CHAPTER 3

DESIGNING CLASSROOM LANGUAGE TESTS

The previous chapters introduced a number of building blocks for designing language tests. You now have a sense of where tests belong in the larger domain of assessment. You have sorted through differences between formal and informal tests, formative and summative tests, and norm- and criterion-referenced tests. You have traced some of the historical lines of thought in the field of language assessment. You have a sense of major current trends in language assessment, especially the present focus on communicative and process-oriented testing that seeks to transform tests from anguishing ordeals into challenging and intrinsically motivating learning experiences. By now, certain foundational principles have entered your vocabulary: practicality, reliability, validity, authenticity, and washback. And you should now possess a few tools with which you can evaluate the effectiveness of a classroom test.

In this chapter, you will draw on those foundations and tools to begin the process of designing tests or revising existing tests. To start that process, you need to ask some critical questions:

1. *What is the purpose of the test?* Why am I creating this test or why was it created by someone else? For an evaluation of overall proficiency? To place students into a course? To measure achievement within a course? Once you have established the major purpose of a test, you can determine its objectives.
2. *What are the objectives of the test?* What specifically am I trying to find out? Establishing appropriate objectives involves a number of issues, ranging from relatively simple ones about forms and functions covered in a course unit to much more complex ones about constructs to be operationalized in the test. Included here are decisions about what language abilities are to be assessed.
3. *How will the test specifications reflect both the purpose and the objectives?* To evaluate or design a test, you must make sure that the objectives are incorporated into a structure that appropriately weights the various competencies being assessed. (These first three questions all center, in one way or another, on the principle of validity.)
4. *How will the test tasks be selected and the separate items arranged?* The tasks that the test-takers must perform need to be practical in the ways defined in

the previous chapter. They should also achieve content validity by presenting tasks that mirror those of the course (or segment thereof) being assessed. Further, they should be able to be evaluated reliably by the teacher or scorer. The tasks themselves should strive for authenticity, and the progression of tasks ought to be biased for best performance.

5. *What kind of scoring, grading, and/or feedback is expected?* Tests vary in the form and function of feedback, depending on their purpose. For every test, the way results are reported is an important consideration. Under some circumstances a letter grade or a holistic score may be appropriate; other circumstances may require that a teacher offer substantive washback to the learner.

These five questions should form the basis of your approach to designing tests for your classroom.

TEST TYPES

The first task you will face in designing a test for your students is to determine the purpose for the test. Defining your purpose will help you choose the right kind of test, and it will also help you to focus on the specific objectives of the test. We will look first at two test types that you will probably not have many opportunities to create as a classroom teacher—language aptitude tests and language proficiency tests—and three types that you will almost certainly need to create—placement tests, diagnostic tests, and achievement tests.

Language Aptitude Tests

One type of test—although admittedly not a very common one—predicts a person's success prior to exposure to the second language. A **language aptitude test** is designed to measure capacity or general ability to learn a foreign language and ultimate success in that undertaking. Language aptitude tests are ostensibly designed to apply to the classroom learning of any language.

Two standardized aptitude tests have been 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 a number of language-related tasks. The MLAT, for example, consists of five different tasks.

Tasks in the Modern Language Aptitude Test

1. Number learning: Examinees must learn a set of numbers through aural input and then discriminate different combinations of those numbers.
2. Phonetic script: Examinees must learn a set of correspondences between speech sounds and phonetic symbols.

3. Spelling clues: Examinees must read words that are spelled somewhat phonetically, and then select from a list the one word whose meaning is closest to the "disguised" word.
4. Words in sentences: Examinees are given a key word in a sentence and are then asked to select a word in a second sentence that performs the same grammatical function as the key word.
5. Paired associates: Examinees must quickly learn a set of vocabulary words from another language and memorize their English meanings.

More information on the MLAT may be obtained from the following website: **http://www.2lti.com/mlat.htm#2**.

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. 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. 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 succeed eventually. 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 *PLLT*, Chapter 4.)

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. Sometimes a sample of writing is added, and more recent tests also include oral production performance. As noted in the previous chapter, such tests often have content validity weaknesses, but several decades of construct validation research have brought us much closer to constructing successful communicative proficiency tests.

Proficiency tests are almost always summative and norm-referenced. They provide results in the form of a single score (or at best two or three subscores, one for

each section of a test), which is a sufficient result for the **gate-keeping** role they play of accepting or denying someone passage into the next stage of a journey. And 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 typical example of a standardized proficiency test is the Test of English as a Foreign Language (TOEFL®) produced by the Educational Testing Service. The TOEFL is used by more than a thousand institutions of higher education in the United States as an indicator of a prospective student's ability to undertake academic work in an English-speaking milieu. The TOEFL consists of sections on listening comprehension, structure (or grammatical accuracy), reading comprehension, and written expression. The new computer-scored TOEFL announced for 2005 will also include an oral production component. With the exception of its writing section, the TOEFL (as well as many other large-scale proficiency tests) is machine-scorable for rapid turnaround and cost effectiveness (that is, for reasons of practicality). Research is in progress (Bernstein et al., 2000) to determine, through the technology of speech recognition, if oral production performance can be adequately machine-scored. (Chapter 4 provides a comprehensive look at the TOEFL and other standardized tests.)

A key issue in testing proficiency is how the *constructs* of language ability are specified. 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 would be wise not 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.

Placement Tests

Certain proficiency tests can act in the role of **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.

The English as a Second Language Placement Test (ESLPT) at San Francisco State University has three parts. In Part I, students read a short article and then write a summary essay. In Part II, students write a composition in response to an article. Part III is multiple-choice: students read an essay and identify grammar errors in it. The maximum time allowed for the test is three hours. Justification for this three-part structure rests largely on the test's content validation. Most of the ESL courses at San Francisco State involve a combination of reading and writing, with a heavy emphasis on writing. The first part of the test acts as both a test of reading comprehension and a test of writing (a summary). The second part requires students to state opinions and to back them up, a task that forms a major component of the

writing courses. Finally, proofreading drafts of essays is a useful academic skill, and the exercise in error detection simulates the proofreading process.

Teachers and administrators in the ESL program at SFSU are satisfied with this test's capacity to discriminate appropriately, and they feel that it is a more authentic test than its multiple-choice, discrete-point, grammar-vocabulary predecessor. The practicality of the ESLPT is relatively low: human evaluators are required for the first two parts, a process more costly in both time and money than running the multiple-choice Part III responses through a pre-programmed scanner. Reliability problems are also present but are mitigated by conscientious training of all evaluators of the test. What is lost in practicality and reliability is gained in the diagnostic information that the ESLPT provides. Statistical analysis of errors in the multiple-choice section furnishes data on each student's grammatical and rhetorical areas of difficulty, and the essay responses are available to teachers later as a preview of their students' writing.

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. Others prefer the performance data available in more open-ended written and/or oral production. The ultimate objective of a placement test is, of course, to correctly place a student into a course or level. Secondary benefits to consider include face validity, diagnostic information on students' performance, and authenticity.

In a recent one-month special summer program in English conversation and writing at San Francisco State University, 30 students were to be placed into one of two sections. The ultimate objective of the placement test (consisting of a five-minute oral interview and an essay-writing task) was to find a performance-based means to divide the students evenly into two sections. This objective might have been achieved easily by administering a simple grid-scorable multiple-choice grammar-vocabulary test. But the interview and writing sample added some important face validity, gave a more personal touch in a small program, and provided some diagnostic information on a group of learners about whom we knew very little prior to their arrival on campus.

Diagnostic Tests

A **diagnostic test** is designed to diagnose specified aspects of a language. 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 needed to focus special attention.

Diagnostic and placement tests, as we have already implied, may sometimes be indistinguishable from each other. The San Francisco State ESLPT serves dual

purposes. Any placement test that offers information beyond simply designating a course level may also serve diagnostic purposes.

There is also a fine line of difference 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 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.

A typical diagnostic test of oral production was created by Clifford Prator (1972) to accompany a manual of English pronunciation. Test-takers are directed to read a 150-word passage while they are tape-recorded. The test administrator then refers to an inventory of phonological items for analyzing a learner's production. After multiple listenings, the administrator produces a checklist of errors in five separate categories, each of which has several subcategories. The main categories include

1. stress and rhythm,
2. intonation,
3. vowels,
4. consonants, and
5. other factors.

An example of subcategories is shown in this list for the first category (stress and rhythm):

a. stress on the wrong syllable (in multi-syllabic words)
b. incorrect sentence stress
c. incorrect division of sentences into thought groups
d. failure to make smooth transitions between words or syllables

(Prator, 1972)

Each subcategory is appropriately referenced to a chapter and section of Prator's manual. This information can help teachers make decisions about aspects of English phonology on which to focus. This same information can help a student become aware of errors and encourage the adoption of appropriate compensatory strategies.

Achievement Tests

An **achievement test** is related directly to classroom lessons, units, or even a total curriculum. Achievement tests 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 period of instruction.

Achievement tests are often summative because they are administered at the end of a unit or term of study. They also play an important formative role. An effective achievement test will offer washback about the quality of a learner's performance in subsets of the unit or course. This washback contributes to the formative nature of such tests.

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,
- practicality issues, such as the time frame for the test and turnaround time, and
- the extent to which the test structure lends itself to formative washback.

Achievement tests range from five- or ten-minute quizzes to three-hour final examinations, with an almost infinite variety of item types and formats. Here is the outline for a midterm examination offered at the high-intermediate level of an intensive English program in the United States. The course focus is on academic reading and writing; the structure of the course and its objectives may be implied from the sections of the test.

Midterm examination outline, high-intermediate

Section A. Vocabulary
Part 1 (5 items): match words and definitions
Part 2 (5 items): use the word in a sentence

Section B. Grammar
(10 sentences): error detection (underline or circle the error)

Section C. Reading comprehension
(2 one-paragraph passages): four short-answer items for each

Section D. Writing
respond to a two-paragraph article on Native American culture

SOME PRACTICAL STEPS TO TEST CONSTRUCTION

The descriptions of types of tests in the preceding section are intended to help you understand how to answer the first question posed in this chapter: What is the