in their totality, including NRTs, CRTs, and performance assessments, and should carefully note the progress of their students on all of these assessments across time, including long-term follow-up on the achievement of their former pupils.

State and Federal Standards

The standards movement, which is closely connected to the 1990s drive towards U.S. school reform, has led to many conflicting messages regarding the assessment of language minority students, as policy decisions at the state and federal level are made. The debates focus on the dilemmas involved in equity issues regarding appropriate assessment practices for underachieving language minority students in general, as well as on the modification of assessment practices for English learners in particular, such as inclusion or exclusion on L_2 measures, and development of L_1 national assessment measures in Spanish (the language of approximately 73 percent of current English learners in the United States [August & McArthur, 1996]). These issues strongly affect teachers and school administrators in bilingual/ESL settings when national or state assessments are mandated for all students, and when each school or school district is held accountable for student performance on the mandated tests.

Minimum Competency Tests

Among the measures that are considered high-stakes tests are the minimum competency tests that have been developed by 46 states during the past two decades (O'Malley & Valdez Pierce, 1994). These assessments generally require that students achieve a passing score on the test in order to receive their high school diploma, demonstrating that "high school graduates possess at least those minimal skills usually deemed necessary for successful survival in the modern world" (Geisinger, 1992, p. 33). For these tests, some states use a norm-referenced test, others a criterion-referenced test, and others have developed performance assessment measures. Some of the dangers of these tests for English learners come with the particular state policies that ignore the inappropriateness of this type of test administered in English for a student who has only been exposed to English for two to three years. When English learners are excluded from participation in extracurricular activities or are ineligible to enroll in academic courses or receive academic credit in high school until they have reached the "passing" score, these types of state policies clearly deny students an equal educational opportunity guaranteed by federal law.

A survey of state policies in the eastern half of the United States, conducted by O'Malley and Valdez Pierce (1994), identified the following modifications that some states made for English learners: (1) precise criteria and procedures were specified for English learners' inclusion or exclusion from the statewide assessments, including measures of a student's proficiency in English; (2) modifications to test administration procedures were made (but in very few states); such as testing by persons familiar to the students, reading the test to students in L_1 or L_2, allowing use of a dictionary, and allowing extra time for test administration; (3) parents were informed of the test and its consequences (but few states required that notification be in the home language, a federal OCR guideline); and (4) alternatives were specified for students not meeting the minimum competency criteria, such as assessment in L_1 (in only two states), or awarding certificates of attendance (which is a questionable practice if it denies students the right to continue schooling). In a subsequent state survey of all 50 states, conducted by Rivera and Vincent (in press), four states—Hawaii, New Jersey, New Mexico, and New York—had modified their policies to permit local school districts to use alternative assessments, such as portfolios of student work.

Mandated Statewide Assessment Across the Curriculum

Other statewide assessment practices mandate that a standardized assessment be conducted in specific grades throughout students' schooling (e.g., grades 4, 6, 8, 11, or every grade, or grades 3, 5, 7, 9, 11). Some states have developed performance assessments across the curriculum, while others use state-developed criterion-referenced tests, and still others use commercial norm-referenced tests with national norms. The same issues discussed above on the state minimum competency measures apply to determining the appropriateness of these assessment measures for English learners. The Texas Assessment of Academic Skills (TAAS) is currently being developed in both Spanish and English. Illinois is now piloting a new assessment system for linguistically and culturally diverse students that includes an English-language proficiency measure that corresponds more to the normal, developmental linguistic and cognitive growth of students of different ages, the Illinois Measure of Annual Growth in English (IMAGE). In general, there is little consistency across states in policies regarding high-stakes assessment of language minority students, and few studies have been conducted regarding the number of English learners being excluded from or included in state tests, or the impact of state policies on these students (Lawton, 1996; O'Malley & Valdez Pierce, 1994).

Federal Standards: Goals 2000

At the federal level, Goals 2000 standards demand that the opportunity to achieve the standards be made available to all students, including language minority students, and that assessment practices be in place to hold schools accountable for meeting the standards (August, Hakuta, & Pompa, 1994). Some of the state-mandated assessment measures are geared to Goals 2000. In addition, the National Assessment of Educational Progress (NAEP) is a congressionally mandated measure that tests a nationally and regionally representative sample of students in grades 4, 8, and 12, across the curriculum, to address Goal 3 of Goals 2000: "American students will leave Grades 4, 8, and 12 having demonstrated competency in challenging subject matter including English, mathematics, science, history, and geography; and every school in America will ensure that all students learn to use their minds well, so they may be prepared for responsible citizenship, further learning, and productive employment in our modern economy" (U.S. Department of Education, 1991, p. 3). The NAEP includes multiple-choice as well as short and extended, constructed-response items. Once again, the same issues apply for English learners being asked to take the NAEP as those for state-mandated assessments. In a conference focused on "inclusion guidelines and accommodations for limited-English-proficient students in the National Assessment of Educational Progress," participants recommended that a Spanish-language version of the NAEP be developed, that only those English learners proficient enough in written English should be included in the assessment, and that English learners' performance be disaggregated from native English speakers' performance (August & McArthur, 1996).

Clearly, assessment practices mandated at the state and federal levels present schools with both problems and possibilities. Shepard (1991b) cautions us to keep in mind the potential dangers of assessment practices that are not done thoughtfully and appropriately:

> High-stakes testing in American schools has always led to decisions about tracking and sorting. It is easy to foresee that challenging tests [for maintenance of high standards for all students], especially those administered in high school, will lead to tracking if admission to test-preparation courses is restricted to those students who are thought to be capable of handling the material . . . Questions of equity remain to be addressed for these new examinations. The issue is not just whether the exams measure fairly but also how they control educational opportunity. (p. 237)

So we must fight for language minority students' rights to appropriate and meaningful assessment practices in both L_1 and L_2, at federal, state, and local levels. Assessment done responsibly, using multiple measures across time, should give us a clear picture of their performance relative to native English speakers of their age. Assessment informs our teaching and administrative practices and encourages high quality, challenging learning environments for all.

General Resources

Throughout this chapter, we have referred to many sources that provide the rich detail of implementation of assessment practices in bilingual/ESL education. Textbooks that provide a comprehensive and current view of assessment from the classroom perspective in general (not focused on bilingual/ESL education) include Airasian (1997), Stiggins (1992), and Wiggins (1993, 1997). Program evaluators are encouraged to read Worthen, Sanders and Fitzpatrick (1997). Teachers implementing authentic assessment with English-language learners will find abundant hands-on, practical strategies in O'Malley & Valdez Pierce (1996). We would also like to encourage all educators to check the World Wide Web site of the National Clearinghouse for Bilingual Education regularly, to keep up to date with the latest information on instructional and assessment practices in bilingual/ESL education. Other sources of up-to-date information are the ERIC Clearinghouses sponsored by the U.S. Department of Education.