CHAPTER 1

ASSESSMENT CONCEPTS AND ISSUES

OBJECTIVES: After reading this chapter, you will be able to

- understand differences between *assessment* and *testing*, along with other basic assessment concepts and terms
- distinguish among five different types of language tests, cite examples of each, and apply them for different purposes and contexts
- appreciate historical antecedents of present-day trends and research in language assessment
- grasp some of the major current issues that assessment researchers are now addressing

Tests have a way of scaring students. How many times in your school days did you feel yourself tense up when your teacher mentioned a test? The anticipation of the upcoming "moment of truth" provoked feelings of anxiety and self-doubt along with a fervent hope that you would come out on the other end with at least a sense of worthiness. The fear of failure is perhaps one of the strongest negative emotions a student can experience, and the most common instrument inflicting such fear is the test. You are not likely to view a test as positive, pleasant, or affirming, and, like most ordinary mortals, you intensely wish for a miraculous exemption from the ordeal.

And yet, tests seem as unavoidable as tomorrow's sunrise in virtually all educational settings around the world. Courses of study in every discipline are marked by these periodic milestones of progress (or sometimes, in the perception of the learner, confirmations of inadequacy) that have become conventional methods of measurement. The gate-keeping function of tests—from classroom achievement tests to large-scale standardized tests—has become an acceptable norm.

Now, just for fun, take the following quiz. All five of the words are found in standard English dictionaries, so you should be able to answer all five items easily, right? Okay, go for it.

Directions: In each of the five items below, select the definition that correctly defines the word. You have two minutes to complete this test!

1. onager
a. a large specialized bit used in the final stages of oil well drilling
b. in cultural anthropology, an adolescent approaching puberty
c. an Asian wild ass with a broad dorsal stripe
d. a phrase or word that quantifies a noun

2. shroff
a. (Yiddish) a prayer shawl worn by Hassidic Jews
b. a fragment of an ancient manuscript
c. (Archaic) past tense form of the verb *to shrive*
d. a banker or money changer who evaluates coin

3. hadal
a. relating to the deepest parts of the ocean below 20,000 feet
b. one of seven stations in the Islamic *hajj*, or pilgrimage, to Mecca
c. a traditional Romanian folk dance performed at spring festivals
d. pertaining to Hades

4. chary
a. discreetly cautious and vigilant about dangers and risks
b. pertaining to damp, humid weather before a rainstorm
c. optimistic, positive, looking on the bright side
d. expensive beyond one's means

5. yabby
a. overly talkative, obnoxiously loquacious
b. any of various burrowing Australian crayfishes
c. a small horse-drawn carriage used in Victorian England for transporting one or two persons
d. in clockwork mechanisms, a small latch for calibrating the correct time

Now, how did that make you feel? Probably just the same as many learners feel when they take multiple-choice (or shall we say multiple-guess?), timed, "tricky" tests. To add to the torment, if this were a commercially administered standardized test, you would probably get a score that, in your mind, demonstrates that you did *worse* than hundreds of people! If you're curious about how you did on the quiz, check your answers by turning to page 22 at the end of this chapter.

Of course, this little quiz on obscure, infrequently used English words is not an appropriate example of classroom-based achievement testing, nor is it intended to be. It was designed to be overly difficult, to offer you no opportunity to use contextual clues, and to give you little chance of deciphering the words from your knowledge of English. It's simply an illustration of how tests make us feel much of the time.

Here's the bottom line: Tests need *not* be degrading or threatening to your students. Can they build a person's confidence and become learning experiences? Can they become an integral part of a student's ongoing classroom development? Can they bring out the best in students? The answer is yes. That's mostly what this book is about: helping you as a teacher create more authentic, intrinsically motivating assessment procedures that are appropriate for their context and designed to offer constructive feedback to your students.

To reach this goal, it's important to understand some basic concepts: What do we mean by *assessment*? What is the difference between assessment and a test? And how do various categories of assessments and tests fit into the teaching-learning process?

ASSESSMENT AND TESTING

Assessment is a popular and sometimes misunderstood term in current educational practice. You might be tempted to think of *assessing* and *testing* as synonymous terms, but they are not. Let's differentiate the two concepts.

Assessment is "appraising or estimating the level or magnitude of some attribute of a person" (Mousavi, 2009, p. 36). In educational practice, assessment is an ongoing process that encompasses a wide range of methodological techniques. Whenever a student responds to a question, offers a comment, or tries out a new word or structure, the teacher subconsciously makes an appraisal of the student's performance. Written work—from a jotted-down phrase to a formal essay—is performance that ultimately is "judged" by self, teacher, and possibly other students. Reading and listening activities usually require some sort of productive performance that the teacher observes and then implicitly appraises, however peripheral that appraisal may be. A good teacher never ceases to assess students, whether those assessments are incidental or intended.

Tests, on the other hand, are a subset of assessment, a genre of assessment techniques. They are prepared administrative procedures that occur at identifiable times in a curriculum when learners muster all their faculties to offer peak performance, knowing that their responses are being measured and evaluated.

In scientific terms, a test is a method of measuring a person's ability, knowledge, or performance in a given domain. Let's look at the components of this definition. A test is first a *method.* It's an instrument—a set of techniques, procedures, or items—that requires performance on the part of the test-taker. To qualify as a test, the method must be explicit and structured: multiple-choice questions with prescribed correct answers, a writing prompt with a scoring rubric, an oral interview based on a question script, or a checklist of expected responses to be filled in by the administrator.

Second, a test must *measure,* which may be defined as a process of quantifying a test-taker's performance according to explicit procedures or rules (Bachman, 1990, pp. 18–19). Some tests measure general ability, whereas others focus on very specific competencies or objectives. A multiskill proficiency test determines a general ability level; a quiz on recognizing correct use of definite articles measures specific knowledge. The way the results or measurements are communicated may vary. Some tests, such as a classroom-based, short-answer essay test, may earn the test-taker a letter grade accompanied by the instructor's marginal comments. Others, particularly large-scale standardized tests, provide a total numerical score, a percentile rank, and perhaps some subscores. If an instrument does not specify a form of reporting measurement—a means for offering the test-taker some kind of result—then that technique cannot appropriately be defined as a test.

Next, a test measures an *individual's* ability, knowledge, or performance. Testers need to understand who the test-takers are. What are their previous experiences and backgrounds? Is the test appropriately matched to their abilities? How should test-takers interpret their scores?

A test measures **performance**, but the results imply the test-taker's ability or, to use a term common in the field of linguistics, **competence**. Most language tests measure one's ability to perform language, that is, to speak, write, read, or listen to a subset of language. On the other hand, it is not uncommon to find tests designed to tap into a test-taker's knowledge about language: defining a vocabulary item, reciting a grammatical rule, or identifying a rhetorical feature in written discourse. Performance-based tests sample the test-taker's actual use of language, but from those samples the test administrator infers general competence. A test of reading comprehension, for example, may consist of several short reading passages each followed by a limited number of comprehension questions—a small sample of a second language learner's total reading behavior. But from the results of that test, the examiner may infer a certain level of general reading ability.

Finally, a test measures a given *domain*. For example, in the case of a proficiency test, even though the actual performance on the test involves only a sampling of skills, the domain is overall proficiency in a language—general competence in all skills of a language. Other tests may have more specific criteria. A test of pronunciation might well be a test of only a limited set of phonemic minimal pairs. A vocabulary test may focus on only the set of words covered in a particular lesson or unit. One of the biggest obstacles to overcome in constructing adequate tests is to measure the desired criterion and not include other factors inadvertently, an issue that is addressed in Chapters 2 and 3.

A well-constructed test is an instrument that provides an accurate measure of the test-taker's ability within a particular domain. The definition sounds fairly simple, but, in fact, constructing a good test is a complex task involving both science and art.

Measurement and Evaluation

A couple other potentially confusing terms often appear in discussions of assessment and testing: *measurement* and *evaluation*. Because the terms lie somewhere in between assessment and testing, they are at times mistakenly used as synonyms of one or the other concept. Let's take a brief look at these two processes.

Measurement is the process of quantifying the observed performance of classroom learners. Bachman (1990) cautioned us to distinguish between quantitative and qualitative descriptions of student performance. Simply put, the former involves assigning numbers (including rankings and letter grades) to observed performance whereas the latter consists of written descriptions, oral feedback, and other nonquantifiable reports.

There are clear advantages to quantification. Numbers allow us to provide exact descriptions of student performance and to compare one student to another more easily. They also can spur us to be explicit in our specifications for scoring student responses, thus leading to greater objectivity. On the other hand, quantifying student performance can work against the teacher or tester, perhaps masking nuances of performance or giving an air of certainty when scoring **rubrics** may actually be quite vague. Verbal or qualitative descriptions may offer an opportunity for a teacher to individualize feedback to a student, such as in marginal comments on a student's written work or oral feedback on pronunciation.

Yet another potentially confusing term that needs explanation is **evaluation.** Is evaluation the same as testing? Evaluation does not necessarily entail testing; rather, evaluation is involved when the results of a test (or other assessment procedure) are used for decision making (Bachman, 1990, pp. 22-23). Evaluation involves the interpretation of information. Simply recording numbers or making check marks on a chart does not constitute evaluation. You evaluate when you "value" the results in such a way that the worth of the performance is conveyed to the test-taker, usually with some reference to the consequences—good or bad—of the performance.

Test scores are an example of measurement, and conveying the "meaning" of those scores is evaluation. If a student achieves a score of 75 percent (measurement) on a final classroom examination, he or she may be told that the score resulted in a failure (evaluation) to pass the course. Evaluation can take place without measurement, as in, for example, a teacher's appraisal of a student's correct oral response with words like "excellent insight, Fernando!"

Assessment and Learning

Returning to our contrast between tests and assessment, we find that tests are a subset of assessment, but they are certainly not the only form of assessment that a teacher can make. Although tests can be useful devices, they are only one among many procedures and tasks that teachers can ultimately use to assess (and measure) students. But now, you might be thinking, if you make assessments every time you teach something in the classroom, does all teaching involve assessment? Are teachers constantly assessing students with no interaction that is assessment-free?

The answer depends on your perspective. For optimal learning to take place, students in the classroom must have the freedom to experiment, to try out their own hypotheses about language without feeling that their overall competence is being judged in terms of those trials and errors. In the same way that tournament tennis players must, before a tournament, have the freedom to practice their skills with no implications for their final placement on that day of days, so also must learners have ample opportunities to "play" with language in a classroom without being formally graded. Teaching sets up the practice games of language learning: the opportunities for learners to listen, think, take risks, set goals, and process feedback from the "coach" and then recycle through the skills that they are trying to master.

At the same time, during these practice activities, teachers (and tennis coaches) are indeed observing students' performance, possibly taking measurements, offering qualitative feedback, and making strategic suggestions. How did the performance compare to previous performance? Which aspects of the performance were better than others? Is the learner performing up to an expected potential? What can the learner do to improve performance the next time? How does the performance compare to that of others in the same learning community? In the ideal classroom, all these observations feed into the way the teacher provides instruction to each student (See Clapham, 2000, for a discussion of the relationship among testing, assessment, and teaching.).

Our discussion of all these overlapping concepts is represented in the following diagram (see Figure 1.1) showing the interrelationships among testing, measurement, assessment, teaching, and evaluation.

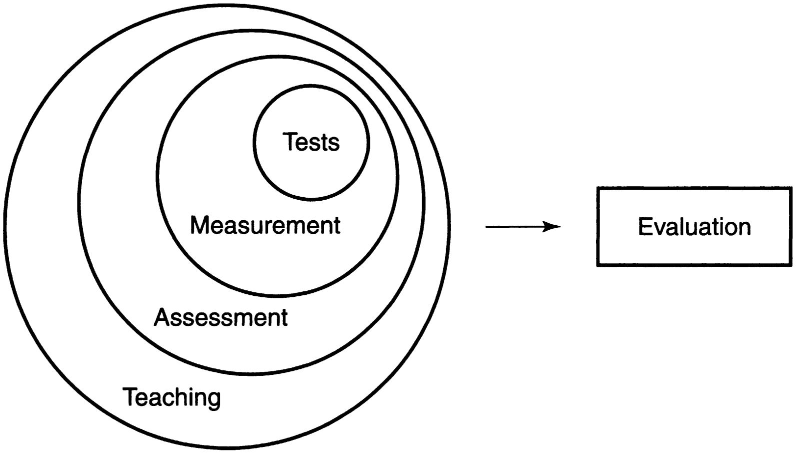

Figure 1.1. Tests, measurement, assessment, teaching, and evaluation

Informal and Formal Assessment

One way to begin untangling the lexical conundrum created by distinguishing among tests, assessment, teaching, and other related concepts is to distinguish between informal and formal assessment. **Informal assessment** can take a number of forms, starting with incidental, unplanned comments and responses, along with coaching and other impromptu feedback to the student. Examples include saying "Nice job!"; "Good work!"; "Did you say *can* or *can't?*"; "I think you meant to say you *broke* the glass, not you *break* the glass"; or putting a smiley face on some homework.

Informal assessment does not stop there. A good deal of a teacher's informal assessment is embedded in classroom tasks designed to elicit performance without recording results and making fixed conclusions about a student's competence. Informal assessment is virtually always nonjudgmental, in that you as a teacher are not making ultimate decisions about the student's performance; you're simply trying to be a good coach. Examples at this end of the continuum are marginal comments on papers, responding to a draft of an essay, offering advice about how to better pronounce a word, suggesting a strategy for compensating for a reading difficulty, or showing a student how to modify his or her notetaking to better remember the content of a lecture.

On the other hand, **formal assessments** are exercises or procedures specifically designed to tap into a storehouse of skills and knowledge. They are systematic, planned sampling techniques constructed to give teacher and student an appraisal of student achievement. To extend the tennis analogy, formal assessments are the tournament games that occur periodically in the course of a regimen of practice.

Is formal assessment the same as a test? We can say that all tests are formal assessments, but not all formal assessment is testing. For example, you might use a student's journal or portfolio of materials as a formal assessment of the attainment of certain course objectives, but it is problematic to call those two procedures "tests." A systematic set of observations of a student's frequency of oral participation in class is certainly a formal assessment, but it too is hardly what anyone would call a test. Tests are usually relatively time-constrained (usually spanning a class period or at most several hours) and draw on a limited sample of behavior.

Formative and Summative Assessment

Another useful distinction to bear in mind is the function of an assessment: How is the procedure to be used? Two functions are commonly identified in the literature: formative and summative assessment. Most of our classroom assessment is **formative assessment**: evaluating students in the process of "forming" their competencies and skills with the goal of helping them to continue that growth process. The key to such formation is the delivery (by the teacher) and internalization (by the student) of appropriate feedback on performance, with an eye toward the future continuation (or formation) of learning.

For all practical purposes, virtually all kinds of informal assessment are (or should be) formative. They have as their primary focus the ongoing development of the learner's language. So when you give a student a comment or a suggestion, or call attention to an error, that feedback is offered to improve the learner's language ability.

Summative assessment aims to measure, or summarize, what a student has grasped and typically occurs at the end of a course or unit of instruction. A summation of what a student has learned implies looking back and taking stock of how well that student has accomplished objectives, but it does not necessarily point the way to future progress. Final exams in a course and general proficiency exams are examples of summative assessment. Summative assessment often, but not always, involves evaluation (decision making).

Ross (2005) cited research to show that the appeal of formative assessment is growing and that conventional summative testing of language-learning outcomes is gradually integrating formative modes of assessing language learning as an ongoing process. Also, Black and William's (1998) analysis of 540 research studies found that formative assessment was superior to summative assessment in providing crucial information to classroom teachers.

One of the problems with prevailing attitudes toward testing is the view that all tests (quizzes, periodic review tests, midterm exams, etc.) are summative. At various points in your past educational experiences, no doubt you've considered such tests summative. You may have thought, "Whew! I'm glad that's over. Now I don't have to remember that stuff anymore!" A challenge to you as a teacher is to change that attitude among your students: Can you instill a more formative quality to what your students might otherwise view as a summative test? Can you offer your students an opportunity to convert tests into "learning experiences"? We will take up that challenge in subsequent chapters in this book.

Norm-Referenced and Criterion-Referenced Tests

Another dichotomy that's important to clarify and that aids in sorting out common terminology in assessment is the distinction between norm-referenced and criterion-referenced testing. In **norm-referenced tests**, each test-taker's score is interpreted in relation to a mean (average score), median (middle score), standard deviation (extent of variance in scores), and/or percentile rank. The purpose of such tests is to place test-takers along a mathematical continuum in rank order. Scores are usually reported back to the test-taker in the form of a numerical score (e.g., 230 out of 300) and a percentile rank (such as 84 percent, which means that the test-taker's score was higher than 84 percent of the total number of test-takers but lower than 16 percent in that administration). Typical of norm-referenced tests are standardized tests such as the Scholastic Aptitude Test (SAT®), the Graduate Record Exam (GRE®), and the Test of English as a Foreign Language (TOEFL® Test), all intended to be administered to large audiences, with results efficiently disseminated to test-takers. Such tests must have fixed, predetermined responses in a format that can be scored mechanically at minimum expense. Cost and efficiency are primary concerns in these tests.

Criterion-referenced tests, on the other hand, are designed to give test-takers feedback, usually in the form of grades, on specific course or lesson objectives. Classroom tests involving students in only one course, and connected to a curriculum, are typical of criterion-referenced testing. A good deal of time and effort on the part of the teacher (test administrator) is sometimes required to deliver useful, appropriate feedback to students, or what Oller (1979, p. 52) called "instructional value." In a criterion-referenced test, the distribution of students' scores across a continuum may be of little concern as long as the instrument assesses appropriate objectives (Brown & Hudson, 2000; Lynch & Davidson, 1994). In *Language*

Assessment, with an audience of classroom language teachers and teachers in training, and with its emphasis on classroom-based assessment (as opposed to standardized, large-scale testing), criterion-referenced testing is of more prominent interest than norm-referenced testing.

TYPES AND PURPOSES OF ASSESSMENT

Assessment instruments, whether formal tests or informal assessments, serve multiple purposes. Commercially designed and administered tests may be used for measuring proficiency, placing students into one of several levels of a course, or diagnosing students' strengths and weaknesses according to specific linguistic categories, among other purposes. Classroom-based teacher-made tests might be used to diagnose difficulty or measure achievement in a given unit of a course. Specifying the purpose of an assessment instrument and stating its objectives is an essential first step in choosing, designing, revising, or adapting the procedure you will finally use.

Tests tend to fall into a finite number of types, classified according to their purpose. Let's take a look at these test types, so that you'll be familiar with them before proceeding with the practical task of creating your own assessments. We'll begin with the most common type for classroom-based assessment.

Achievement Tests

The most frequent purpose for which a classroom teacher will use a test is to measure learners' ability within a classroom lesson, unit, or even total curriculum. Commonly called **achievement tests**, they are (or should be) limited to particular material addressed in a curriculum within a particular time frame and are offered after a course has focused on the objectives in question. Achievement tests can also serve the diagnostic role of indicating what a student needs to continue to work on in the future, but the primary role of an achievement test is to determine whether course objectives have been met—and appropriate knowledge and skills acquired—by the end of a given period of instruction.

Achievement tests are often summative because they are administered at the end of a lesson, unit, or term of study. They also play an important formative role, because an effective achievement test will offer feedback about the quality of a learner's performance in subsets of the unit or course. The specifications for an achievement test should be determined by

- the objectives of the lesson, unit, or course being assessed
- the relative importance (or weight) assigned to each objective
- the tasks employed in classroom lessons during the unit of time
- the time frame for the test and turnaround time
- its potential for formative feedback

Achievement tests range from 5- or 10-minute quizzes to three-hour final examinations, with an almost infinite variety of item types and formats. More will be said in Chapter 3 about choosing assessment methods for achievement tests.

Diagnostic Tests

The purpose of a **diagnostic test** is to diagnose aspects of a language that a student needs to develop or that a course should include. A test in pronunciation, for example, might diagnose the phonological features of English that are difficult for learners and should therefore become part of a curriculum. Usually, such tests offer a checklist of features for the administrator (often the teacher) to use in pinpointing difficulties. A writing diagnostic would elicit a writing sample from students that would allow the teacher to identify those rhetorical and linguistic features on which the course needs to focus special attention.

It's tempting to blur the line of distinction between a diagnostic test and a general achievement test. Achievement tests analyze the extent to which students have acquired language features that have already been taught; diagnostic tests should elicit information on what students need to work on in the future. Therefore, a diagnostic test will typically offer more detailed, subcategorized information on the learner. In a curriculum that has a grammatical form-focused phase, for example, a diagnostic test might offer information about a learner's acquisition of verb tenses, modal auxiliaries, definite articles, relative clauses, and the like. Likewise, a course in oral production might start off with a read-aloud passage (Prator, 1972) or an elicitation of a free speech sample (Celce-Murcia, Brinton, & Goodwin, 1996, p. 346), either of which could give a teacher an advance sense of a learner's ability to produce stress and rhythm patterns, intonation, and segmental phonemes.

Placement Tests

Some achievement tests and proficiency tests (see the next page) can act as **placement tests**, the purpose of which is to place a student into a particular level or section of a language curriculum or school. A placement test usually, but not always, includes a sampling of the material to be covered in the various courses in a curriculum; a student's performance on the test should indicate the point at which the student will find material neither too easy nor too difficult but appropriately challenging.

Some would argue that an effective placement test should be diagnostic as well. If an institution is going to the effort and expense to administer a test for placing students into one of several possible levels of a curriculum, a beneficial side effect of such a test would be a breakdown of strengths and weaknesses that students showed. A tally of correct and incorrect responses, categorized by modules in a curriculum, can provide teachers with useful information on what may or may not need to be emphasized in the weeks to come. Thus a placement test takes on a formative role.

Placement tests come in many varieties—assessing comprehension and production, responding through written and oral performance, open-ended and limited responses, selection (e.g., multiple-choice) and gap-filling formats—depending on the nature of a program and its needs. Some programs simply use existing standardized proficiency tests because of their obvious advantage in practicality: cost, speed in scoring, and efficient reporting of results. Other programs prefer specific course-based assessments that double as diagnostic instruments. Although the ultimate objective of a placement test is to correctly place a student into a course or level, a very useful secondary benefit is diagnostic information on a student's performance, which in turn gives teachers a head start on assessing their students' abilities.

Proficiency Tests

If your aim is to test global competence in a language, then you are, in conventional terminology, testing proficiency. A **proficiency test** is not limited to any one course, curriculum, or single skill in the language; rather, it tests overall ability. Proficiency tests have traditionally consisted of standardized multiple-choice items on grammar, vocabulary, reading comprehension, and aural comprehension. Many commercially produced proficiency tests—the Test of English as a Foreign Language (TOEFL® Test), for example—include a sample of writing as well as oral production performance.

Proficiency tests are almost always summative and norm-referenced. They provide results in the form of a single score (and usually two or three subscores, one for each section of a test), which, to many, is sufficient for the **gate-keeping** role they play of accepting or denying someone passage into the next level of education. Also, because they measure performance against a norm, with equated scores and percentile ranks taking on paramount importance, they are usually not equipped to provide diagnostic feedback.

A key issue in testing proficiency is how the **constructs** of language ability are specified. (See Chapter 2 for further discussion of assessment constructs.) The tasks that test-takers are required to perform must be legitimate samples of English language use in a defined context. Creating these tasks and validating them with research is a time-consuming and costly process. Language teachers should not attempt to create an overall proficiency test on their own. A far more practical method is to choose one of a number of commercially available proficiency tests.

Aptitude Tests

This last type of test no longer enjoys the widespread use it once had. An **aptitude test** is designed to measure capacity or general ability to learn a foreign language *a priori* (before taking a course) and ultimate predicted success in that undertaking. Language aptitude tests were ostensibly designed to apply to the classroom learning of any language.

Two standardized aptitude tests were once used in the United States: the Modern Language Aptitude Test (MLAT; Carroll & Sapon, 1958) and the Pimsleur Language Aptitude Battery (PLAB; Pimsleur, 1966). Both are English language tests and require students to perform language-related tasks such as number learning, distinguishing speech sounds, detecting grammatical functions, and memorizing paired associates. (See Appendix for more information on the MLAT.)

The MLAT and PLAB show some significant correlations with ultimate performance of students in language courses (Carroll, 1981). Those correlations, however, presuppose a foreign language course in which success is measured by similar processes of mimicry, memorization, and puzzle-solving. There is no research to show unequivocally that those kinds of tasks predict communicative success in a language, especially untutored acquisition of the language.

Because of this limitation, standardized aptitude tests are seldom used today, with the exception, perhaps, of identifying foreign language-learning disability (Stansfield & Reed, 2004). Instead, attempts to measure language aptitude more often provide learners with information about their preferred styles and their potential strengths and weaknesses, with follow-up strategies for capitalizing on the strengths and overcoming the weaknesses (Robinson, 2005; Skehan, 2002). Any test that claims to predict success in learning a language is undoubtedly flawed, because we now know that with appropriate self-knowledge, active strategic involvement in learning, and/or strategies-based instruction, virtually everyone can eventually succeed. To pigeon-hole learners *a priori,* before they have even attempted to learn a language, is to presuppose failure or success without substantial cause. (A further discussion of language aptitude can be found in H. D. Brown's [2007a] *Principles of Language Learning and Teaching [PLLT],* Chapter 4.)[1]

ISSUES IN LANGUAGE ASSESSMENT: THEN AND NOW

Before moving on to the practicalities of creating classroom tests and assessments, you will better appreciate the intricacies of the process by taking a brief historical look at language testing over the past half-century and taking note of some current issues in the field.

Historically, language-testing trends and practices have followed the shifting sands of teaching methodology (for a description of these trends, see H. D. Brown [2007b], *Teaching by Principles* [*TBP*], Chapter 2). For example, in the 1940s and 1950s, an era of behaviorism and special attention to contrastive analysis, language tests focused on specific linguistic elements such as the phonological, grammatical, and lexical contrasts between two languages. In the 1970s and 1980s, communicative theories of language brought with them a more integrative view of testing in which specialists claimed that "the whole of the communicative event

[1] Frequent references are made in this book to companion volumes by H. Douglas Brown. *Principles of Language Learning and Teaching* (*PLLT*; fifth edition, 2007) is a basic teacher reference book on essential foundations of second language acquisition on which pedagogical practices are based. *Teaching by Principles* (*TBP*; third edition, 2007) spells out that pedagogy in practical terms for the language teacher.

was considerably greater than the sum of its linguistic elements" (Clark, 1983, p. 432). Today, test designers are still challenged in their quest for more authentic, valid instruments that simulate real-world interaction (Leung & Lewkowicz, 2006).

Behavioral Influences on Language Testing

Through the middle of the twentieth century, language teaching and testing were both strongly influenced by behavioral psychology and structural linguistics. Both traditions emphasized sentence-level grammatical paradigms, definitions of vocabulary items, and translation from first language to second language and placed only minor focus on real-world authentic communication. Typically, tests consisted of grammar and vocabulary items in multiple-choice format along with a variety of translation exercises ranging from words to sentences to short paragraphs.

Such **discrete-point** formats still prevail today, especially in large-scale standardized "entrance examinations" used to admit students to institutions of higher education around the world. (See Barnwell, 1996, and Spolsky, 1978, 1995, for a summary.) Essentially, assessments were designed on the assumption that language can be broken down into its component parts and that those parts can be tested successfully. These components are the skills of listening, speaking, reading, and writing and the various units of language (discrete points) of phonology/graphology, morphology, lexicon, syntax, and discourse. It was claimed that an overall language proficiency test, then, should sample all four skills and as many linguistic discrete points as possible.

Discrete-point testing provided fertile ground for what Spolsky (1978, 1995) called the **psychometric-structuralist** approach to language assessment, in which test designers seized the tools of the day to focus on issues of validity, reliability, and objectivity. Standardized tests of language blossomed in this scientific climate, and the language teaching/testing world saw such tests as the Michigan Test of English Language Proficiency (1961) and the Test of English as a Foreign Language (1963) become extraordinarily popular. The science of measurement and the art of teaching appeared to have made a revolutionary alliance.

Integrative Approaches

In the midst of this fervor, language pedagogy was rapidly moving in more communicative directions, and testing specialists were forced into a debate that would soon respond to the changes. The discrete-point approach presupposed a decontextualization that was proving to be inauthentic. So, as the profession emerged into an era of emphasizing communication, authenticity, and context, new approaches were sought. John Oller (1979) argued that language competence was a unified set of interacting abilities that could not be tested separately. His claim was that communicative competence is so global and requires such integration that it cannot be captured in additive tests of grammar, reading, vocabulary, and other discrete points of language. Others (among them Cziko, 1982, and Savignon, 1982) soon followed in their support for what became known as **integrative testing**.

What does an integrative test look like? Two types of tests were, at the time, claimed to be examples of integrative tests: cloze tests and dictations. A **cloze** test is a reading passage (perhaps 150 to 300 words) in which roughly every sixth or seventh word has been deleted; the test-taker is required to supply words that fit into those blanks. (See Chapter 8 for a full discussion of cloze testing.)

Oller (1979) claimed that cloze test results were good measures of overall proficiency. According to theoretical constructs underlying this claim, the ability to supply appropriate words in blanks requires competence in a language, which includes knowledge of vocabulary, grammatical structure, discourse structure; reading skills and strategies; and an internalized "expectancy" grammar (enabling one to predict an item that will come next in a sequence). It was argued that successful completion of cloze items taps into all of those abilities, which were said to be the essence of global language proficiency.

Dictation, in which learners listen to a short passage and write what they hear, is a familiar language-teaching technique that evolved into a testing technique. (See Chapter 6 for a discussion of dictation as an assessment device.) Supporters argued that dictation was an integrative test because it taps into grammatical and discourse competencies required for other modes of performance in a language. Success on a dictation test requires careful listening, reproduction in writing of what is heard, efficient short-term memory, and, to an extent, some expectancy rules to aid the short-term memory. Further, dictation test results tend to correlate strongly with other tests of proficiency. For large-scale testing, the usually classroom-centered dictation technique can be practical and reliable through the design of multiple-choice items.

Proponents of integrative test methods soon centered their arguments on what became known as the **unitary trait hypothesis**, which suggested an "indivisible" view of language proficiency: that vocabulary, grammar, phonology, the "four skills," and other discrete points of language could not be disentangled from each other in language performance. The unitary trait hypothesis argued that there is a general factor of language proficiency such that all the discrete points do not add up to that whole. However, in a series of debates and research evidence (Farhady, 1982; Oller, 1983), the unitary trait hypothesis was abandoned.

Communicative Language Testing

By the mid-1980s, especially in the wake of Canale and Swain's (1980) seminal work on communicative competence, the language-testing field had begun to focus on designing **communicative** language-testing tasks. Bachman and Palmer (1996) included among "fundamental" principles of language testing the need for a correspondence between language test performance and language use: "In order for a particular language test to be useful for its intended purposes, test performance must correspond in demonstrable ways to language use in non-test situations" (p. 9). The problem that language assessment experts faced was that tasks tended to be artificial, contrived, and unlikely to mirror language use in real life. As Weir (1990) noted, "Integrative tests such as cloze only tell us about a candidate's linguistic competence. They do not tell us anything directly about a student's performance ability" (p. 6).

Thus a quest for authenticity was launched, as test designers centered on communicative performance. Following Canale and Swain's (1980) model, Bachman (1990) proposed a model of language competence consisting of organizational and pragmatic competence, respectively subdivided into grammatical and textual components and into illocutionary and sociolinguistic components (see Figure 1.2).

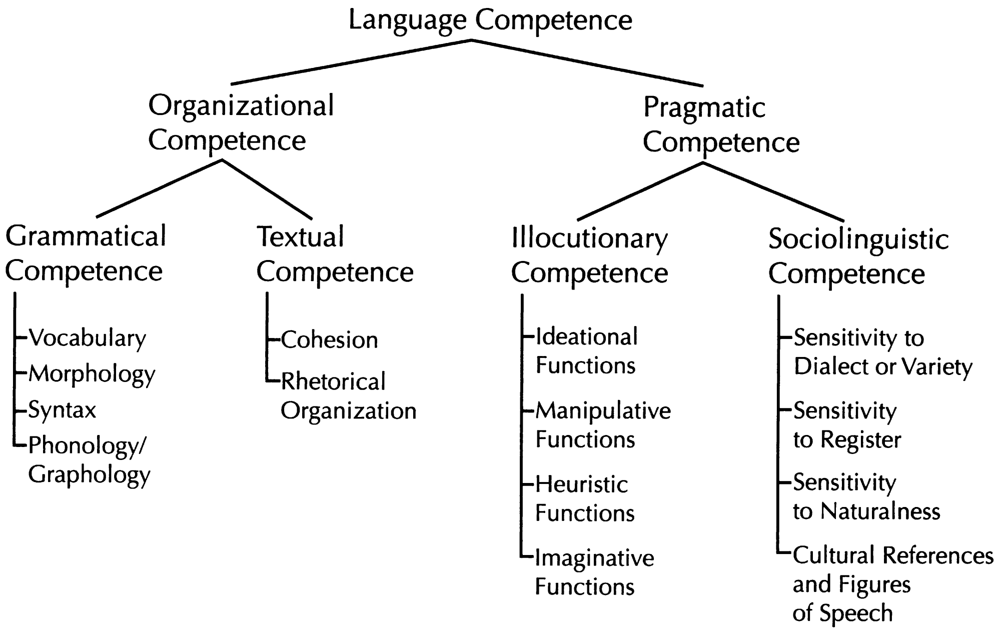

Figure 1.2. Components of language competence (Bachman, 1990, p. 87)

Bachman and Palmer (1996, pp. 70–75) also emphasized the importance of **strategic competence** (the ability to employ communicative strategies to compensate for breakdowns as well as enhance the rhetorical effect of utterances) in the process of communication. All elements of the model, especially pragmatic and strategic abilities, needed to be included in the constructs of language testing and in the actual performance required of test-takers.

Communicative testing presented challenges to test designers, as we will see in subsequent chapters of this book. Test designers began to identify the kinds of real-world tasks that language-learners were called on to perform. It was clear that the contexts for those tasks were extraordinarily widely varied and that the sampling of tasks for any assessment procedure needed to be validated by what language users actually do with language. Weir (1990) reminded his readers that "to measure language proficiency . . . account must now be taken of: where, when, how, with whom, and why language is to be used, and on what topics, and with what effect" (p. 11). The assessment field also became more concerned with the authenticity of tasks and the genuineness of texts. (See Skehan, 1988, 1989, and Fulcher, 2000, for surveys of communicative testing research.)

Performance-Based Assessment

In language courses and programs around the world, test designers are now tackling this new and more student-centered agenda (Alderson & Banerjee, 2001, 2002; Bachman, 2002; Leung & Lewkowicz, 2006; Weir, 2005). Instead of offering paper-and-pencil multiple-choice tests of a plethora of separate items, **performance-based assessment** of language typically involves oral production, written production, open-ended responses, integrated performance (across skill areas), group performance, and other interactive tasks. To be sure, such assessment is time-consuming and therefore expensive, but those extra efforts result in more direct and more accurate testing because students are assessed as they perform actual or simulated real-world tasks. In technical terms, higher content validity (see Chapter 2 for an explanation) is achieved because learners are measured in the process of performing the targeted linguistic acts.

In an English language teaching context, performance-based assessment means you may have a difficult time distinguishing between formal and informal assessment. If you rely a little less on formally structured tests and a little more on evaluation while students are performing various tasks, you will be taking some steps toward meeting the goals of performance-based assessment. (See Chapter 10 for a further discussion of performance-based assessment.)

A characteristic of many (but not all) performance-based language assessments is the presence of interactive tasks, hence the alternative term, **task-based assessment**, for such approaches. In such cases, the assessments involve learners in actually performing the behavior that we want to measure. In interactive tasks, test-takers are measured in the act of speaking, requesting, responding, or in combining listening and speaking, and in integrating reading and writing. Paper-and-pencil tests certainly do not elicit such communicative performance.

A prime example of an interactive language assessment procedure is an oral interview. The test-taker is required to listen accurately to someone else and respond appropriately. If care is taken in the test design process, language elicited and volunteered by the student can be personalized and meaningful, and tasks can approach the authenticity of real-life language use (see Chapter 7).

CURRENT "HOT TOPICS" IN CLASSROOM-BASED ASSESSMENT

Designing communicative, performance-based assessment rubrics continues to challenge assessment experts and classroom teachers alike. In addition, three new issues in the field are shaping our current understanding of effective assessment.

These are: (1) the effect of new theories of intelligence on the testing industry in general, (2) the advent of what has come to be called "alternative" assessment, and (3) the increasing use of computer technology in assessments of various kinds. We briefly explore these issues here.

Multiple Intelligences

Intelligence was once viewed strictly as the ability to perform (a) linguistic and (b) logical-mathematical problem solving. This IQ (intelligence quotient) concept of intelligence has permeated the Western world and its way of testing for almost a century. Because "smartness" in general is measured by timed, discrete-point tests consisting of a hierarchy of separate items, why shouldn't every field of study be so measured? For many years, we have lived in a world of standardized, norm-referenced tests that are timed in a multiple-choice format and consist of a multiplicity of logic-constrained items, many of which are inauthentic.

However, in the last two decades of the twentieth century, research on intelligence began to turn the psychometric world upside down. Howard Gardner (1983, 1999), for example, extended the traditional view of intelligence to eight different components.[2] He accepted the traditional conceptualizations of linguistic intelligence and logical-mathematical intelligence on which standardized IQ tests are based but included other "frames of mind" in his theory of **multiple intelligences**: spatial, musical, kinesthetic, naturalist, interpersonal, and intrapersonal. Robert Sternberg (1988, 1997) also charted new territory in intelligence research in recognizing creative thinking and manipulative strategies as part of intelligence. Likewise, Daniel Goleman's (1995) concept of EQ (emotional quotient) spurred us to underscore the importance of the emotions in our cognitive processing.

These conceptualizations of intelligence were not universally accepted by the academic community (see White, 1998, for example). After all, how does one objectively measure such hypothetical constructs as interpersonal intelligence, creativity, and self-esteem? Nevertheless, their intuitive appeal infused educators with a sense of both freedom and responsibility in their teaching and testing agendas, as evidenced by educational reforms at the time (Armstrong, 1994).

In the language assessment field in particular, the recognition of multiple intelligences has had an indirect effect. On the one hand, communicative classroom activities in textbooks and programs have paid increasing attention to diversity of learning abilities and styles. Christison (2005), for example, offered more than 150 activities for language learners, each emphasizing specific intelligences. On the other hand, in classroom assessment, new views on intelligence have helped to free language instruction programs from relying exclusively on timed, discrete-point, analytical tests in measuring language. Classroom language teachers have been collectively prodded to cautiously combat the potential tyranny of "objectivity" and its

[2] For a summary of Gardner's theory of intelligence, see H. D. Brown (2007a, pp. 107–110).

accompanying impersonal approach. Teachers and administrators have also been urged to measure whole language skills, learning processes, and the ability to negotiate meaning. Our challenge continues to be designing assessments that tap into interpersonal, creative, communicative, and interactive skills and in doing so place some trust in our subjectivity and intuition.

Traditional and Alternative Assessment

Implied in some of the earlier description of performance-based classroom assessment is a trend to supplement traditional test designs with alternatives that are more authentic in their elicitation of meaningful communication. Table 1.1 highlights differences between the two approaches (adapted from Armstrong, 1994, and Bailey, 1998, p. 207).

Two caveats need to be stated here. First, the concepts in Table 1.1 represent some overgeneralizations and should therefore be interpreted with caution. It is difficult, in fact, to draw a clear line of distinction between what Armstrong (1994) and Bailey (1998) called traditional and **alternative assessment**. Many forms of assessment fall in between the two, and some combine the best of both.

Second, it is obvious that the table shows a bias toward alternative assessment, and one should not be misled into thinking that everything on the left-hand side is tainted whereas the list on the right-hand side offers salvation to the field of language assessment. As Brown and Hudson (1998) aptly pointed out, the assessment traditions available to us should be valued and utilized for the functions they provide. At the same time, we might all be stimulated to look at the right-hand list and ask ourselves if, among those concepts, there are alternatives to assessment that we can use constructively in our classrooms.

It should be noted here that considerably more time and higher institutional budgets are required to administer and score assessments that presuppose more subjective evaluation, more individualization, and more interaction in the process

Table 1.1. Traditional and alternative assessment

Traditional Assessment	Alternative Assessment
Standardized exams	Continuous long-term assessment
Timed, multiple-choice format	Untimed, open-ended responses
Decontextualized test items	Contextualized communicative tasks
Scores suffice for feedback	Individualized feedback
Norm-referenced scores	Criterion-referenced scores
Focus on discrete answers	Open-ended, creative answers
Summative	Formative
Oriented to product	Oriented to process
Noninteractive performance	Interactive performance
Fosters extrinsic motivation	Fosters intrinsic motivation

of offering feedback. The payoff for the latter, however, comes with more useful feedback to students, the potential for intrinsic motivation, and ultimately a more complete description of a student's ability. (See Chapter 6 for a complete treatment of alternatives in assessment.) More educators and advocates for educational reform are arguing for a de-emphasis of large-scale standardized tests in favor of contextualized, communicative, performance-based assessment that will better facilitate learning in our schools. (In Chapters 4 and 5, issues surrounding standardized testing are addressed at length.)

Computer-Based Testing

Recent years have seen a burgeoning of computer technology and applications of that technology to language learning and teaching. Virtually every language learner worldwide is, to a lesser or greater extent, a user of computers, the Internet, iPods, cell phones, the Web, and other common cybertechnology. It's no surprise, then, that an overwhelming number of language courses utilize some form of **computer-assisted language learning (CALL)** to achieve their goals, as recent publications show (Chapelle, 2005; Chapelle & Jamieson, 2008; de Szendeffy, 2005; H. D. Brown, 2007b).

The assessment of language learning is no exception to the mushrooming growth of computer technology in educational contexts (see Chapelle & Douglas, 2006; Douglas & Hegelheimer, 2008; Jamieson, 2005, for overviews of computer-based second language testing). Some computer-based tests are small-scale, "homegrown" tests available on a plethora of Web sites. Others are standardized, large-scale tests in which tens of thousands of test-takers may be involved. Students receive prompts (or probes, as they are sometimes referred to) in the form of spoken or written stimuli from preprogrammed algorithm and are required to type (or, in some cases, speak) their responses. Most computer-based test items have fixed, closed-ended responses; however, tests such as the Test of English as a Foreign Language (TOEFL® Test) now offer a written essay section and an oral production section, both of which are scored by humans (as opposed to automatic, electronic, or machine scoring).

Recent developments in computer-based assessment include contributions of **corpus linguistics** in providing more authenticity, the design of more complex tasks in computer-delivered tests, the utilization of speech and writing recognition software to score oral and written production (Jamieson, 2005), and some intriguing questions about "whether and how the delivery medium [of computer-based language testing] changes the nature of the construct being measured" (Douglas & Hegelheimer, 2008, p. 116).

A specific type of computer-based test, a **computer-adaptive test (CAT)**, has been available for many years but has recently gained momentum. In a computer-adaptive test, each test-taker receives a set of questions that meet the test specifications and are generally appropriate for his or her performance level. The CAT starts with questions of moderate difficulty. As test-takers answer each question, the computer scores the question and uses that information, as well as the responses to previous questions, to determine which question will be presented next. As long as examinees respond correctly, the computer typically selects questions of greater or equal difficulty. Incorrect answers, however, typically bring questions of lesser or equal difficulty. The computer is programmed to fulfill the test design as it continuously adjusts to find questions of appropriate difficulty for test-takers at all performance levels. In CATs, the test-taker sees only one question at a time, and the computer scores each question before selecting the next one. As a result, test-takers cannot skip questions, and once they have entered and confirmed their answers, they cannot return to questions or to any earlier part of the test.

Computer-based testing, with or without CAT technology, offers these advantages:

- a variety of easily administered classroom-based tests
- self-directed testing on various aspects of a language (vocabulary, grammar, discourse, one or all of the four skills, etc.)
- practice for upcoming high-stakes standardized tests
- some individualization, in the case of CATs
- large-scale standardized tests that can be administered easily to thousands of test-takers at many different stations, then scored electronically for rapid reporting of results
- improved (but imperfect) technology for automated essay evaluation and speech recognition (Douglas & Hegelheimer, 2008)

Of course, some disadvantages are present in our current predilection for computer-based testing. Among them:

- Lack of security and the possibility of cheating are inherent in unsupervised computerized tests.
- Occasional "homegrown" quizzes that appear on unofficial Web sites may be mistaken for validated assessments.
- The multiple-choice format preferred for most computer-based tests contains the usual potential for flawed item design (see Chapter 3).
- Open-ended responses are less likely to appear due to (a) the expense and potential unreliability of human scoring or (b) the complexity of recognition software for automated scoring.
- The human interactive element (especially in oral production) is absent.
- Validation issues stemming from test-takers approaching tasks as test tasks rather than as real-world language use (Douglas & Hegelheimer, 2008).

Some argue that computer-based testing, pushed to its ultimate level, might mitigate recent efforts to return testing to its artful form of (a) being tailored by teachers for their classrooms, (b) being designed to be performance-based, and (c) allowing a teacher–student dialogue to form the basis of assessment. This need not be the case. While "computer-assisted language tests [CALTs]