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*^{TM} (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*^{TM}, *Lotus 1-2-3*^{TM}, 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 world, but they have been supplemented (and in some cases, supplanted) by what Carroll (1972) labeled integrative tests.

The **integrative movement** has its roots in the argument that language is creative. More precisely, language professionals began to believe that language is more than the sum of the discrete parts being tested during the psychometric-structuralist movement. Beginning with the work of sociolinguists like Hymes (1967a), it was felt that the development of communicative competence depended on more than simple grammar control; communicative competence also hinged on knowledge of the language appropriate for different situations. Tests typical of this movement were the cloze test and dictation, both of which assess the student's ability to manipulate language within a context of extended text rather than in a collection of discrete-point questions. The possibility of testing language in context led to further arguments for the benefits of integrative tests with regard to **pragmatics,** the ways that linguistic and extra-linguistic elements of language are interrelated and relevant to human experience (see Oller 1979). The integrative-sociolinguistic movement is probably most important because it questions the linguistic assumptions of the previous structuralist movement, yet uses the psychometric tools made available by that movement to explore language testing techniques designed to assess contextualized language.

In Hinofotis's discussion of trends for the 1980s, she suggests that the influence of notional-functional syllabuses and English for specific purposes have added new elements to language testing including new attempts to define communicative competence. She refers to Brière (1979) and Canale and Swain (1981). I will include this sort of testing here as the **communicative movement**, and expand her references to include at least Savignon (1972), Canale and Swain (1980), Canale (1983a & b), and Bachman (1990). I will go into more detail on this movement because it is the current bandwagon of choice and because, in my view, it is still developing. (For different perspectives on these issues, see Allison 1999, pp. 42–56; Brown, Hudson, Norris, & Bonk 2002; or the articles in Norris 2002.)

The communicative tests advocated within this movement were new and different in the 1980s, because they promoted certain characteristics that initially proved novel. Tests typical of this movement would include role plays, problem-solving tests, group tests, and task-based tests. Based on my reading and experiences trying to create such communicative tests, I would list their characteristics in two categories as shown in Table 2.2: test-setting requirements and bases for ratings. As for the communicative test-setting requirements, insofar as possible, the communication that is required of the students should be meaningful to the students as individuals, that is, it should include functions of the language that are useful to them. Also, in order for communication to be meaningful, it will probably be necessary to create a situation that is as authentic as possible. Moreover, the students should encounter unpredictable language input and be put in a position where they must produce creative language output (in the same sense that language input in real life is unpredictable and therefore language output must be creative, whether in a first or second language). Finally, just like in real life, students should be using all four language skills, including reading, writing, listening, and speaking.

Table 2.2 Characteristics of communicative tests

Communicative test-setting requirements:

Meaningful communication
Authentic situation
Unpredictable language input
Creative language output
All language skills (including reading, writing, listening, & speaking)

Bases for ratings:

Success in getting meanings across
Use focus rather than usage
New components to be rated

Three characteristics exist for the bases for rating such tests. Because of the need to somehow assign a score or grade for feedback on such productive and oral tests, ratings by teachers or testers become a normal part of the testing process. To begin with, those ratings should also be based, at least to some degree, on students' relative *success* in getting their meanings across. In addition, the ratings should focus on language *use* rather than usage, which means in some cases that the focus is on fluency rather than accuracy. Finally, ratings should perhaps include *new rating components* (in addition to the traditional phonemes/graphemes, vocabulary, and grammar) like suprasegmentals, paralinguistic features, proxemics, pragmatics, strategy use, and so forth. (For more on these topics see the feedback scales in Mendelsohn 1992; Brown (with contributions by LAIRDIL) 1995; or Brown 1996c.) In short, a communicative test would necessarily create a situation involving "...a coming together of organized knowledge structures with a set of procedures for adapting this knowledge to solve new problems of communication that do not have ready-made and tailored solutions" (Candlin 1986, p. 40).

To clarify by counter-example, during a meeting about communicative testing, a language teacher at my university once volunteered that he was already doing communicative testing because he had his students memorize dialogues and perform them in front of the class. Unfortunately, though his dialogue "communicative test" was oral and productive, it did not require any meaningful communication on the part of the students. It was not set in an authentic situation, had no unpredictable or creative elements at all, and was not rated for anything but accuracy. Hence, it clearly does not qualify as a communicative test, at least as that sort of test is defined here.

In addition to the test-setting and rating characteristics of communicative tests, they are sometimes discussed in terms of the components of language that they should assess. For instance Candlin (1986) cites Hymes (1967b; 1972) augmented view of the components of communicative competence (pp. 40-41), which included grammar, semantics, and sociolinguistic components. He also cites Halliday's (1979) model of communicative competence (pp. 42-44), which included textual (linguistic), ideational (semantic), interpersonal (pragmatic), and discoursal "capacity" (psycholinguistic) components.

Probably the best known model of the components of communicative competence is the one offered by Canale and Swain (Canale & Swain 1980; Canale 1983a & b). The version of that model outlined in Table 2.3 (from Canale 1983b) is still relevant today. Notice that under grammatical competence the model covers the elements of language that have traditionally been taught, that is, the ones that even lay people recognize as important aspects of language: phonology, orthography, vocabulary, word formation, sentence formation. Note also that, like Hymes, Canale and Swain include a sociolinguistic component (with two subcomponents: expressing and understanding appropriate social meanings and grammatical forms in different contexts) and, like Halliday, they include a discourse component (with cohesion and coherence subcomponents). However, in addition, they include strategic components (that is, the abilities necessary to overcome grammatical, sociolinguistic, discourse, and performance difficulties).

Table 2.3 The components of communicative competence

A. Grammatical competence
  1. Phonology
  2. Orthography
  3. Vocabulary
  4. Word formation
  5. Sentence formation
B. Sociolinguistic competence: Expressing and understanding *appropriate:*
  1. Social meanings
  2. Grammatical forms in different sociolinguistic contexts
C. Discourse competence
  1. Cohesion in different genres
  2. Coherence in different genres
D. Strategic competence for
  1. Grammatical difficulties
  2. Sociolinguistic difficulties
  3. Discourse difficulties
  4. Performance factors

Because of the need to address both the characteristics of communicative testing (listed in Table 2.2) and the components of communicative competence as just discussed (and summarized in Table 2.3), a natural part of this communicative movement has been the development of performance assessment and task-based assessment, which, in my view, are both ways of designing communicative tests, or assessment procedures. **Performance assessment**, according to Norris, Brown, Hudson, and Yoshioka (1998, p. 8), is distinguished from other types of testing in that: "(a) examinees must perform tasks, (b) the tasks should be as authentic as possible, and (c) success or failure in the outcome of the tasks, because they are performances, must usually be rated by qualified judges." They then point out that, "These three characteristics might just as well serve as a working definition...that will help us to distinguish already existing performance assessments, such as essays, interviews, extensive reading tasks, and so forth from integrative tests like dictations and cloze tests which do not fully meet any of the three criteria."

One type of performance assessment, **task-based assessment**, is defined by Brown, Hudson, Norris, and Bonk (2002, p. 9) as follows:

> In task-based language assessment, then, we are interested in eliciting and evaluating students' abilities to accomplish particular tasks or task types in which target language communication is essential. Such assessment is obviously performance assessment because a student's second language performance on the task is that which gets evaluated.

Why knowing about these movements is important

The methodology issue, initially described in terms of language teaching practices ranging from structuralist to communicative, has serious implications in thinking about historical movements within language testing, as well as important ramifications for the decisions that teachers make about which types of tests to use in their language programs.

To begin with, it is important to recognize that different theoretical views on linguistics and language teaching may exist in any program. These views might vary from teachers who still believe in a structural approach to others who passionately argue for communicative language teaching—with the bulk of the teachers falling somewhere in between. The degree to which different teachers believe in various language teaching theories (even if they do not know what they are called) can strongly influence the teaching in a program, and also the choices made in testing. Thus, a program will have to come to grips with such differences before any serious efforts can be made to implement tests of one type or another.

As a result, the content of any given test and the types of test questions used will be determined by the language teaching view(s) that underpin the test. As a result, understanding these movements and their relationships to language teaching is important for understanding the very purpose of your test and the degree to which the test is meeting that purpose, that is, the validity of your test (see Chapter 10).

The Competence/Performance Issue

Much has been made in linguistics of the distinction originally proposed by Chomsky between competence and performance. Chomsky (1965, p. 4) differentiates between the two as follows: "*competence* (the speaker-hearer's knowledge of his language) and *performance* (the actual use of language in concrete situations)." This distinction has some interesting ramifications for language testing. If linguistic performance is viewed as imperfect and full of flaws (even in native speakers), such performances can only be taken to be the outward manifestations of the underlying, but unobservable, linguistic competence. And, if such a difference exists for native speakers of a language, the difference may be even more pronounced in non-native speakers.

This distinction can help teachers to realize that tests are at best fairly artificial observations of a student's performance, and performance is only an imperfect reflection of the underlying competence. Since both competence and performance are of interest to language teachers, teachers must be very careful in their interpretation of test results to remember that performance is only part of the picture—a part that is a second-hand observation of competence.

In testing circles, the underlying competence is more often described in terms of a psychological construct (see Chapter 10). An example of a **psychological construct** in the ESL field is the notion of overall English as a foreign language proficiency. Thus, a student's competence in EFL might more readily be discussed as overall EFL proficiency, which is a psychological construct. However, even a relatively successful attempt to test this construct, as with the TOEFL, only provides an estimate of the student's performance, which is only a reflection of the underlying construct, or competence. The important thing to remember, in my view, is that language testing can provide an estimate of a student's performance (sometimes from various angles as in listening, reading, and grammar subtests), but never provides a direct measure of the actual competence that underlies the performance.

The Discrete-Point/Integrative Issue

Another issue which concerns language testers has to do with the different types of tests, which can range from discrete-point tests to integrative tests. Various combinations of these two types are possible as well.

Discrete-point tests are those which measure the small bits and pieces of a language as in a multiple-choice test made up of questions constructed to measure students' knowledge of different structures. One question on such an ESL test might be written to measure whether the students know the distinction between *a* and *an* in English. A major assumption that underlies the use of test questions like this is that a collection of such discrete-point questions covering different structures (or other language learning points), if taken together as a single score, will produce a measure of some global aspect of language ability. In other words, a teacher who believes in discrete-point tests would argue that scores based on the administration of fifty narrowly defined discrete-point multiple-choice questions covering a variety of English grammatical structures will reveal something about the students' overall proficiency in grammar. Anyone holding the psychometric-structuralist view of language teaching and testing would probably be comfortable developing a test along these lines. A corollary to this general view would be that the individual skills (reading, writing, listening, and speaking) can be tested separately, and that different aspects of these skills (like pronunciation, grammar, vocabulary, culture, and so forth) can also be assessed as isolated phenomena.

As noted above, however, not all testers and teachers are so comfortable with the discrete-point view of testing. **Integrative tests** are those designed to use several skills at one time. Consider dictation as a test type. The student is usually asked to listen carefully and write down a short prose passage as it is read aloud three times (with or without pauses) by the teacher, or played on a tape. The skills involved are at least listening comprehension and writing, but different aspects of these two skills come into play as well. Sometimes handwriting is a factor; certainly distinguishing between phonemes is important as are grammar, vocabulary, and spelling knowledge. In short, dictation is testing many different things at the same time and does so in the context of extended text. Advocates of the integrative-sociolinguistic movement would argue that such a test is complex in a similar fashion to the ways actual language use is complex. They would also argue that the language tested in integrative procedures like dictation, cloze test, and writing samples is being tested in the more natural, or at least larger, context of extended text.

Along the continuum between the most discrete-point types of tests and the most integrative tests, other kinds of tests are in a sense both integrative and discrete-point in nature. Consider a typical reading test in which the student is asked to read a passage and then answer multiple-choice fact, vocabulary, and inference questions about the passage. Viewing this task as a combination of reading a passage and integrating that reading into answering questions at different conceptual levels (that is, fact, vocabulary, and inference) might lead a teacher to conclude that reading comprehension is an integrative test. Yet looking at the focused nature of the fact and vocabulary questions, a discrete-point label would come to mind. The point is that the sometimes useful distinction between discrete-point and integrative tests is not always clear.

PRACTICAL ISSUES

The **practical issues** that I will address have to do with physically putting tests into place in a program. Teachers may find themselves concerned with the degree to which tests are fair in terms of objectivity. Or they may have to decide whether to keep the tests cheap or fight for the resources necessary to do a quality job of testing. Teachers may also be concerned about the logistics of testing. For instance, they may be worried about the relative difficulty of constructing, administering, and scoring different types of tests. In discussing each of these practical issues, I will illustrate how each works and how it interrelates with the other practical issues.

THE FAIRNESS ISSUE

Fairness can be defined as the degree to which a test treats every student the same, or the degree to which it is impartial. Teachers would generally like to ensure that their personal feelings do not interfere with fair assessment of the students or bias the assignment of scores. The aim in maximizing objectivity is to give each student an equal chance to do well. Therefore, teachers and testers often do everything in their power to find test questions, administration procedures, scoring methods, and reporting policies that optimize the chances that each student will receive equal and fair treatment. This tendency to seek objectivity has led to the proliferation of "objective" tests, which is to say tests, usually multiple-choice, which minimize the possibility of varying treatment for different students. Since such tests can be and often are scored by machine, the process is maximally dispassionate and therefore viewed as objective.

However, many of the elements of any language course may not be testable in the most objective test types, such as multiple-choice, true-false, and matching. Whether teachers like it or not, one day they will have to recognize that they are not able to measure everything impartially and objectively. Consider what would happen if a group of adult education ESL teachers decide that they want to test their incoming students' communicative abilities. In thinking through such a placement test, they will eventually have to recognize that a multiple-choice format is not appropriate and that, instead, they need to set up situations, probably role plays, in which the students will use the spoken language in interactions with other students (or with native speakers if they can convince some to help out). Having set up the testing situations, they will then have to decide how the performance of each student will be scored and compared to the performances of all other students.

They might begin by designing some sort of scale, which includes descriptions of what they are looking for in the language use of their adult education students, that is, whether they want to score for grammar accuracy, fluency, clear pronunciation, ability to use specific functions, or any of the myriad other possible focuses. The teachers may then have to further analyze and describe each area that they decide to focus on in order to provide descriptive categories that will help them to assign so many points for excellent performance, fewer points for mediocre performance, and no points for poor performance. All this is possible and even admirable if their methodological perspective is communicative. The problem is not with the scale itself, but rather with the person, or rater, who will inevitably assign the scores on such a test. Can any person ever be completely objective when assigning such ratings? Of course not.

There are a number of test types that necessitate rater judgments like that just described. These tend to be toward the integrative end of the discrete-point to integrative continuum and include tests like oral interviews, translations, and compositions. Such tests ultimately require someone to use some scale to rate the written or spoken language that the students produce. Since the results must eventually be rated by some scorer, there is always a threat to objectivity when these types of tests are used. The question is not whether the test is objective, but rather the degree of subjectivity that the teachers are willing to accept. For example, the University of Hawaii ELI placement test mixes relatively objective subtests like multiple-choice reading, multiple-choice proofreading, and multiple-choice academic listening subtests with a fairly judgmental, and therefore relatively subjective, composition subtest. There are also cloze and dictation subtests which cannot be classed as entirely objective (because some judgments must be made) nor completely subjective (because the range of possibilities for those judgments is fairly restricted).

Thus, teachers may find that their thinking about this issue cannot be framed in absolutes, but rather must center on the trade-offs that are sometimes necessary in testing theoretically desirable elements of student production while trying to maintain a relatively high degree of objectivity.

THE COST ISSUES

In the best of all possible worlds, unlimited time and funds would be available for teaching and testing languages. Unfortunately, this is rarely true. Most teachers are underpaid and overworked and must constantly make decisions which are based on how expensive some aspect of teaching, or testing, may turn out to be. This issue affects all the other issues covered in this chapter so it cannot be ignored even if it seems self-evident. Lack of funds can cause the abandonment of otherwise well-thought-out theoretical and practical positions that teachers have taken (and cause them to do things that they would previously have found detestable).

Consider the example of the adult education ESL communicative test that I discussed above. The teachers may have decided, for sound and defensible theoretical reasons, that they want to include a communicative test in their placement battery. They have also agreed that they are willing to tolerate a certain amount of subjectivity in order to achieve their collective theoretical ends. They develop a scale and procedures for administering the test and take them proudly to the department head, who says that it is absolutely impossible to conduct these interviews because of the time (and therefore cost) involved in paying teachers to do the ratings.

Something happens to teachers when they become administrators. I know that this is true because I watched it happen to me. When I first became a language teacher, I staunchly detested multiple-choice tests because I could not see how they represented students' abilities to actually use language in real situations. After all, people rarely communicate in real life with four optional answers provided. However, when I became an administrator I found myself arguing for large-scale placement testing in machine scorable multiple-choice formats—a position based on the fact that such testing is relatively easy and cheap to administer and score. While testing each student individually may sometimes be desirable, teachers must recognize that it is very expensive in terms of both time and money. Nevertheless, if a group of teachers decides that interviews or role plays are worth doing, they must somehow find adequate funding to do such testing well.

Ease of Test Construction

Special considerations with regard to test construction can range from deciding how long the test should be to considering what types of questions to use. All things being equal, a long test of 100 questions is likely to be better in terms of the consistency and accuracy of what is being measured than a shorter one. This is logical given that a one-question multiple-choice test is not likely to be as accurate in assessing students' performance as a two-question test, or a ten-question test, or a fifty-question test. Which test should teachers have the most confidence in? The fifty-question test, right? The problem is that this characteristic of tests is in direct conflict with the fact that short tests are easier to write than long ones. One goal of many test development projects is to find the "happy medium," that is, the shortest test length that does a consistent and accurate job of testing the students.

Another test construction issue involves the degree to which different types of tests are easy or difficult to produce. Some test types, for instance a composition test, are relatively easy to construct. A teacher needs only to think of a good topic for the students to write on and make up some test directions that specify how long the students will have to write and perhaps the types of things that the teacher will be looking for in scoring the writing samples. Dictation tests are also easy to construct: just find an appropriate passage, provide paper, read the passage aloud (perhaps once straight through, a second time in phrases with pauses so that students can write, and a third time straight through for proofreading), and have the students write the passage down. Short-answer questions and translations are also relatively easy to construct. Constructing a cloze test is somewhat more difficult: one must find an appropriate passage and type it up replacing every n^{th} word with a numbered blank (for evidence that this process is not quite as easy as it seems, see Brown 2002).

Writing fill-in, matching, true-false and multiple-choice questions, as I will explain in the next chapter, is more difficult. Most language testers find that writing sound multiple-choice questions is the most difficult of these. Anyone who does not find that to be the case might want to look very carefully at their questions to see if they are indeed sound and effective. With these more restricted receptive types of test questions, questions must be carefully constructed so that the correct answers are truly correct and incorrect answers are really wrong. Any teacher who has ever tried this will verify that the process of writing such questions can quickly become time-consuming.

Ease of Test Administration

My experience also indicates that ease of administration is a very important issue because testing is a human activity, which is very prone to mix-ups and confusion. Perhaps this problem results from the fact that students are often nervous during a test and teachers are under pressure. The degree to which a test is easy to administer will depend on the amount of time it takes, on the number of subtests involved, on the amount of equipment and materials required to administer it, and on the amount of guidance that the students need during the test. A short 30-question, 15-minute, one-page cloze test with clear directions is relatively easy to administer. A one-hour lecture listening test based on a video tape that requires the students to write an essay will probably be relatively difficult to administer.

Ease of Test Scoring

Ease of scoring is an important issue because a test that is easy to score is cheaper and is less likely to result in scorers making simple tallying, counting, and copying mistakes that might affect the students' scores. Most teachers will agree that such scoring mistakes are undesirable because they are not fair to the students, but I am willing to wager that any teacher who has served as a scorer in a pressure-filled testing situation has made such scoring mistakes. In one composition scoring situation, I found that ten language teachers made numerous mistakes resulting in adding five two-digit subscores to find each student's total score. These mistakes affected about 20 percent of the compositions and no teacher (myself included) was immune. The best that teachers can hope to do is to minimize mistakes in scoring by making the processes as simple and clear as humanly possible and by double and triple checking those parts of the process that are error prone.

Ease of scoring seems to be inversely related to the ease of constructing a test type. In other words, the easiest types of tests to construct initially (composition, dictation, translation, and so forth) are usually the most difficult to score and least objective, while those test types which are more difficult to construct initially (multiple-choice, true-false, matching, and so forth) are usually the easiest to score and most objective.

Interactions of Theoretical Issues

While it may seem redundant, I must stress the importance of recognizing that each of the theoretical issues discussed above can and will interact with all the others—sometimes in predictable patterns and at other times in unpredictable ways. For instance, if a group of high school language teachers wants to develop a test that, from a theoretical point of view, is communicative yet integrative and measures productive skills, they may have to accept that the test will be relatively subjective, expensive, and hard to administer and score. Thus, they must be willing to put in the effort to create a test that validly assesses the aspects of language learning they think are important.

If, on the other hand, they decide they want a test that is very objective, easy to administer, and easy to score, they may have to accept the fact that the questions must be relatively discrete-point (and therefore difficult to write) so that the answer sheets can be machine scorable. This decision will naturally result in a test that is not communicative and that focuses mostly on receptive skills. Hence, they may be sacrificing the validity of their test to practical considerations simply because they are not giving testing much priority in terms of resources and energy.

I am not arguing for one type of test or another. I am, however, arguing that all of these trade-offs are inevitably linked to the many testing issues discussed in this chapter as well as to the issues of test reliability and validity that I will discuss in Chapters 8 to 10.

ADOPT, ADAPT, OR DEVELOP?

In adopting, adapting, or developing language tests for a particular situation, teachers may be surprised at the diversity of opinion that exists, even within a specific institution, about what a good test should include. Some teachers may have naive views of what a test should be, while others hold very sophisticated, or idealistic, or impractical views. For instance, those teachers who studied languages in the audio-lingual tradition often think of a language test as a longer and more varied form of the transformation drill, while colleagues who have recently graduated from M.A. or Ph.D. programs may be talking about communicative, task-based procedures, which take two teachers 20 minutes to administer to each student.

The appropriate managerial strategies for developing tests must, of course, be tailored to each situation. But every management strategy falls somewhere along a continuum that ranges from authoritarian to democratic. Since most language teachers of my acquaintance do not take well to dictatorial administrative practices, I find that the best strategies to employ are those which involve the teachers in the process of adopting, adapting, or developing tests. An additional benefit, of course, is that they can usually be drawn into contributing more than just their ideas and opinions. Since testing sometimes involves long hours of work (often with no extra pay), any help colleagues can give will help.

A consensus must first be built about the purpose and type of test to employ. Then a strategy must be worked out that will maximize the quality and effectiveness of the test that will eventually be put into place. In the best of all possible worlds, each program would have a resident testing expert, whose entire job is to develop tests especially tailored for that program. But even in the worst of all possible worlds, rational decisions can be made in selecting commercially available tests if certain guidelines are followed. In many cases, any rational approach to testing will be a vast improvement over the existing conditions. Between these two extremes of developing tests from scratch or adopting them from commercial sources on pure faith is the notion of adapting existing tests and materials so that they better serve the purposes of the program.

The main point here is that many tests are, or should be, situation-specific. That is to say, a test can be very effective in one situation with one particular group of students and be virtually useless in another. In other words, teachers cannot simply go out and buy a test and automatically expect it to work with their students. Any particular commercial test may have been developed for an entirely different type of student and for entirely different purposes. The goal of this section of the chapter is to provide teachers with rational bases for adopting, adapting, or developing language tests so they will be maximally useful in their specific language programs.

ADOPTING LANGUAGE TESTS

The tests that are used in language programs are often adopted from sources outside of the program. This may mean that the tests are bought from commercial publishing houses, adopted from other language programs, or pulled straight from the current textbook. Given differences that exist among the participants in the various language programs around the world (for instance, differences in gender, number of languages previously studied, types of educational background, educational level, levels of proficiency, differences in native languages, and so forth), it is probable that many of the tests which have been acquired from external sources are being used with students quite different from those envisioned when the tests were originally developed and standardized. Using tests with the wrong types of students can result in mismatches between the tests and the abilities of the students as well as between the tests and the purposes of the program. For instance, I have seen situations where a proficiency test like the TOEFL is used for making placement decisions in a program with narrowly defined ability levels. Such practices are irresponsible and should be corrected whenever they are discovered, because the decisions are being based on test questions that are to a large extent too easy or too difficult for the students involved. Thus, the test items are quite unrelated to the needs of the particular students in the given language program or unrelated to the curriculum being taught in that program.

Selecting good tests to match the purposes of a particular language program is therefore very important. However, making these matches properly is often difficult because of the technical aspects of testing that many language teachers find intimidating. In searching for tests that are suitable for a program, teachers and administrators may therefore wish to begin by looking for help from testing experts by reading test reviews. Test reviews are useful in the same way that book reviews are. That is, they provide at least one other person's informed opinion about the test. However, a good reviewer may also explain key concepts for the reader and point to what features of a test are important to consider. Test reviews sometimes appear in the review sections of language teaching journals along with reviews of textbooks and professional volumes. Naturally, testing is not the focus of these journals, so test reviews tend to appear infrequently. *Language Testing* is a journal that specializes in articles on testing and, therefore, is more likely to provide test reviews. These particular reviews are sometimes fairly technical because the intended audience is testing specialists. For those teachers in ESL/EFL, Alderson, Krahnke, and Stansfield (1987), though somewhat dated now, is the only book I know of that provides a collection of practical and useful test reviews specifically designed for them. Most of the major tests available for ESL at that time are reviewed. One other source for language test reviews is available in any full-fledged research library: It is commonly referred to as *Buros Mental Measurements Yearbook*, a book of reviews of all kinds of published tests (including language tests) that comes out every two or three years (for full names, see Plake & Impara 2001; Plake, Impara, & Spies 2003).

Other approaches that teachers might want to use to improve their abilities to select quality tests for their programs would include: informing themselves about language testing through taking a course or reading up on it; hiring a new teacher, who also happens to have an interest in, or already knows about, the subject of testing; and giving one member of the faculty release time to become informed on the topic. In all cases, the checklist provided in Table 2.4 should (with some background in testing) help in selecting tests that match the purposes for which a particular language program needs them.

In using the checklist, teachers should look at the test manual provided by the publisher and begin by considering the general facts about the test. What is the title? Who wrote it? Where and when was it published? As shown in the table, the theoretical orientation of the test should probably be reviewed next. Is it in the correct family of tests (NRT or CRT) for the program's purposes? Is it designed for the type of decisions

Table 2.4 Test evaluation checklist

A. General background information
  1. Title
  2. Author(s)
  3. Publisher and date of publication
  4. Published reviews available

B. Your theoretical orientation
  1. Test family—Norm-referenced or criterion-referenced (see Chapter 1)
  2. Purpose of decision—placement, proficiency, achievement, diagnostic (see Chapter 1)
  3. Language methodology orientation—structural ←→ communicative
  4. Type of test—discrete-point ←→ integrative

C. Your practical orientation
  1. Objective ←→ subjective
  2. Expensive ←→ inexpensive
  3. Logistical issues—easy ←→ difficult
     a. Test construction
     b. Test administration
     c. Test scoring

D. Test characteristics
  1. Item description (see Chapter 3)
     a. Receptive mode (written text, picture, cassette tape, CD, and so on)
     b. Productive mode (marking choice, speaking, writing, and so on)
  2. Norms (see Chapter 6)
     a. Standardization sample (nature, size, method of selection, generalizability of results, availability of established norms for subgroups based on nationality, native language, gender, academic status, and so on)
     b. Number of subtests and separate scores
     c. Type of standardized scores (percentiles, and so on)
  3. Descriptive information (see Chapter 5)
     a. Central tendency (mean, mode, and median)
     b. Dispersion (low-high scores, range, and standard deviation)
     c. Item characteristics (facility and discrimination)
  4. Reliability/dependability (see Chapters 8 & 9)
     a. Types of reliability procedures used (test-retest, equivalent forms, internal consistency, interrater, intrarater, and so on)
     b. Degree of reliability for no. 4.a. above
     c. Standard error of measurement
  5. Validity (see Chapter 10)
     a. Types of validity procedures used (content, construct, and/or predictive/concurrent criterion-related validity)
     b. Degree to which you find convincing the validity statistics and argument(s) referred to above
  6. Actual practicality of the test
     a. Cost of test booklets, audio components, manual, answer sheets, scoring templates, scoring services, and any other necessary test components
     b. Quality of items listed in number 6.a. above (paper, printing, audio clarity, durability, and so on)
     c. Ease of administration (time required, proctor/student ratio, proctor qualifications, equipment necessary, availability and quality of directions for administration, and so on)
     d. Ease of scoring (method of scoring, amount of training necessary, time per test, score conversion information, and so on)
     e. Ease of interpretation (quality of guidelines for the interpretation of scores in terms of norms or other criteria)

involved? Does it match the methodological orientation of the teachers and the goals of the curriculum? What types of subtests are involved? Are they discrete-point or integrative, or some combination of the two?

From a practical point of view, a number of other issues must be considered. For instance, to what degree is the test objective? Will allowances have to be made for subjectivity? What about cost? Is the test too expensive for the program, or just about right? What about logistics? Is the test going to be easy to put together, administer, and score?

In terms of test characteristics, the nature of the test questions must be considered. What are the students confronted with in the receptive mode? And what are they expected to do in the productive mode? If the test is designed for norm-referenced decisions, is information about norms and standardized scores provided? Does the test seem to be aimed at the correct group of students and organized to test the skills that are taught in the program? How many parts and separate scores will there be, and are they all necessary? Do the types of test questions reflect the productive and receptive types of techniques and exercises that are used in the program? Is the test described clearly and does the description make sense? Is the test reliable and valid?

There are other practical considerations that are also important. What are the initial and ongoing costs of the test? How good is the quality of the audio program, test booklets, answer sheets, and so forth? Are there preview booklets or other sorts of preparatory materials available to give out to the students? Is the test easy to administer? Is the scoring reasonably easy relative to the types of test questions being used? Is the interpretation of scores explained with guidelines for reporting and clarifying the scores to the students and teachers involved?

In short, there are many factors that must be considered even in adopting an already published test for a particular program. Many of these issues can be addressed by any thoughtful language teacher, but others, such as examining the degree to which the test is reliable and valid, will take more knowledge and experience with language tests. (For a quick idea of the scope of what a teacher must know to decide about the relative reliability and validity of a test, take a brief glance through Chapters 8 to 10.) However, for commercial test products, it is the publisher's responsibility to convince potential test users that the test is worth adopting. The test users should, therefore, expect to find clearly explained arguments supporting the quality of the test. If such is not the case, then they should probably be suspicious of what the publisher is hiding and seriously ask themselves if they want to adopt such a poorly defended test.

Adapting Language Tests

A newly developed test may work fairly well in a program, but perhaps not as well as was originally hoped. Such a situation would call for further adapting of the test so that it better fits the needs and purposes of the particular language program. A number of strategies are described in the next chapter, which will help teachers to use qualitative and statistical analyses of test results to revise and improve tests. Generally, however, the process of adapting a test to a specific situation will involve some variant of the following steps:

1. Administer the test in the particular program, using the appropriate teachers and students;

2. Select those test questions that work well at spreading out the students (for NRTs), or are efficient at measuring the learning of the objectives (for CRTs) in this particular program;
3. Develop a shorter, more efficient revision of the test—one that fits the program's purposes and works well with its students (some new questions may be necessary, ones similar to those which worked well, in order to have a long enough test); and
4. Evaluate the quality of the newly revised test (see Table 2.4, p. 32).

With the basic knowledge provided in this book, any language teacher can accomplish all these steps. In fact, following the guidelines given in Chapter 4 will enable any teacher to adapt a test to a specific set of program goals and decision-making purposes. However, in the interest of fair advertising, I must provide the warning that test development is hard work and can be time-consuming. Nevertheless, in the end, the hard work is worthwhile because of the useful information that is gained and the satisfaction that is derived from making responsible decisions about students' lives. The point is that, before teachers begin a test revision project, they should insure that they will have enough time and help to do the job well.

DEVELOPING LANGUAGE TESTS

In an ideal situation, teachers will have enough resources and expertise available in their program so that proficiency, placement, achievement, and diagnostic tests can be developed and fitted to the goals of the program and to the ability levels and needs of the students enrolled there. The guidelines offered in this book should help with that process.

If a group of teachers decides to develop their own tests, they will need to begin by deciding which tests to develop first. Perhaps those tests which were identified as most program-specific in the previous chapter should have priority. That would mean developing tests of achievement and diagnosis first because they will tend to be based entirely and exclusively on the objectives of the particular program. In the interim, while developing these achievement and diagnostic tests, previously published proficiency and placement tests could be adopted as needed. Later, these teachers may wish to develop their own placement test so that the test questions being used to separate students into levels of study are related to the objectives of the courses and to what the students are learning in the program. However, because of their inter-programmatic nature, proficiency tests may necessarily always be adopted from outside sources so that comparisons between and among various institutions will make sense.

Somewhere in the process of developing tests, teachers may want to stop and evaluate them on the basis of the checklist provided in Table 2.4 (p. 32). Teachers should always be willing to be just as critical of their own tests as they are of commercial tests. The fact that a test is developed by and for a specific program does not necessarily make it a good test. So evaluation of test quality should be an integral part of the test development process.

PUTTING SOUND TESTS IN PLACE

Having decided to adopt, adapt, or develop tests, teachers are in a position to actually put them into place to help with decision making. The checklist shown in Table 2.5 should help successfully put tests into place. To begin with, make sure that the purposes for administering the various tests are clear to the curriculum developers and to the teachers (and eventually to the students). This presupposes that these purposes are already clearly defined in both theoretical and practical terms that are understood and agreed to by a majority of the staff.

The next step is to insure that all the necessary physical conditions for the test have been met. This might entail making sure that there is a well-ventilated and quiet place to give the test with enough time in that space for some flexibility and clear scheduling. Also, make sure that the students have been properly notified and have

Table 2.5 A testing program checklist

A. Establishing purposes of test
   1. Clearly defined (from both theoretical and practical orientations)
   2. Understood and agreed upon by staff

B. Evaluating the test itself (see Table 2.4)

C. Arranging the physical needs
   1. Adequate, well-ventilated, and quiet space
   2. Enough time in that space for some flexibility
   3. Clear scheduling

D. Making pre-administration arrangements
   1. Students properly notified of test
   2. Students signed up for test
   3. Students given precise information (where and when test will be, what they should do to prepare, and what they should bring with them, especially identification if required)

E. Administering the test
   1. Adequate materials in hand (test booklets, answer sheets, audio components, pencils, scoring templates, and so on) plus extras
   2. All necessary equipment in hand and tested (cassette/CD players, microphones, public address system, videotape/DVD players, blackboard, chalk, and so on) with backups where appropriate
   3. Proctors trained in their duties
   4. All necessary information distributed to proctors (test directions, answers to obvious questions, schedule of who is to be where and when, and so on)

F. Scoring
   1. Adequate space for all scoring to take place
   2. Clear scheduling of scoring and notification of results
   3. Sufficient qualified staff for all scoring activities
   4. Staff adequately trained in all scoring procedures

G. Interpreting
   1. Clearly defined purpose for results
   2. Provision for helping teachers use scores and explain them to students
   3. A well-defined place for the results in the overall curriculum

H. Record keeping
   1. All necessary resources for keeping track of scores
   2. Ready access to the records for administrators and staff
   3. Provision for eventual systematic termination of records

I. Test analyses
   1. Item analyses for test revision and improvement for future uses
   2. For reliability and validity
   3. Report the results to interested parties

J. Ongoing research
   1. Test results used to full advantage for research
   2. Test results incorporated into overall program evaluation plan

signed up in advance for the test. Perhaps students should be given precise written information that answers their most pressing questions. Where and when will the test be administered? What should they do to prepare for the test? What should they bring with them? Should they bring picture identification? This type of information prepared in advance in the form of a handout or pamphlet may save answering the same questions hundreds of times.

Before actually administering the test, check that there are adequate materials on hand, perhaps with a few extras of everything. All necessary equipment should be ready and checked to see that it works (with backups if that is appropriate). Proctors must be trained in their duties and have sufficient information to do a professional job of test administration.

After the test has been administered, provision must be made for scoring. Again, adequate space and scheduling are important so that qualified staff can be properly trained and carry out the scoring of the test(s). Equally important is the interpretation of results. The purpose of the results must be clear, and provision must be made for helping teachers use the scores and explain the scores to the students. Ideally, there will be a well-defined purpose for the results of the test in the overall curriculum planning.

Record keeping is often forgotten in the process of test giving. Nevertheless, all necessary resources must be marshaled for keeping track of scores including sufficient clerical staff, computers and software, or just some type of ledger book. In all cases, staff members should have ready access to the records. Provision must also be made for the eventual destruction or long-term storage of these records.

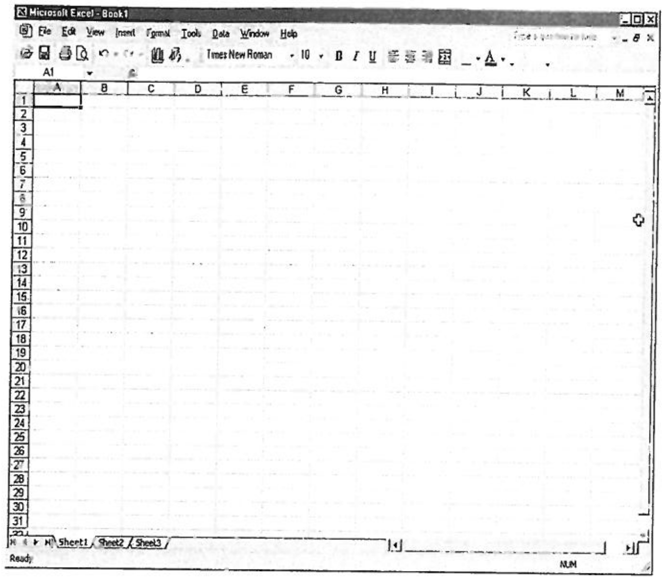
Test analysis is another essential part of test administration. Just as the unexamined life may not be worth living, the unanalyzed test may not be worth administering. As you will see in other chapters of this book, the pertinent analyses will most often include item analyses for purposes of revising and improving the test for future administrations (see Chapter 4), as well as analysis of the reliability and validity of the test (see Chapters 8 to 10). Naturally the results of these analyses should be reported to all interested parties.

Last but not least, an ongoing plan for research should be developed to utilize the information generated by test scores. Such research should take full advantage of the test results so that the new information can be effectively incorporated into the overall curriculum development process (see Chapter 11).

GETTING STARTED WITH YOUR SPREADSHEET PROGRAM

In this chapter, I will be asking you to get on a computer, open a spreadsheet program (preferably *Excel*™ because the directions I give here will be directly applicable), move around the spreadsheet, and enter sample test scores. You will benefit most from what follows if you do it while sitting at the computer. So now is the time to get on a computer and open up the *Excel* spreadsheet program.

Screen 2.1 Opening screen for *Excel*



On the opening screen, you will notice the following features when using the *Excel* spreadsheet:

Cells. Your spreadsheet is made up of **cells**, which are squares made by the intersections of the rows and columns in your spreadsheet. Cells are used to store data, such as numbers, names, or dates.

Rows and Columns. *Excel* stores and calculates data using a row and column format. Rows are labeled with numbers to the left, and columns are labeled with capital letters at the top (A through Z, then, AA, AB, AC, etc.). In a typical *Excel* spreadsheet, there are a total of 65,536 rows and 230 columns. Use your mouse or arrow keys to explore the rows and columns in the spreadsheet.

Cell Addresses. Each cell has an **address**, which is made up of column letter(s) and row numbers. Each cell has its own distinct address that is different from all the other cells' addresses. The cell in the upper left corner of the spreadsheet is labeled A1, and the address of the next cell is B1. The cell at the far right of the spreadsheet is labeled IV1. If you move down the spreadsheet 10 rows, the address is IV10, and if you move to the furthest column to the left, the address is A10.

Moving around the spreadsheet

To move around the spreadsheet, hold down the keys, described below, in quick succession.

END and **RIGHT ARROW** (→) keys will move you to the last column in the spreadsheet. The last column is labeled IV.

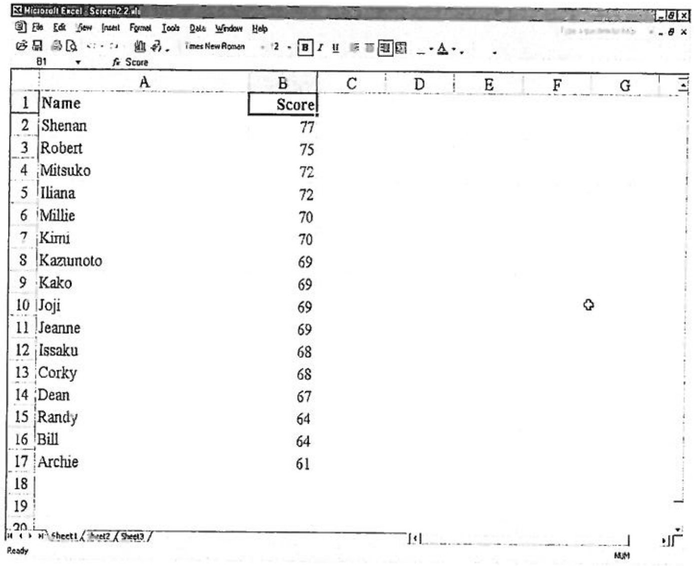
END and **DOWN ARROW** (↓) keys will move you to the last row in the spreadsheet. The last row is labeled 65536, which means that there are 65,536 total rows in the spreadsheet.

CTRL (or **CONTROL**) and **HOME** keys will move you to the upper-left hand corner of the spreadsheet, where you originally started when the spreadsheet was opened.

Creating a sample spreadsheet

In the following exercise, you will enter student names and test scores to create a sample spreadsheet. In the steps listed below, items that are in bold type are entered into the cells (i.e. **Name**). You may use the keyboard shortcuts by pressing the ALT key, followed by the underlined letters in the menu choices (i.e., ALT *f* to access the File item in the *Excel* menu, as shown by **File**). Items that are located in a specific menu will appear with a comma between each menu item (i.e., **File, Exit** to exit the *Excel* program).

Screen 2.2 Spreadsheet to track student scores


Microsoft Excel - Screen2.2.xls
File Edit View Insert Format Tools Data Window Help
Times New Roman
12
B1
Score
Sheet1 Sheet2 Sheet3
Ready
NUM

	A	B	C	D	E	F	G
1	Name	Score					
2	Shenan	77					
3	Robert	75					
4	Mitsuko	72					
5	Iliana	72					
6	Millie	70					
7	Kimi	70					
8	Kazunoto	69					
9	Kako	69					
10	Joji	69					
11	Jeanne	69					
12	Issaku	68					
13	Corky	68					
14	Dean	67					
15	Randy	64					
16	Bill	64					
17	Archie	61					
18							
19							
20							

Entering test score data to create a spreadsheet

1. Open the *Excel* program on your computer.
2. Click Cell A1, type **Name**, and then press ENTER.
3. In Cells A2 through A17, type the names of the students, as shown in Screen 2.2.
4. Click Cell B1, type **Score**, and then press ENTER.
5. In Cells B2 through B17, type the student scores in the cells, as shown in Screen 2.2.
6. To align the heading **Score** with the numbers, click Cell B1, and then click the ALIGN RIGHT button on the toolbar, located at the top of your screen. When the data are aligned, the button will have a pushed-down appearance. *Excel* aligns alphabetical data to the left side of the cells by default, and numerical data to the right.
7. To save the spreadsheet, click **File** from the menu bar, and then select **Save**. In the ***File name*** box, type an appropriate name for the spreadsheet, and then verify that the file will be saved in the correct location by checking the directory name listed in the ***Save in*** box.
8. Click **Save**, and then click **File**, **Exit** to close the *Excel* program.

You are now able to create spreadsheets to track test scores for students. In the next chapter, you will learn how to use a spreadsheet for analyzing the quality of the questions you use in your tests. But first, a disclaimer: this book is not designed to teach you all the details of using a spreadsheet, but to enhance your understanding of how a spreadsheet can help you perform better language testing. I encourage you to use the manual for your spreadsheet, and to get a good book that explains the ins and outs of your spreadsheet to answer any questions that may arise. Explore the menus and buttons on your spreadsheet to find out their functions, and use the help screens when you run into trouble. Try using your spreadsheet to do different things in your everyday teaching life, like entering and keeping track of your students' attendance and grades, or keeping track of your checks. A spreadsheet is a very useful tool, but it is important for you to establish a playful relationship when using the program. If you fight your spreadsheet and fear it, it will sense your fear and take control. So try playing with it in various ways. I've never known a student to break his or her spreadsheet, and at worst, you might have to reboot, so why not just try some things.

REVIEW QUESTIONS

1. What are the theoretical and practical issues that must be considered in developing language tests? How are the theoretical issues different in general from those classified as practical?
2. On a continuum of methodological choices that ranges from structural language teaching to communicative, where would your philosophy of teaching fit? What about your philosophy of testing? Are you prescientific? Are you a psychometric-structuralist? An integrative-sociolinguist? Or are you part of the communicative wave of the future?
3. How are performance testing and task-based testing related? Different? How are they communicative in nature?
4. What is the difference between competence and performance as discussed by Chomsky? And why might this distinction be important to think about with regard to language testing?
5. What is the fundamental difference between a discrete-point test and an integrative one? Can you think of at least one example of each? Would you prefer to use discrete-point or integrative tests for purposes of placing students into the levels of a language program? Why?
6. What are the five characteristics of a communicative test? What are the three bases for rating communicative language performance? And, what are the four main components of communicative competence (according to Canale 1983b)?
7. Why is objectivity important to language testers? Under what conditions could you justify sacrificing some degree of objectivity? And why?
8. What are some of the logistical conditions that you should consider in any testing project? Which of the three logistical conditions discussed in this book (ease of construction, administration, and scoring) do you think is the most important? How are ease of test construction and ease of scoring inversely related?
9. What are the factors that you must consider in looking at the quality of a test? Which do you think are the most important?
10. What are the factors that you must keep in mind in putting together a successful testing program? Which factors do you think are the most important?
11. What is an address in a spreadsheet? What is a cell? How many columns does your spreadsheet have? How many rows?

APPLICATION EXERCISES

A. Locate a test that you think might be useful in a language program in which you are now working, or if you have never taught, find a test for an elementary, secondary, adult education, commercial, or university language program. Examine the test very carefully using Table 2.5 (p. 35), keeping in mind all the theoretical and practical issues discussed in this chapter. Perhaps you should consult with several colleagues and find out what they think of it. What differences do you now have with your colleagues in your views on testing?

B. What theoretical and practical issues would be of particular importance for implementing the test that you selected for the above application exercise (see Table 2.5, p. 35)?

DEVELOPING GOOD QUALITY LANGUAGE TEST ITEMS

INTRODUCTION

In this chapter, I will begin explaining the elements that make up a good test. The basic unit of any test is the test item, so I will begin the chapter with a broad definition of this crucial concept. Then I will continue with guidelines for item format analysis including four separate sets: general guidelines for all types of test items; guidelines for receptive response items (true-false, multiple-choice, and matching); guidelines for productive response items (fill-in, short-response, and task); and personal response items (self-assessments, conferences, and portfolios). As usual, I will end the chapter with review questions and applications exercises.

WHAT IS A TEST ITEM?

The Multilingual Glossary of Language Testing Terms (ALTE 1998, p. 149) defines an *item* as follows: "Each testing point in a test which is given a separate mark or marks." That is fine as far as it goes, but what is a "testing point" and what is a "separate mark"? I think there is a clearer way to look at test items.

In the same sense that the phoneme is a basic unit in phonology and the morpheme is a basic unit in syntax, an *item* is the basic unit of language testing. Like the linguistic units above, the item is sometimes difficult to define. Some types of items, like multiple-choice or true-false items, are relatively easy to identify because they are the individual test questions that anyone can recognize as discrete units. An item may prove more difficult to identify for the more integrative types of language tests such as dictations, interviews, role plays, and compositions, or for more personal assessments like conferences, self-assessments, or portfolios. To accommodate the variety of discrete-point, integrative, and personal item types found in language testing, I will define the term **item** very broadly as the smallest unit that produces distinctive and meaningful information or feedback on a test when it is scored or rated. This definition will be general enough to work for every type of language test from multiple-choice to portfolio, yet will be specific enough to also prove useful.

Since the *item* is the basic unit, or building block, in testing, one way to improve a test is to examine the individual items and revise the test so that only those items that are performing well remain in the revised version of the test. Teachers often look at the total scores of their students on a test, but careful examination of the individual items that contributed to the total scores can also prove very illuminating. This process of carefully inspecting individual test items is called item analysis.

More formally, **item analysis** is the systematic evaluation of the effectiveness of the individual items on a test. This is usually done for purposes of selecting the "best" items which will remain on a revised and improved version of the test. Sometimes, however, item analysis is performed simply to investigate how well the items on a test are working with a particular group of students. Item analysis can take numerous forms, but when testing for norm-referenced purposes, there are three types of analyses that are typically applied: item format analysis, item facility analysis, and item discrimination analysis. In developing CRTs, three other concerns become paramount: item quality analysis, the item difference index, and the *B*-index for each item.

GUIDELINES FOR ITEM FORMAT ANALYSIS

In analyzing **item format**, testers focus on the degree to which each item is properly written so that it measures all and only the desired content. Such analyses often involve making judgments about the adequacy of item formats. Consider the following multiple-choice grammar item:

The apple is located somewhere on or around ______.

Ⓐ a table　　Ⓒ the table

Ⓑ an table　　Ⓓ table

This item has two possible answers (A and C), is wordier than it needs to be ("located somewhere...or around" may be difficult, distracting, and superfluous), and repeats the word "table" inefficiently. Item format analysis could lead us to correct these problems and produce a better item as follows:

Do you see the chair and table? The apple is on ______ table.

Ⓐ a　　Ⓒ the

Ⓑ an　　Ⓓ (no article)

Now, the first sentence makes "the" the only correct answer; the item has been reworded to avoid difficult, distracting, and superfluous words; and the word *table* is moved up into the main part of the item so it is not repeated four times in the A-D options. The item may still be imperfect because other teachers have not given feedback on it, but it is considerably better than it was when first written.

The guidelines provided in this chapter are designed to help teachers make well-informed and relatively objective judgments about how well items are formatted. The first set of guidelines is a very general set that teachers can apply to virtually all types of items. A second set will help guide teachers to analyze receptive response item formats (true-false, multiple-choice, and matching items). A third set will aid with the different types of productive response item formats (fill-in, short-response, and task), and a fourth set will aid teachers in formatting personal response item formats (conferences, portfolios, self-assessments). In all cases, the purpose is to help teachers improve the formatting of the items that they use in their language tests.

Table 3.1 General guidelines for most item formats

Checklist Questions	Yes	No
1. Is the item format correctly matched to the purpose and content of the item?	□	□
2. Is there only one correct answer?	□	□
3. Is the item written at the students' level of proficiency?	□	□
4. Have ambiguous terms and statements been avoided?	□	□
5. Have negatives and double negatives been avoided?	□	□
6. Does the item avoid giving clues that could be used in answering other items?	□	□
7. Are all parts of the item on the same page?	□	□
8. Is only relevant information presented?	□	□
9. Have race, gender, and nationality bias been avoided?	□	□
10. Has at least one other colleague looked over the items?	□	□

GENERAL GUIDELINES

Table 3.1 shows some general guidelines, which are applicable to most language testing formats. They are in the form of questions that teachers can ask themselves when writing or critiquing any type of item format. In most cases, the purpose of asking these questions is to insure that the students score high or low on the item type for the right reasons. In other words, the students should answer the items correctly only if they know the concept or skill being tested or have the skill involved. By extension, the students should answer incorrectly only if they do not know the material or lack the skill being tested. Let's consider each question in Table 3.1.

1. Is the item format correctly matched to the purpose and content of the item?

Teachers will, of course, want their item formats to match the purpose and content of the item. In part, this means matching the right type of item to what is being tested in terms of modes (productive or receptive) and channels (written or oral language). For instance, teachers may want to avoid using a multiple-choice format, which is basically receptive mode (students read and select, but produce nothing), for testing productive skills like writing and speaking. Similarly, it would make little sense to require the students to read aloud (productive) the individual letters of the words in a book in order to test the receptive skill of reading comprehension. Such a task would be senseless, in part because the students would be using both receptive and productive modes mixed with both oral and written channels when the purpose of the test, reading comprehension, is essentially receptive mode and written channel. A second problem would arise because the students would be too narrowly focused in terms of content on reading the letters of the words. To avoid mixing modes and channels and to focus the content at the comprehension level of the reading skill, teachers might more profitably have the students read a written passage and use receptive-response items in the form of multiple-choice comprehension questions. In short, teachers must think about what they are trying to test in terms of all the dimensions discussed in the previous chapter and try to match their purpose with the item format that most closely resembles it.