CHAPTER 1

TYPES AND USES OF LANGUAGE TESTS

INTRODUCTION

Before getting into the nuts and bolts of doing language testing, I need to first lay some groundwork by discussing the differences between the two basic families of tests found in language testing. Then, I will define and discuss the four primary functions that these tests serve in language programs. Next, I will explain how administrators and teachers can best match the four basic types of language tests to the purposes and decision-making needs of their own language programs and courses. I will then explain why it is impossible to create a single test that can fulfill the functions of all four basic types of language tests. Finally, I will give a brief introduction to the Microsoft *Excel*™ spreadsheet program. As in all the chapters of this book, I will end with a series of review questions, that will help to summarize the chapter, and a set of application exercises.

TWO FAMILIES OF LANGUAGE TESTS

The first and most basic distinction in language testing involves two families of tests that perform two very different functions: one family helps administrators and teachers make program level decisions, such as proficiency and placement decisions, and the other family helps teachers make classroom-level decisions, such as diagnostic and achievement decisions. In the technical jargon of testing, these two families are called norm-referenced tests and criterion-referenced tests. The concepts underlying norm-referenced testing have been fully developed in educational measurement circles for most of the twentieth century, and many language teachers have been exposed to this category of testing in one way or another. However, the idea of criterion-referenced testing did not surface in educational measurement circles until 1963 when Glaser first mentioned the idea. The distinction between norm-referenced and criterion-referenced tests has only gradually entered the language testing literature, starting in the sixties (see Cartier 1968), skipping the seventies, reappearing in the early eighties (Cziko 1982, 1983; Brown 1984a; Hudson & Lynch 1984), becoming more prominent in the late eighties (Delamere 1985; Henning 1987; Bachman 1987, 1989; Brown 1988a, 1989a; Hudson 1989a, 1989b), gaining more prominence in the nineties (Bachman 1990; Cook 1990; Davidson & Lynch 1993; Griffee 1995; Brown 1990a, 1990b, 1992, 1993, 1995a, 1995b, 1996a; Lynch & Davidson 1994, 1997), and continuing into the new millennium (Brown & Hudson 2002; Davidson & Lynch 2002).

In recent years, the distinction between norm-referenced and criterion-referenced testing has continued to be important in educational and psychological measurement. I hope this will continue because an understanding of the fundamental differences and similarities between these two types of tests can help language program administrators and language teachers make much better decisions about their students.

Norm-Referenced Tests

In brief, a **norm-referenced test** (NRT) is designed to measure global language abilities (i.e., overall English language proficiency, academic listening ability, reading comprehension, and so on). Each student's score on such a test is interpreted relative to the scores of all other students who took the test. Such comparisons are usually done with reference to the concept of the **normal distribution** (familiarly known as the "bell curve"; for more on this concept, see Chapters 5 and 6). The purpose of an NRT is to spread students out along a continuum of scores so that those with low abilities in a general area such as reading comprehension are at one end of the normal distribution, while those with high abilities are at the other end (with the bulk of the students falling between the extremes). In addition, while students may know the general format of the questions on an NRT (for example, multiple-choice, true-false, dictation, or essay), they will typically not know before the test what specific content or skills will be covered by those questions.

Criterion-Referenced Tests

In contrast, a **criterion-referenced test** (CRT) is usually produced to measure well-defined and fairly specific instructional objectives. Often these objectives are specific to a particular course, program, school district, or state. An example of a very strict instructional objective would be the following: By the end of the course the students will be able to underline the sentence containing the main idea of an academic paragraph of 200–250 words at the eleventh grade readability level with 60 percent accuracy. However, objectives come in many forms. Other objectives might be defined in terms of tasks we would expect the students to be able to perform by the end of the term, or experiences we would expect them to go through. For example: by the end of the term the students will watch at least five English language movies with no subtitles. (For more example objectives, see Chapter 3 of Brown 1995a).

The interpretation of scores on a CRT is considered absolute in the sense that each student's score is meaningful without reference to the other students' scores. In other words, a student's score on a particular objective indicates the percent of the knowledge or skill in that objective that the student has learned. Moreover, the distribution of scores on a CRT need not necessarily be normal. If all the students know 100 percent of the material on all the objectives, then all the students should receive the same score with no variation at all. The purpose of a CRT is to measure the amount of learning that a student has accomplished on each objective. In most cases, the students should know in advance what types of questions, tasks, and content to expect for each objective because the question content should be implied (if not explicitly stated) in the objectives of the course.

A more detailed step-by-step comparison of norm-referenced and criterion-referenced tests will help to clarify the distinction. The six characteristics listed in the first column of Table 1.1 indicate that norm-referenced and criterion-referenced tests contrast in: the ways that scores are interpreted, the kinds of things that they are used to measure, the purposes for testing, the ways that scores are distributed, the structures of the test, and the students' knowledge of test question content.

Table 1.1 Norm-referenced and criterion-referenced test differences

Characteristic	Norm-Referenced	Criterion-Referenced
Type of Interpretation	Relative (A student's performance is compared to those of all other students in percentile terms.)	Absolute (A student's performance is compared only to the amount, or percentage, of material learned.)
Type of Measurement	To measure general language abilities or proficiencies	To measure specific objectives-based language points
Purpose of Testing	Spread students out along a continuum of general abilities or proficiencies	Assess the amount of material known or learned by each student
Distribution of Scores	Normal distribution of scores around the mean	Varies; often non-normal. Students who know the material should score 100%.
Test Structure	A few relatively long subtests with a variety of item contents	A series of short, well-defined subtests with similar item contents
Knowledge of Questions	Students have little or no idea of what content to expect in test items.	Students know exactly what content to expect in test items.

Type of interpretation

In terms of the type of interpretation, one essential difference between these two categories of tests is that each student's performance on a CRT is compared to a particular criterion in absolute terms. Some confusion has developed over the years about what the *criterion* in criterion-referenced testing refers to. This confusion is understandable because two definitions have evolved for criterion. For some authors, the material that the students are supposed to learn in a particular course is the criterion against which they are being measured. For other authors, the term criterion refers to the **standard**, also called a criterion level or cut-point (see Chapter 10), against which each student's performance is judged. For instance, if the cut-point for passing a CRT is set at 70 percent, that is the criterion level.

Regardless of which version of the term is being applied in a given situation, the primary focus in interpreting CRT scores is on how much of the material each student has learned in absolute terms. For example, the following would be a characteristic CRT score interpretation: a student scored 85 percent, which means that the student knew 85 percent of the material. Notice that no reference is made to the performances of other students in that score interpretation.

In contrast, on an NRT, each student's performance is interpreted relative to the performances of the other students in the norm group. In fact, NRT scores are sometimes expressed with no reference to the actual number of test questions answered correctly. For example, the following would be a typical NRT score interpretation: a student scored in the 84th percentile, which means that the student scored better than 84 out of 100 students in the group as a whole (and by extension, worse than 16 out of 100 students). How many questions did the student answer correctly? We have no way of knowing because a percentile score only expresses the student's position relative to the other students.

One key to understanding the difference between NRT and CRT score interpretations is captured in the terms percentage and percentile. On CRTs, teachers are primarily concerned with how much of the material the students know. That is, they focus on the **percentage** of material known, which tells them the proportion that each student has learned without reference to the performances of the other students. In other words, the teachers only care about the percentage of questions the students answered correctly (or percentage of tasks the students correctly completed) in connection with the material at hand and perhaps in relationship to a previously established criterion level. The percentages are interpreted directly without reference to the students' positions vis-à-vis each other. Hence, a high percentage score means that the test was easy for the students, which may in turn mean that the students knew the material being tested very well or that the test questions were written at too low a level. Similarly, a low percentage score means that the test was difficult for the students, which may in turn mean that the students did not know the material being tested or that the test questions were written at too high a level of difficulty.

On NRTs, the concern is entirely different. Teachers focus instead on how each student's performance relates to the performances of all other students. Thus, in one way or another, they are interested in the student's **percentile** score, which tells them the proportion of students who scored above and below the student in question. For instance, a student with a percentile score of 70 performed better than 70 out of 100 students but worse than 30 out of 100. If another NRT were administered to the same students but had much more difficult questions on it, the percentage of correct answers would be lower for all students, but their positions relative to each other in terms of percentile scores could be virtually the same. Similarly, if another NRT had easy questions on it, the percentage of correct answers would be high for all students, but their positions relative to each other in terms of percentile scores could be very similar.

In short, CRTs look at the amount of material known by the students in percentage terms, while NRTs examine the relationship of a given student's performance to those of all other students in percentile terms.

Type of measurement

With regard to type of measurement, NRTs are typically most suitable for measuring general abilities. Examples would include reading ability in French, listening comprehension in Chinese, and overall English language proficiency. The *Test of English as a Foreign Language*, more commonly known as the TOEFL, is a good example of such a test. While the TOEFL paper-and-pencil version does have three subtests measuring listening comprehension, writing and analysis, and reading comprehension and vocabulary (ETS 2002a, 2003), the computer-based TOEFL has four subtests: listening, structure, reading, and writing (ETS 2000)—all of which must necessarily be considered general abilities.

In contrast, CRTs are better suited to providing precise information about each individual's performance on well-defined learning points. For instance, if a language course focuses on a structural syllabus, the CRT for that course might contain four subtests on: subject pronouns, the *a/an* distinction, the third person *-s*, and the use of present tense copula. However, CRTs are not limited to grammar points. Subtests on a CRT for a notional-functional language course might consist of a short interview where ratings are made of the student's abilities to: perform greetings, agree or disagree, express an opinion, and end a conversation. The variety and types of test questions used on a CRT are limited only by the imagination of the test developer(s).

Purpose of the testing

In terms of the purpose of the testing, major differences clearly exist in the way scores are interpreted on NRTs and CRTs. As mentioned above, NRT interpretations are relative, while CRT interpretations are absolute. The purpose of an NRT is, therefore, to generate scores that spread the students out along a continuum of general abilities so that any existing differences among the individuals can be distinguished. However, since the purpose of a CRT is to assess the amount of knowledge or skill learned by each student, the focus is on the individuals' knowledge or skills, not on distributions of scores. As a result, the distributions of scores for NRTs and CRTs can be quite different.

Distributions of scores

Since NRTs must be constructed to spread students out along a continuum or distribution of scores, the manner in which test questions for an NRT are generated, analyzed, selected, and refined (see Chapter 4) will usually lead to a test that produces scores which fall into a normal distribution. Such a distribution is desirable so that any existing differences among the students will be clearly revealed. For instance, if you want your students to be accurately placed into levels of study in your institution, you would want to do so on the basis of tests that reveal clear differences in their abilities. In other words, if there is variation in the group with regard to the knowledge or skill being tested, any differences among students should be reflected in their scores so the students will be placed in a fair and equitable manner.

In contrast, on a criterion-referenced final examination, students who have learned all the course material should all be able to score 100 percent on the final examination. Thus, very homogeneous scores can occur on a CRT. In other words, very similar scores among students on a CRT may be perfectly logical, acceptable, and even desirable if the test is administered at the end of a course. In this situation, a normal distribution of scores may not appear. In fact, a normal distribution in CRT scores may even be a sign that something is wrong with the test, with the curriculum, or with the teaching (see Chapters 4, 5, & 11).

Test structure

Differences also arise in the test structure for the two families of tests. Early on, Popham and Husek (1969) contended that "…it is not possible to tell a NRT from a CRT by looking at it." However, even though you may not be able to tell whether an item is NRT or CRT in orientation by looking at it, I would argue that you can tell an NRT from a CRT in terms of the structure and organization of the test. Typically, an NRT is relatively long and contains a wide variety of question content types. Indeed, the content can be so diverse that students find it difficult to know exactly what is being tested. Such a test is usually made up of a few subtests on rather general language skills like reading comprehension, listening comprehension, grammar, writing, and so forth. Each of these subtests is relatively long (30–50 questions) and covers a wide variety of different contents.

In contrast, CRTs usually consist of numerous shorter subtests. Each subtest will typically represent a different instructional objective. If a course has twelve instructional objectives, the associated CRT will usually have twelve subtests. Sometimes, in courses with many objectives, for reasons of practicality, only a sub-sample of the objectives will be tested. For example, in a course with 30 objectives, it might be necessary due to time constraints to randomly select 15 of the objectives for

testing, or to pick the 15 most important objectives (as judged by the teachers). Because of the number of subtests involved in most CRTs, the subtests are usually kept short (i.e., three to ten test items).

For reasons of economy of time and effort, the subtests on a CRT will sometimes be collapsed together, which makes it difficult for an outsider to identify the subtests. For example, on a reading comprehension test, the students might be required to read five passages and answer four multiple-choice questions on each passage. If on each passage there is one fact question, one vocabulary question, one cohesive device question, and one inference question, the teachers will most likely consider the five fact questions (across the five passages) together as one subtest, the five vocabulary questions together as another subtest, the five cohesive device questions together as yet another subtest, and the five inference questions together as the last subtest. In other words, the teachers will be focusing on the question types as subtests, not the passages, and this fact might not be obvious to an outside observer.

Finally, the two families of tests differ in the *knowledge of the questions* that students are expected to have. Students rarely know in any detail what content to expect on an NRT. In general, they might know what question formats to expect (for example, multiple-choice, true-false, and so forth), but seldom would the actual language points be predictable. This unpredictability of the question content results from the general nature of what NRTs are measuring and the wide variety of question content types that are typically used.

On a CRT, good teaching practice is more likely to lead to a situation in which the students can predict not only the question formats on the test, but also the language points that will be tested. If the instructional objectives for a course are clearly stated, if the students are given those objectives, if the objectives are addressed by the teacher, and if the language points involved are adequately practiced and learned, then the students should know exactly what to expect on the test, unless for some reason the criterion-referenced test is not properly referenced to the criteria (i.e., the instructional objectives).

This can often lead to complaints that the development of CRTs will cause teachers to "teach to the test" to the exclusion of other more important ways of spending classroom time. While I acknowledge that not all elements of the teaching and learning process can be tested, I argue that teaching to the test should nevertheless be a major part of what teachers do. If the objectives of a language course are worthwhile and have been properly constructed to reflect the needs of the students, then tests based on those objectives should reflect the important language points that are being taught. Teaching to such a test should help teachers and students stay on track, and the test results should provide useful feedback to both groups on the effectiveness of the teaching and learning processes. In short, teaching to the test, if the test is a well-developed CRT, should help the teacher and students rather than constrain them.

A very useful side effect of teaching to the test is that the information gained can have what Oller (1979, p. 52) termed **instructional value**, that is, the test-derived information can "enhance the delivery of instruction in student populations." In other words, such CRTs can provide useful information for evaluating the effectiveness of the needs analysis, the objectives, the tests themselves, the materials, the teaching, the students study habits, and so forth. In short, CRTs will prove enlightening in the never-ending evaluation process (see Brown 1995a).

I am not arguing that teachers should only address a very restricted set of objectives in a language course. Flexibility and time must be allowed in any curriculum for the teachers to address problems and learning points that arise along the way.

Nevertheless, if a common core of objectives can be developed for a course, a CRT can then be developed to test those objectives, and a number of benefits will accrue to the teachers, the students, and the curriculum developers alike (see Brown 1995a).

CRTs are not better than NRTs. Both categories of tests are very important for the decision-making processes in a language program, but for different types of decisions. Understanding the distinction between NRTs and CRTs can help teachers to match the correct type of test with any decision purpose.

Matching Tests to Decision Purposes

A variety of decisions are made in almost any language program, and language tests of various kinds can help in making such decisions (e.g., placement decisions, pass/fail decisions, etc.). In order to test appropriately, administrators and teachers must be very clear about their purpose for making a given decision and then match the correct type of test to that purpose. In this section, I will summarize the main points that administrators and teachers must keep in mind when matching the appropriate measuring tool (NRT or CRT) with the types of decisions they must make about their students. The main points to consider are shown in Table 1.2. As the discussion develops, I will briefly cover each point as it applies to four types of decisions.

Table 1.2 Matching tests to decision purposes

Test Qualities	***Type of Decision***			
	Norm-Referenced		Criterion-Referenced	
	Proficiency	Placement	Achievement	Diagnostic
Detail of Information	Very general	General	Specific	Very specific
Focus	Usually general skills prerequisite to entry	Learning points from all levels & skills of program	Terminal objectives of course or program	Terminal and enabling objectives of courses
Purpose of Decision	To compare an individual's overall ability with other individuals	To find each student's appropriate level	To determine the degree of learning for advancement or graduation	To inform students and teachers of objectives needing more work
Relationship to Program	Comparisons with other institutions or programs	Comparisons within program	Directly related to objectives	Directly related to objectives still needing work
When Administered	Before entry and sometimes at exit	Beginning of program	End of courses	Beginning and/or middle of courses
Interpretation of Scores	Spread of wide range of scores	Spread of narrower, program-specific range of scores	Overall number and percentage of objectives learned	Percentage of each objective in terms of strengths and weaknesses

In administering and teaching in language programs, I have found myself making four basic kinds of decisions: proficiency, placement, achievement, and diagnostic. Since these are also the four types of tests identified in Alderson, Krahnke, and Stansfield (1987) as the most commonly used types of tests in our field, I will call them the primary **language testing functions** and focus on them in the remainder of this chapter. These testing functions correspond neatly to the NRT and CRT categories as follows: NRTs help in making program-level decisions (proficiency and placement), and CRTs are useful in making classroom-level decisions (diagnostic and achievement). They provide a useful framework for thinking about decision making in language programs.

Generally speaking, the program-level proficiency decisions (usually for admissions) and placement decisions are the prerogative of administrators. That is, administrators are most interested in and usually responsible for seeing to it that students are properly admitted to their institutions, and then that students are properly placed in the correct level of study. In contrast, the classroom-level decisions for diagnosis and achievement are the prerogative of classroom teachers. That is, teachers are usually most interested in and responsible for determining the individual student's strengths and weaknesses through diagnostic testing and the individual student's level of attainment through achievement testing.

Of course, other categories of tests do exist. For instance, aptitude tests, intelligence tests, learning strategy tests, and attitude tests do not fit neatly into these four language testing functions. However, those other types of tests are not generally administered in language programs, and so are not relevant to the topic of this book.

Program-level proficiency decisions

Sometimes, administrators need to make decisions based on the students' general levels of language proficiency. The focus of such decisions is usually on the general knowledge or skills prerequisite to entry or exit from some type of institution, for example, American universities. Such **proficiency decisions** are necessary in setting up entrance and exit standards for a curriculum, in adjusting the level of program objectives to the students' abilities, or in making comparisons between programs. Proficiency decisions are often based on proficiency tests specifically designed for such decisions. By definition, then, **proficiency tests** assess the general knowledge or skills commonly required or prerequisite to entry into (or exemption from) a group of similar institutions. One example is the *Test of English as a Foreign Language* (TOEFL), which is used by many American universities that have English language proficiency prerequisites in common (see ETS 1997, 2000, 2001, and 2002a). Understandably, such tests are very general in nature and cannot be related to the goals and objectives of any particular language program. Another example of the general nature of proficiency tests is the *ACTFL Proficiency Guidelines* from the American Council on the Teaching of Foreign Languages (ACTFL 1986, 2004). Though proficiency tests may contain subtests for different language skills, the testing of those skills remains very general, and the resulting scores can only serve as overall indicators of proficiency.

Since proficiency decisions require knowing the general level of proficiency of language students in comparison to other students, the test must provide scores that form a wide distribution so that interpretations of the differences among students will be as fair as possible. Thus, proficiency decisions should be made on the basis of norm-referenced proficiency tests, because NRTs have all the qualities desirable for such decisions (see Table 1.1, p. 3).

Proficiency decisions based on large scale standardized tests may sometimes seem unfair to teachers and administrators because of the arbitrary way that they are handled in some settings. However, such proficiency decisions are often necessary: to protect the integrity of the institutions involved, to keep students from getting in over their heads, and to prevent students from entering programs that they really do not need.

Proficiency decisions most often occur when a program must relate to the external world in some way. The students are arriving. How will they fit into the program? And when the students leave the program. Is their level of proficiency high enough so they can succeed linguistically in other institutions?

Sometimes, comparisons are also made among different language programs. For instance, since proficiency tests, by definition, are general in nature, rather than geared to any particular program, they could serve to compare regional branches of a particular language teaching system. Consider what would happen if the central office for a nationwide chain of ESL business English schools wanted to compare the effectiveness of all its centers. To make such decisions about the relative merit of the various centers, the administrators in charge would probably want to use some form of business English proficiency test.

Because such tests are not geared to any particular language program, extreme care must be exercised in making comparisons among different language programs. By chance, the test could fit the teaching and content of one program relatively closely, and as a consequence, the students in that program might score high on average. By chance, the test might not match the curriculum of another program quite so well, and consequently, the students would score low on that particular proficiency test. The question is: Should one program be judged less effective than another simply because the teaching and learning that is going on in that program (though perfectly effective and useful) is not adequately assessed by the test? Of course not. Hence, **program fair tests** (after Baretta 1986) must be used in such comparisons. That is, great care must be used in making such comparisons to make sure the test(s) involved appropriately match the curriculum goals and objectives of the programs involved.

Because of the general nature of proficiency decisions, a proficiency test must be designed so that the general abilities or skills of students are reflected in a wide distribution of scores. Only with such a wide distribution can decision-makers make fair comparisons among the students, or groups of students. This need for a wide spread of scores most often leads testers to create tests that produce normal distributions of scores. All of which is to argue that proficiency tests should usually be norm-referenced.

Proficiency decisions should never be undertaken lightly. Instead, these decisions must be based on the best obtainable proficiency test scores as well as other multiple sources of information about the students (for example, other test scores, grade point averages, interviews, recommendation letters, statements of purpose, research papers written by the students, etc.). Proficiency decisions can dramatically affect students' lives, so slipshod decision making in this area would be particularly unprofessional.

Program-level placement decisions

Placement decisions usually have the goal of grouping students of similar ability levels together. Teachers benefit from placement decisions because they end up with classes that have students with relatively homogeneous ability levels. As a result, teachers can focus on the problems and learning points appropriate for that level of students. To that end, placement tests are designed to help decide what each student's

appropriate level will be within a specific program, skill area, or course. The purpose of such tests is to reveal which students have more or less of a particular knowledge or skill so that students with similar levels of ability can be grouped together.

Examining the similarities and differences between proficiency and placement testing will help to clarify the role of placement tests. At first glance, a proficiency test and a placement test might look very similar because they are both testing fairly general material. However, a proficiency test will tend to be very, very general in character because it is designed to assess extremely wide bands of abilities, from say beginning to near-native-speaker levels. In contrast, **placement tests** must be more specifically related to a given program, particularly in terms of the relatively narrow range of abilities assessed and the content of the curriculum, so that it efficiently separates the students into level groupings within that program.

Put another way, a general proficiency test might be useful for determining which language program is most appropriate for a student, but once in that program, a placement test would be necessary to determine the level of study that the student would most benefit from. Both proficiency and placement tests should be norm-referenced instruments because decisions must be made on the students' relative knowledge or skill levels. However, the degree to which a test is effective in spreading students out is directly related to the degree to which that test fits the ability levels of the students.

Consider, for example, the English Language Institute (ELI) at the University of Hawaii at Manoa (UHM). All the international students at UHM have been fully admitted by the time they arrive. In order to have been admitted, they must have taken the TOEFL (a proficiency test) and scored at least 500 on the paper-and-pencil version (or 173 on the computer-based version). From the ELI's point of view, language proficiency test scores are used to determine whether these students are eligible to study in the ELI and follow a few courses at UHM. Those students who score 600 or above on the paper-and-pencil TOEFL (or 250 on the computer-based version) are told that they are completely exempt from ELI training. Thus, I can safely say that most of the ELI students at UHM have scored between 500 and 600 on the paper-and-pencil TOEFL or between 173 and 250 on the computer-based version.

Within the ELI, there are three tracks, each of which is focused on one skill (reading, writing, or listening) with two skill levels in each track. As a result, the placement decisions and the tests upon which they are based must be much more focused than the information provided by TOEFL scores. The placement tests must provide information on each of the three skills involved as well as on the language needed by students in the relatively narrow proficiency range reflected in their TOEFL scores. While the contrasts between proficiency and placement decisions may not be quite so clear in all programs, these definitions and ways of distinguishing between proficiency and placement decisions should help teachers and administrators think about the program level decisions and testing in their own language programs.

If a particular program is designed with levels that include true beginners as well as very advanced learners, a general proficiency test *might* adequately serve as a placement test. However, such a wide range of abilities is not common in most language programs and, even when appropriately measuring such general abilities, each test must be examined in terms of how well it fits the abilities of the students and how well it matches what is actually taught in the classrooms.

If there is a mismatch between the placement test and what is taught in a program (as reported in Brown 1981), the danger is that the groupings of similar ability levels

will simply not occur. For instance, consider an elementary school ESL program in which a general grammar test is used for placement. If the focus of the program is on oral communication at three levels and a pencil-and-paper test is used to place the children into those levels, numerous problems may arise. Such a test is placing the children into levels on the basis of their *written grammar* abilities. While grammar ability may be related to oral proficiency, other factors may be more important to successful oral communication. The result of such testing practices might be that the oral abilities of the children in all three of the (grammar-placed) levels could turn out to be about the same on average.

Some form of oral placement procedure, for example, the oral proficiency scale of the American Council on the Teaching of Foreign Languages (ACTFL 1986, 2004), might more accurately separate the children into three ability-level groups for the purposes of teaching them oral communication skills. However, the ACTFL scale was designed for assessing overall language proficiency and, therefore, may be too general for making responsible placement decisions in this particular elementary school program. In addition, the ACTFL scale may only be tangentially related to the goals and purposes of this particular school. Most importantly, the ACTFL scale was designed with adult university students in mind so it may not be at all appropriate for elementary school children. Clearly then, the purpose of a program, the range of abilities within the program, and the type of students involved are all factors that may make a proficiency test inappropriate for purposes of testing placement. Typically, placement decisions should be based on placement tests that have either been designed with a specific program in mind or been seriously examined for their appropriateness for the program in question.

Classroom-level achievement decisions

All language teachers are in the business of fostering achievement in the form of language learning. In fact, the purpose of most language programs is to maximize the possibilities for students to achieve a high degree of language learning. As a result, most language teachers will sooner or later find themselves interested in making achievement decisions. **Achievement decisions** are decisions about the amount of learning that students have accomplished. Such tests are typically administered at the end of the term, and such decisions may take the form of deciding which students will be advanced to the next level of study, determining which students should graduate, or simply for grading the students. Teachers may find themselves wanting to make rational decisions that will help improve their students' achievement. Or they may need to make and justify changes in curriculum design, staffing, facilities, materials, equipment, and so on. Such decisions should most often be made with the help of achievement test scores.

Making decisions about the achievement of students and about ways to improve their achievement will at least partly involve testing to find out how much each person has learned within the program. Thus, **achievement tests** should be designed with very specific reference to a particular course. This link with a specific course usually means that the achievement tests will be directly based on course objectives and will therefore be criterion-referenced. Such tests will typically be administered at the end of a course to determine how effectively students have mastered the instructional objectives.

Achievement tests must not only be very specifically designed to measure the objectives of a given course, but also must be flexible enough to help teachers readily respond to what they learn from the tests about the students' abilities, the students' needs, and the students' learning of the course objectives. In other words, a good achievement test can tell teachers a great deal about their students' achievement *and*

about the adequacy of the course. Hence, while achievement tests should definitely be used to make decisions about students' levels of learning, they can also be used to affect curriculum changes.

Classroom-level diagnostic decisions

From time to time, teachers may also take an interest in assessing the strengths and weaknesses of each individual student in terms of the instructional objectives for the purpose of correcting an individual's deficiencies "before it is too late." To that end, **diagnostic decisions** are typically made at the beginning or middle of the term and are aimed at fostering achievement by promoting strengths and eliminating the weaknesses of individual students. Naturally, the primary concern of the teacher must be the entire group of students collectively, but some attention can also be given to each individual student. Clearly, this last category of decision is concerned with diagnosing problems that students may be having in the learning process. While diagnostic decisions are definitely related to achievement, diagnostic testing often requires more detailed information about which specific objectives students can already do well and which they still need to work on. The purpose is to help students and their teachers to focus their efforts where they will be most effective.

As with achievement tests, **diagnostic tests** are designed to determine the degree to which the specific instructional objectives of the course have already been accomplished. Hence, they should be criterion-referenced in nature. While achievement decisions are usually focused on the degree to which the objectives have been accomplished at the end of the program or course, diagnostic decisions are normally made along the way as the students are learning the language. As a result, diagnostic tests are typically administered at the beginning or in the middle of a language course. In fact, if well constructed to reflect the instructional objectives, one CRT in three equivalent forms could serve as a diagnostic tool at the beginning and midpoints in a course and as an achievement test at the end.

Perhaps the most effective use of a diagnostic test is to report the performance level on each objective (in a percentage) to each student so that they can decide how and where to most profitably invest their time and energy. For example, telling a student that she scored 100 percent on the first objective (selecting the main idea of a paragraph) but only 20 percent on the second objective (guessing vocabulary from context) would tell that student that she is good at finding the main idea of a paragraph but needs to focus her energy on guessing vocabulary from context.

It would also be useful to report the average performance level for each class on each objective (in percentage terms) to the teacher(s) along with indications of which students have particular strengths or weaknesses on each objective.

Why a Single Test Cannot Fulfill All Four Functions

In my various contacts with language educators around the world, I have found that what many administrators and teachers would really like would be a proficiency-placement-diagnostic-achievement test that they could use for all kinds of decisions. Wouldn't that be wonderful? Why can't we have such a proficiency-placement-diagnostic-achievement test? Basically, there are at least two reasons why such a test could never be created: differences in ranges of ability and differences in variety of content.

Differences in ranges of ability

First, the ranges of ability tested by the four types of tests are very different. Typically, norm-referenced proficiency tests are designed to measure a very wide range of abilities as represented by the entire width of the outside box in Figure 1.1. In English for instance, the paper-and-pencil TOEFL measures from virtually no English (that is to say "guessing on the test") at 200 to native, or native-like ability at 677. That range is appropriate for passing students from institution to institution for admissions decisions and for comparing different institutions.

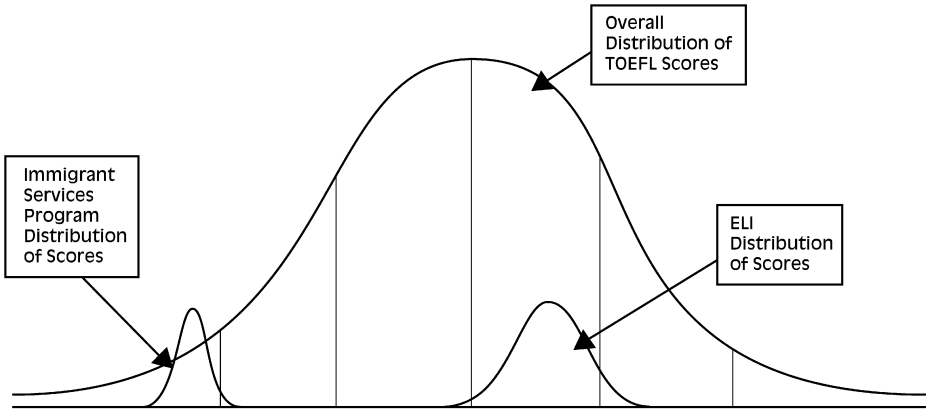

Figure 1.1 Distributions on the TOEFL for various groups of students

Placement tests would normally be very different in the range of abilities they assess, usually limited to the range of abilities handled by the particular institution involved. For example, Figure 1.1 shows the overall distribution of scores on the TOEFL proficiency test, and inside that distribution the distribution of TOEFL scores for two different institutions, one a survival-level ESL immigrant services program and the other a university English language institute (ELI). In both cases, the ranges of abilities within each of the institutions are much narrower than the range of abilities on the overall TOEFL and are very different from each other in terms of the overall abilities of the groups of students. Thus, to make placement decisions within either of these institutions, a much more narrowly focused placement test would be necessary. Also, note that a placement test developed for one institution would not be appropriate in level for the students in the other institution.

In addition, a proficiency test like the TOEFL would not be appropriate for such placement decisions for two reasons. One, because a proficiency test is designed to measure a very wide range of abilities, many of the items would be far too easy or far too difficult, or both, for the students in the particular institution. Two, because only a small subset of items would actually be at the appropriate level for placement decisions; the test items that discriminate at all would probably be too few to provide reliable enough measurement for making responsible decisions (see Chapter 9).

Conversely, a placement test designed for the specific range of abilities in a particular institution would not be of much use in making proficiency decisions between institutions. In other words, a placement test designed for a particular institution would probably not have a wide enough range of item difficulties to be useful for making admissions decisions that must by definition include students with wide spans of abilities from many different institutions.

The criterion-referenced diagnostic and achievement tests in the courses of the immigrant services program or in the courses of the ELI would have to be even more narrowly defined in terms of the ranges of abilities they test because they would typically be developed to measure the very specific levels of material taught in the particular courses within the particular institution. Hence, the range of abilities would be even narrower than that for the placement test used to put the students into that level of study, not to mention the proficiency test that was used to put the students in that institution. It wouldn't make any sense to use a very broad scale inter-institutional proficiency test, or even an inter-course placement test for diagnostic or achievement testing in a particular course. Conversely, such a course diagnostic or achievement test would be far too narrowly defined for placement into different courses, or for proficiency testing across institutions.

Differences in variety of content

The content of a proficiency test is also very broadly defined so that it will not favor one institution or another, but rather will cover the whole range of content types and ability levels covered across many institutions. In contrast, the content of a placement test should be more narrowly defined to meet the needs of the particular program in which it is being used. For instance, if one program has an overall grammar-translation orientation, the placement test should reflect that orientation across the range of abilities of the students within that program; if another program has a task-based orientation, the placement test should reflect that orientation, again across the appropriate range of abilities of the students in that program.

The content of diagnostic or achievement tests for a course should be even more narrowly defined to reflect the exact content of the course, perhaps as expressed in the goals and objectives for that course. For example, if an intermediate reading course has 15 intensive reading objectives and one extensive reading objective, the 15 *intensive reading objectives* (e.g., getting the main idea, reading for facts, reading for inferences, identifying language functions, etc.) should be directly reflected on the diagnostic and achievement tests with perhaps three items for each, while the single *extensive reading objective* (e.g., each student will read at least three books from the library) might take quite a different form (e.g., student summaries of the three books they read), because teachers simply need to check off something as students turn in their summaries in order to verify that each student has accomplished the objective. Thus, simply checking off the students' achievement of the objective can become a part of the testing system.

As with ability ranges, the variety of *contents* in an inter-institutional proficiency test would not be appropriate for placement testing, nor would the *contents* in a proficiency or placement test be appropriate for diagnostic or achievement purposes in most programs or courses. Conversely, a single course diagnostic or achievement test would be far too narrowly defined to use for placement into multiple courses, just as either a course or placement test would be far too narrowly defined to use for proficiency testing across institutions.

All in all, norm-referenced proficiency and placement tests have important, but different roles to play in language education vis-à-vis the ranges of ability and types of contents involved. Criterion-referenced diagnostic and achievement tests also have important roles in language education that are quite different from each other and different from proficiency or placement tests, again vis-à-vis the ranges of ability and types of contents involved. Trying to mix these purposes is likely to simply make a mess. (For example, see what happens when a company attempts to use a proficiency test for achievement testing purposes in Childs 1995.)

To bring this discussion home to you, consider how you would prefer to be tested on the material in this book when you have finished reading it. Would you prefer a test that is designed to spread people out so the results will be a few grades of *F*, some *D*s, many *C*s, some *B*s, and a few *A*s? That would be a norm-referenced approach. Or would you prefer a purely criterion-referenced achievement test that measures your knowledge of the language testing concepts in this book, designed such that anyone who knows the material will score well? If you would prefer the latter approach, you cannot in good conscience advocate the use of a norm-referenced test for assessing your students' achievement.

USING SPREADSHEET PROGRAMS IN LANGUAGE TESTING

The statistical analyses in this book will be explained in conceptual terms such that the reader can do them with a pencil and paper, or a calculator if necessary. However, they will also be described in terms of how they can be done on a spreadsheet program, which is a much easier way to proceed. If you have never worked with a spreadsheet program, you might reasonably ask three questions: What is a spreadsheet program? How will you personally benefit from using a spreadsheet program in this book? How can you get started with your spreadsheet program?

What Is a Spreadsheet Program?

A **spreadsheet program** is a very flexible computer tool that allows you to enter rows and columns of numbers, then manipulate, analyze, and present them in any way you like. *Excel*™ (Microsoft, 2003) is the spreadsheet program that most people use today. Regardless of which computer platform or which version of *Excel* you use, the layout, menus, commands, functions, etc., are generally the same. Thus, because it is ubiquitous and fairly standard, *Excel* is the logical choice to use as the example program in this book. However, if you are using a different spreadsheet program (e.g., *Quattro Pro*™, *Lotus 1-2-3*™, etc.) the processes will be very similar, so you will be able to work by analogy as long as you have a copy of the manual and/or a good book explaining how to use that particular spreadsheet program.

How Will You Personally Benefit from Using a Spreadsheet Program in This Book?

You can use a spreadsheet to enter your students' responses to the items on a test, analyze those responses to see which items are working and which are not (as explained in Chapter 4), calculate the students' total scores and descriptive statistics as well as their standardized scores (see Chapters 5 and 6), work out the correlation between their scores on the test and those from some other measure (see Chapter 7), estimate the reliability or dependability of the test (see Chapters 8 and 9), investigate the **validity** of the test (see Chapter 10), and keep records of their progress through the entire language program (see Chapter 11). All of this will prove relatively easy when using a spreadsheet program and very useful for any language teacher or administrator. While many of the above uses of a spreadsheet program may sound very complicated and difficult, they will all be explained step-by-step in the subsequent chapters so that, before you know it, these concepts will all be clear to you and become tools you can use in your classroom or program-level testing projects.

In the next chapter, you will be asked to get on a computer, actually open such a spreadsheet program, and have a look around. So you might want to begin now to get access both to a computer and a spreadsheet program.[1] I'm sure you will enjoy using a spreadsheet once you learn how. One warning, however, spreadsheets can be so addictive that they have been known to ruin relationships, marriages, and lives. So please use your spreadsheet prudently and only with the utmost restraint.

[1] If you don't already have a spreadsheet program at home or at work, you might consider buying *Excel* or downloading a program from the Internet by searching the phrase "free spreadsheet." Naturally, the *Excel* spreadsheet program will better match the instructions in this book.

REVIEW QUESTIONS

1. For which type of test (NRT or CRT) would you expect the interpretation to be absolute? For which type would it be relative?
2. For which type of test (NRT or CRT) would you expect the scores to spread students out along a continuum of general abilities or proficiencies?
3. For which type of test (NRT or CRT) would you expect all the students to be able to score 100 percent if they knew all of what was taught?
4. For which type of test (NRT or CRT) would the students usually have little or no idea what content to expect in questions?
5. For which type of test (NRT or CRT) would you expect to find a series of short, well-defined subtests with fairly similar test questions in each?
6. For which type of decision (proficiency, placement, diagnostic, or achievement) would you use a test that is designed to find each student's appropriate level within a particular program?
7. For which type of decision (proficiency, placement, diagnostic, or achievement) would you use a test that is designed to inform students and teachers of objectives needing attention?
8. For which type of decision (proficiency, placement, diagnostic, or achievement) would you use a test that is designed to determine the degree of learning (with respect to the program objectives) that had taken place by the end of a course or program?
9. For which type of decision (proficiency, placement, diagnostic, or achievement) would you use a test that is designed to compare an individual's overall performance with that of groups/individuals at other institutions?
10. Do you think that the concepts behind CRTs and NRTs can be mixed into one test? In other words, do you think it is possible to create a proficiency-placement-diagnostic-achievement test? If so, why do you think that is desirable? And how on earth would you go about doing it?

APPLICATION EXERCISES

A. Consider a specific language teaching situation in an elementary school, a secondary school, a commercial language center, a university intensive program, or other language teaching setting. Think of one type of decision that administrators and teachers must make in that program. Decide what type of decision it is (proficiency, placement, diagnostic, or achievement).

B. Now describe the test that you would recommend using to make the decision that you selected in Question A. Decide what type of test you would use and what it should be like in terms of overall characteristics, as well as the skills tested, level of difficulty, length, administration time, scoring, and type of report given to teachers and students.

C. Best of all, if you have the opportunity, match a real test to a real decision in some language program; administer, score, interpret, and report the results of the test; and make or help others make the appropriate decisions so that they minimize any potential negative effects on the students' lives.

CHAPTER 2

ADOPTING, ADAPTING, AND DEVELOPING LANGUAGE TESTS

INTRODUCTION

Numerous considerations influence the kinds of choices teachers and administrators must make if they want to develop an effective testing program at their institution. I explore these considerations in this chapter as a series of theoretical and practical testing issues, each of which can be described and thought about separately. The theoretical issues include language teaching methodology issues, the distinction between competence and performance, and the difference between discrete-point and integrative tests. The practical issues include fairness issues, cost issues, and logistical issues.

Though they are discussed separately, all of these issues must be considered simultaneously when addressing the next topic of the chapter: whether you want to adopt, adapt, or develop language tests for your language program. After a brief discussion of the important factors necessary for putting sound tests in place, I will end the chapter by showing how to get started with your spreadsheet program.

THEORETICAL ISSUES

The **theoretical issues** that I will address have to do with what tests should look like and what they should do. These issues have a great deal to do with how a group of teachers feels their course or program fits pedagogically within the overall field of language teaching, and how well they communicate their beliefs about teaching and testing with each other. After all, it is only through communication that teachers can create curriculum and tests that are at least modestly coordinated within and between courses so that students do not face a bewildering array of disconnected teaching and testing methods.

Theoretical issues may include pedagogical beliefs in various language teaching methodologies ranging from grammar-translation to communicative language teaching, or beliefs in the relative importance of the skills that teachers teach and test in their program (written or oral, productive or receptive, and various combinations of the four). Other theoretical issues may range from the linguistic distinction between competence and performance to the purely testing distinction among the various types of tests that are available in language teaching. These test types range from what are called discrete-point to integrative tests and various combinations of the two. I will discuss each of these issues in turn, then, look at some of the ways in which they may interact with each other. Remember, they are theoretical viewpoints on what tests should look like and what they should do.

One problem that arises is that language teaching professionals often disagree on these issues. Since tests are instruments developed by people to make decisions about other people, test development and test administration are inherently political activities. Thus, the policies of a given program on the various testing issues should be decided

consciously and purposefully by the teachers and administrators involved, whether by consensus, by majority vote, or by executive decree. Regardless of the strategy used, healthy discussions can help clarify the issues involved whenever new tests are put into place. Recognizing the political nature of testing early in the process can stave off many problems later.

Language Teaching Methodology Issues

Since views of what constitutes good language teaching vary widely throughout the profession, ideas about what constitutes good testing (or a good test) will also differ. Consider how a teacher like the mythical Miss Fiditch (of the granny glasses, hair-in-a-bun, ruler-in-hand, structuralist school of language teaching) might argue with the much more real, and realistic, Sandra Savignon, one of the early advocates of communicative teaching and testing (see Savignon 1972, 1985; Bachman & Savignon 1986). Miss Fiditch would tolerate only strict testing of knowledge of grammar rules, probably having students translate a selection from one of the "great books" of the target language into their mother tongue. In contrast, Savignon (1972) advocated testing "the students' ability to communicate in four different communicative contexts: discussion, information-getting, reporting, and description" (p. 41). How did language testing get from the extreme views of Miss Fiditch to the more modern views of Savignon?

An exceptionally short history of language testing

Spolsky (1978) and Hinofotis (1981) both pointed out early on that language testing can be broken into periods, or trends, of development. Hinofotis labeled them the prescientific period, the psychometric-structuralist period, and the integrative-sociolinguistic period. As shown in Table 2.1, I will use the term **movements** instead of periods to describe them because these movements overlap chronologically and can be said to all co-exist today in different parts of the world. I will also add one movement, which I will label the communicative movement. (For very different takes on the history of language testing, see Spolsky 1995 and Barnwell 1996.)

Table 2.1 Language testing movements

Testing Movement	Linguistic Basis
Prescientific	Ability to translate
Psychometric-structuralist	Ability to manipulate grammatical structures
Integrative-sociolinguistic	Ability to use sociolinguistic aspects of language
Communicative	Ability to communicate functions/notions and perform tasks with language

The **prescientific movement** in language testing is associated with the grammar-translation approaches to language teaching. Since such approaches have existed for ages, the end of this movement is usually delimited rather than its beginning. I infer

from Hinofotis's article that the prescientific movement ended with the onset of the psychometric-structuralist movement, but clearly such movements have no end in language teaching because, without a doubt, such teaching and testing practices are going in many places in the world today (e.g., the current grammar-translation tests in the *yakudoku* language teaching tradition found in many of Japan's prestigious high school and university entrance examinations; see Brown & Yamashita 1995a & 1995b; Brown 1996b, 1999a).

The prescientific movement is characterized by translation and essay tests developed exclusively by the classroom teachers, who are on their own when it comes to developing and scoring tests. One problem that arises with these types of tests is that they are relatively difficult to score objectively. Thus, subjectivity becomes an important factor in scoring such tests. Perhaps mercifully, no language testing specialists were involved in the prescientific movement. Hence, there was little concern with the application of statistical techniques such as item analysis, descriptive statistics, reliability coefficients, validity studies, and so forth (see Chapters 4 to 10). Some teachers may think back to such a situation with a certain nostalgia for its simplicity, but along with the lack of concern with statistics came an attendant lack of concern with concepts like objectivity, reliability, and validity, that is, a lack of concern with making fair, consistent, and correct decisions about the lives of the students involved. Most teachers would protect their own students from such unfair testing practices and would complain even more vigorously if such lax practices were applied to themselves as students in a teacher training course. How would you like to have to show your knowledge of the material in this book (after you have read it) by taking a test that is subjective, inconsistent, and based on material unrelated to the book? That would seem unfair, right? Wouldn't any decisions based on such a test be unreliable, arbitrary, and unfair? Those are the types of problems the next movement was designed to rectify.

With the onset of the **psychometric-structuralist movement** of language testing, worries about the objectivity, reliability, and validity of tests began to arise. Psychological and educational measurement specialists interacted with linguists, and language tests were created that were increasingly scientific, reliable, and precise, that is to say, they were state-of-the-art for their day. Psychometric-structuralist tests typically set out to measure the discrete structural points (Carroll 1972) being taught in the audio-lingual and related teaching methods of the time. Like the language teaching methods of the day, these tests were influenced by behavioral psychology. The psychometric-structuralist movement saw the rise of the first carefully designed and standardized tests like the *Test of English as a Foreign Language* (first introduced in 1963), the *Michigan Test of English Language Proficiency: Form A* (University of Michigan 1961), *Modern Language Association Foreign Language Proficiency Tests for Teachers and Advanced Students* (ETS 1968), *Comprehensive English Language Test for Speakers of English as a Second Language* (Harris & Palmer 1970), and others. Such tests, usually in multiple-choice format, are easy to administer and score and are carefully constructed to be objective, reliable, and valid. Thus, they were felt to be an improvement on the test design and scoring practices of the prescientific movement.

The psychometric-structuralist movement is important because, for the first time, language test development follows scientific principles. In addition, psychometric-structuralist test development is squarely in the hands of trained linguists and language testers. As a result, statistical analyses are used for the first time (as described in Lado 1961). Psychometric-structuralist tests are still very much in evidence around the