

## LANGUAGE ASSESSMENT I: BASIC CONCEPTS IN TEST DEVELOPMENT

So far, if you have been reading this book chapter by chapter from the beginning, you have gathered a great deal of information about the process of the classroom teaching of second language learners: principles underlying a sound approach, contextual considerations, lesson design and classroom management, and teaching language skills. In all these discussions, the notion of language assessment has emerged implicitly on a number of occasions, but not explicitly to the point of examining what the discipline knows about testing language ability and looking closely at various options available for periodic classroom-based assessment of students' developmental progress in a course. This and the following chapter will do just that.

This chapter focuses on basic concepts and constructs in language assessment. The focus will be on what we traditionally think of as a "test" rather than the broader notion of "assessment," and special attention will be given to large-scale standardized testing as opposed to classroom testing. Criteria for measuring a test, types of test, and a synthesis of issues in testing will be centered on **formal** measurements of language: those designated moments during which we administer a prepared instrument to students for the purpose of measuring their language competence. Chapter 22 will look at practical classroom contexts for assessment; these include some formal measurements as well as **informal** assessment. The latter includes moment-by-moment incidental and intended judgments of students' performance, techniques that are not traditionally thought of as assessment devices, and other procedures that have come to be called "alternative" assessment methods.

### WHAT IS A TEST?

A test, in plain words, is a method of measuring a person's ability or knowledge in a given domain. The definition captures the essential components of a test. A test is first a *method*. It is a set of techniques, procedures, and items that constitute an instrument of some sort that requires performance or activity on the part of the test-

taker (and sometimes on the part of the tester as well). The method may be intuitive and informal, as in the case of a holistic impression of someone's authenticity of pronunciation. Or it may be quite explicit and structured, as in a multiple-choice technique in which correct responses have already been specified by some "objective" means.

Next, a test has the purpose of *measuring*. Some measurements are rather broad and inexact, while others are quantified in mathematically precise terms. The difference between formal and informal assessment (which will be discussed in detail in Chapter 22) exists to a great degree in the nature of the quantification of data. The informal, everyday intuitive judging that we do as laypersons or teachers is difficult to quantify. Judgments are rendered in somewhat global terms. For example, it is common to speak of a "good" tennis player, a "fair" performance by an actor in a play, or a "poor" reader. Formal tests, in which carefully planned techniques of assessment are used, rely more on quantification, especially for comparison either within an individual (say, at the beginning and the end of a course) or across individuals.

A test measures a *person's* ability or knowledge. Care must be taken in any test to understand who the test-takers are. What is their previous experience and background? Is the test appropriate for them? How are scores to be interpreted for individuals?

Also being measured in a test is *ability* or competence. A test samples performance but infers certain competence. A driving test for a driver's license is a test requiring a sample of performance, but that performance is used by the tester to infer someone's general competence to drive a car. A language test samples language behavior and infers general ability in a language. A test of reading comprehension may consist of some questions following one or two paragraphs, a tiny sample of a second language learner's total reading behavior. From the results of that test the examiner infers a certain level of general reading ability.

Finally, a test measures a given *domain*. In the case of a proficiency test, even though the actual performance on the test involves only a sampling of skills, that domain is overall proficiency in a language—general competence in all skills of a language. Other tests may have more specific criteria. A test of pronunciation might well be a test only of a particular phonemic minimal pair in a language. One of the biggest obstacles to overcome in constructing adequate tests is to measure the desired **criterion** and not inadvertently include other factors.

How do you know if a test is a "good" test or not? Is it administrable within given constraints? Is it dependable? Does it accurately measure what you want it to measure? These questions can be answered through three classic criteria for "testing a test": practicality, reliability, and validity.

## PRACTICALITY

A good test is **practical**. It is within the means of financial limitations, time constraints, ease of administration, and scoring and interpretation. A test that is prohibitively expensive is impractical. A test of language proficiency that takes a student ten hours to complete is impractical. A test that requires individual one-to-one proctoring is impractical for a group of 500 people and only a handful of examiners. A test that takes a few minutes for a student to take and several hours for an examiner to evaluate is impractical for most classroom situations. A test that can be scored only by computer is impractical if the test takes place a thousand miles away from the nearest computer. The value and quality of a test are dependent upon such nitty-gritty, practical considerations.

The extent to which a test is practical sometimes hinges on whether a test is designed to be **norm-referenced** or **criterion-referenced**. In norm-referenced tests, each test-taker's score is interpreted in relation to a mean, median, standard deviation, and/or percentile rank. The purpose in such tests is to place test-takers along a mathematical continuum in rank order. Typical of norm-referenced tests are **standardized** tests intended to be administered to large audiences, with results quickly disseminated to test-takers. Such tests must have fixed, predetermined responses in a format that can be electronically scanned. Practicality is a primary issue.

Criterion-referenced tests, on the other hand, are designed to give test-takers feedback on specific course or lesson objectives, that is, the "criteria." Classroom tests involving smaller numbers, and connected to a curriculum, are typical of criterion-referenced testing. Here, more time and effort on the part of the teacher (test administrator) are usually required in order to deliver the feedback. One could say that criterion-referenced tests may, in the opinion of some, consider practicality as a secondary issue in the design of the test; teachers may sacrifice time and effort in order to offer students appropriate and useful feedback, or what John Oller (1979: 52) called "instructional value." Testing and teaching are interrelated, as we shall see in the next chapter in a discussion of the role of **washback** in classroom assessment.

## RELIABILITY

A **reliable** test is consistent and dependable. Sources of unreliability may lie in the test itself or in the scoring of the test, known respectively as test reliability and rater (or scorer) reliability. If you give the same test to the same subject or matched subjects on two different occasions, the test itself should yield similar results; it should have **test reliability**. I once witnessed the administration of a test of aural comprehension in which a tape recorder played items for comprehension, but because of street noise outside the testing room, students in the room who were sitting next

to windows were prevented from hearing the tape accurately. That was a clear case of unreliability. Sometimes a test yields unreliable results because of factors beyond the control of the test writer, such as illness, a "bad day," or no sleep the night before.

**Scorer reliability** is the consistency of scoring by two or more scorers. If very subjective techniques are employed in the scoring of a test, one would not expect to find high scorer reliability. A test of authenticity of pronunciation in which the scorer is to assign a number between one and five might be unreliable if the scoring directions are not clear. If scoring directions are clear and specific as to the exact details the judge should attend to, then such scoring can become reasonably consistent and dependable. In tests of writing skills, as was noted in Chapter 19, scorer reliability is not easy to achieve since writing proficiency involves numerous traits that are difficult to define. But as J.D. Brown (1991) pointed out, the careful specification of an analytical scoring instrument can increase scorer reliability.

## VALIDITY

By far the most complex criterion of a good test is **validity**, the degree to which the test actually measures what it is intended to measure. A valid test of reading ability is one that actually measures reading ability and not, say, 20/20 vision, previous knowledge of a subject, or some other variable of questionable relevance. To measure writing ability, one might conceivably ask students to write as many words as they can in fifteen minutes, then simply count the words for the final score. Such a test would be easy to administer (practical), and the scoring quite dependable (reliable). But it would hardly constitute a valid test of writing ability unless some consideration were given to the communication and organization of ideas, among other factors. Some have felt that standard language proficiency tests, with their context-reduced, CALP (cognitive academic language proficiency)-oriented language and limited stretches of discourse, are not valid measures of language "proficiency" since they do not appear to tap into the communicative competence of the learner. There is good reasoning behind such criticism; nevertheless, what such proficiency tests lack in validity, they gain in practicality and reliability. We will return to the question of large-scale proficiency later in this chapter.

How does one establish the validity of a test? Statistical correlation with other related measures is a standard method. But ultimately, validity can be established only by observation and theoretical justification. There is no final, absolute, and objective measure of validity. We have to ask questions that give us convincing evidence that a test accurately and sufficiently measures the test-taker for the particular objective, or **criterion**, of the test. If that evidence is there, then the test may be said to have **criterion validity**.

In tests of language, validity is supported most convincingly by subsequent personal observation by teachers and peers. The validity of a high score on the final



exam of a foreign language course will be substantiated by “actual” proficiency in the language. A classroom test designed to assess mastery of a point of grammar in communicative use will have validity if test scores correlate either with observed subsequent behavior or with other communicative measures of the grammar point in question.

How can teachers be somewhat assured that a test, whether it is a standardized test or one constructed for classroom use, is indeed valid? Three types of validation are important in your role as a classroom teacher: content validity, face validity, and construct validity.

## Content Validity

If a test actually samples the subject matter about which conclusions are to be drawn, if it requires the test-taker to perform the behavior that is being measured, it can claim **content validity**. You can usually determine content validity, observationally, if you can clearly define the achievement that you are measuring. A test of tennis competency that asks someone to run a 100-yard dash lacks content validity. If you are trying to assess a person’s ability to speak a second language in a conversational setting, a test that asks the learner to answer paper-and-pencil multiple-choice questions requiring grammatical judgments does not achieve content validity. A test that requires the learner actually to speak within some sort of authentic context does.

In most human situations, we are best tested in something when we are required to perform a sampling of the criterion behavior. But there are a few highly specialized and sophisticated testing instruments that do not have high content validity yet are nevertheless valid. Projective personality tests are a prime example. The Thematic Apperception Test and the Rorschach “inkblot” tests have little content validity, yet they have been shown to be accurate in assessing certain types of deviant personality behavior. A test of field independence as a prediction of language success in the classroom may have potentially good criterion validity but poor content validity in that the ability to detect an embedded geometric figure bears little direct resemblance to the ability to speak and hear a language. As already noted, standard proficiency tests often don’t get high scores on content validity.

## Face Validity

A concept that is very closely related to content validity is **face validity**, which asks the question “Does the test, on the ‘face’ of it, appear from the learner’s perspective to test what it is designed to test?” To achieve “peak” performance on a test, a learner needs to be convinced that the test is indeed testing what it claims to test. Once I administered a dictation test and a cloze test (see below for a discussion of cloze tests) as a placement test for an experimental group of learners of English as a second language. Some learners were upset because such tests, on the face of it, did not appear to them to test their true abilities in English. Face validity is almost

always perceived in terms of content: if the test samples the actual content of what the learner has achieved or expects to achieve, then face validity will be perceived.

## Construct Validity

A third category of validity that teachers must be aware of in considering language tests is **construct validity**. One way to look at construct validity is to ask the question "Does this test actually tap into the theoretical construct as it has been defined?" "Proficiency" is a construct. "Communicative competence" is a construct. "Self-esteem" is a construct. Virtually every issue in language learning and teaching involves theoretical constructs. Tests are, in a manner of speaking, operational definitions of such constructs in that they operationalize the entity that is being measured (see Davidson, Hudson, & Lynch 1985).

A teacher needs to be satisfied that a particular test is an adequate definition of a construct. Let's say you have been given a procedure for conducting an oral interview. The scoring analysis for the interview weighs several factors into a final score: pronunciation, fluency, grammatical accuracy, vocabulary use, and sociolinguistic appropriateness. The justification for these five factors lies in a theoretical construct that claims those factors as major components of oral proficiency. So, on the other hand, if you were asked to conduct an oral proficiency interview that accounted only for pronunciation and grammar, you could be justifiably suspicious about the construct validity of such a test.

Most of the tests that you will encounter as a classroom teacher can be validated adequately through content; if the test samples the outcome behavior, then validity will have been achieved. But when there is low, or questionable, content validity in a test, it becomes very important for a teacher to be assured of its construct validity. Standardized tests designed to be given to large numbers of students typically suffer from poor content validity but are redeemed through their construct validation. The TOEFL, for example, does not sample oral production, yet oral production is obviously an important part of succeeding academically in a university course of study. The TOEFL's absence of oral production content is justified by research that has shown positive correlations between oral production and the behaviors (listening, reading, grammaticality detection, and writing) actually sampled on the TOEFL. Because of the crucial need to offer a financially affordable proficiency test and the high cost of administering and scoring oral production tests, the omission of oral content from the TOEFL has been accepted as a necessity in the professional community.

Validity is a complex concept, yet it is indispensable to the teacher's understanding of what makes a "good" test. If in your language teaching you can attend to the practicality, reliability, and validity of tests of language, whether those tests are classroom tests related to a part of a lesson, or final exams, or proficiency tests, then you are well on the way to making accurate judgments about the competence of the learners with whom you are working.

## KINDS OF TESTS

There are many kinds of tests, each with a specific purpose, a particular criterion to be measured. Below you will find descriptions of five test types that are in common use in language curricula. Explanations here are only for the purpose of helping you to identify and differentiate among types, not to serve as a manual for designing such tests.

### 1. Proficiency tests

If your aim in a test is to tap global competence in a language, then you are, in conventional terminology, testing **proficiency**. A proficiency test is not intended to be limited to any one course, curriculum, or single skill in the language. Proficiency tests have traditionally consisted of standardized multiple-choice items on grammar, vocabulary, reading comprehension, aural comprehension, and sometimes a sample of writing. Such tests often have content validity weaknesses as already noted above, but after several decades of construct validation research, some great strides have been made toward constructing communicative proficiency tests.

A typical example of a standardized proficiency test is the Test of English as a Foreign Language (TOEFL) produced by the Educational Testing Service. It is used by nearly 1000 institutions of higher education in the US as an indicator of a prospective student's ability to undertake academic work in an English medium. The TOEFL consists of sections on listening comprehension, grammatical accuracy, written expression, reading, vocabulary, and on the recently introduced computer-based TOEFL, writing. With the exception of the writing section, the TOEFL and virtually all other large-scale proficiency tests are machine-scorable for rapid turnaround and cost effectiveness.

### 2. Diagnostic tests

A **diagnostic** test is designed to diagnose a particular aspect of a language. A diagnostic test in pronunciation might have the purpose of determining which phonological features of English are difficult for a learner and should therefore become a part of a curriculum. Usually, such tests offer a checklist of features for the administrator (often the teacher) to use in pinpointing difficulties. A writing diagnostic would first elicit a writing sample from students. Then, the teacher would identify, from a list of rhetorical features that are already present in a writing course, those on which a student needs to have special focus. It is not advisable to use a general achievement test (see below) as a diagnostic, since diagnostic tests need to be specifically tailored to offer information on student need that will be worked on imminently. Achievement tests are useful for analyzing the extent to which students have acquired language features that have already been taught.

### 3. Placement tests

Certain proficiency tests and diagnostic tests can act in the role of **placement tests**, whose purpose is to place a student into an appropriate level or section of a lan-



guage curriculum or school. A placement test typically includes a sampling of material to be covered in the curriculum (that is, it has content validity), and it thereby provides an indication of the point at which the student will find a level or class to be neither too easy nor too difficult, but appropriately challenging.

#### 4. Achievement tests

An **achievement test** is related directly to classroom lessons, units, or even a total curriculum. Achievement tests are limited to particular material covered in a curriculum within a particular time frame, and are offered after a course has covered the objectives in question. Achievement tests can serve as indicators of features that a student needs to work on in the future, but the primary role of an achievement test is to determine acquisition of course objectives at the end of a period of instruction.

#### 5. Aptitude tests

Finally, we need to consider the type of test that is given to a person prior to *any* exposure to the second language, a test that **predicts** a person's future success. A language **aptitude test** is designed to measure a person's capacity or general ability to learn a foreign language and to be successful in that undertaking. Aptitude tests are considered to be independent of a particular language. Two standardized aptitude tests have been used in the US—the *Modern Language Aptitude Test* (MLAT) (Carroll & Sapon 1958) and the *Pimsleur Language Aptitude Battery* (PLAB) (Pimsleur 1966). Both are English language tests and require students to perform such tasks as memorizing numbers and vocabulary, listening to foreign words, and detecting spelling clues and grammatical patterns.

Because of a number of psychometric issues, standardized aptitude tests are seldom used today. Instead, the measurement of language aptitude has taken the direction of providing learners with information about their preferred styles and their potential strengths and weaknesses. Any test that claims to *predict* success in learning a language is undoubtedly flawed, because we now know that with appropriate self-knowledge, active strategic involvement in learning, and/or strategies-based instruction, virtually everyone can succeed eventually. (A full discussion of language aptitude and aptitude tests can be found in *PLLT*, Chapter 4.)

Within each of the five categories of tests above, there are a variety of different possible techniques and procedures. These range from

- objective to subjective scoring procedures,
- open-ended to structured response options,
- multiple-choice to fill-in-the-blank item design formats,
- written to oral performance modes.

Tests of each of the modes of performance can be focused on a continuum of linguistic units, from smaller to larger: phonology and orthography, words, sentences, and discourse. In interpreting a test it is important to note which linguistic units are being tested. Oral production tests can be tests of overall conversational flu-



ency or pronunciation of a particular subset of phonology, and can take the form of imitation, structured responses, or free responses. Similarly, listening comprehension tests can concentrate on a particular feature of language or on overall listening for general meaning. Tests of reading can cover the range of language units and can aim to test comprehension of long or short passages, single sentences, or even phrases and words. Writing tests can take on an open-ended form with free composition, or be structured to elicit anything from correct spelling to discourse-level competence.

## **HISTORICAL DEVELOPMENTS IN LANGUAGE TESTING**

Historically, language testing trends and practices have followed the changing winds and shifting sands of methodology described earlier in this book (Chapter 2). For example, in the 1950s, an era of behaviorism and special attention to contrastive analysis, testing focused on specific language elements such as the phonological, grammatical, and lexical contrasts between two languages. In the 1970s and '80s, communicative theories of language brought on more of an integrative view of testing in which testing specialists claimed that "the whole of the communicative event was considerably greater than the sum of its linguistic elements" (Clark 1983: 432). Today, test designers are still challenged in their quest for more authentic, content-valid instruments that simulate real-world interaction while still meeting reliability and practicality criteria.

This historical perspective underscores two major approaches to language testing that still prevail, even if in mutated form, today: the choice between **discrete point** and **integrative** testing methods. Discrete-point tests were constructed on the assumption that language can be broken down into its component parts and those parts adequately tested. Those components are basically the skills of listening, speaking, reading, writing, the various hierarchical units of language (phonology/graphology, morphology, lexicon, syntax, discourse) within each skill, and subcategories within those units. So, for example, it was claimed that a typical proficiency test with its sets of multiple choice questions divided into grammar, vocabulary, reading, and the like, with some items attending to smaller units and others to larger units, can measure these discrete points of language and, by adequate sampling of these units, can achieve validity. Such a rationale is not unreasonable if one considers types of testing theory in which certain constructs are measured by breaking down their component parts.

The discrete-point approach met with some criticism as we emerged into an era of emphasizing communication, authenticity, and context. The earliest criticism (Oller 1979) argued that language competence is a unified set of interacting abilities that cannot be tested separately. The claim was, in short, that communicative competence is so global and requires such integration (hence the term "integrative" testing) that it cannot be captured in additive tests of grammar and reading and

Table 22.2 Traditional and alternative assessment (adapted from Armstrong 1994 and Bailey 1998: 207)

<u>Traditional Assessment</u>	<u>Alternative Assessment</u>
<u>One-shot, standardized exams</u>	<u>Continuous long-term assessment</u>
<u>Timed, multiple-choice format</u>	<u>Untimed, free-response format</u>
<u>Decontextualized test items</u>	<u>Contextualized communicative tasks</u>
<u>Scores suffice for feedback</u>	<u>Formative, interactive feedback</u>
<u>Norm-referenced scores</u>	<u>Criterion-referenced scores</u>
<u>Focus on the "right" answer</u>	<u>Open-ended, creative answers</u>
<u>Summative</u>	<u>Formative</u>
<u>Oriented to product</u>	<u>Oriented to process</u>
<u>Non-interactive performance</u>	<u>Interactive performance</u>
<u>Fosters extrinsic motivation</u>	<u>Fosters intrinsic motivation</u>

It should be noted here that traditional assessment offers significantly higher levels of practicality. Considerably more time and higher institutional budgets are required to administer and evaluate assessments that presuppose more subjective evaluation, more individualization, and more interaction in the process of offering feedback. The payoff for the latter, however, comes with more useful feedback to students, better possibilities for intrinsic motivation, and ultimately greater validity.

## PRINCIPLES FOR DESIGNING EFFECTIVE CLASSROOM TESTS

For many language learners, the mention of the word *test* evokes images of walking into a classroom after a sleepless night, of anxiously sitting hunched over a test page while a clock ticks ominously, and of a mind suddenly gone empty as they vainly attempt to "multiple guess" their way through the ordeal.

How can you, as a classroom teacher and designer of your own tests, correct this image? Consider the following four principles for converting what might be ordinary, traditional tests into authentic, intrinsically motivating learning opportunities designed for learners' best performance and for optimal feedback.

### 1. Strategies for test-takers

The first principle is to offer your learners appropriate, useful strategies for taking the test. With some preparation in test-taking strategies, learners can allay some of their fears and put their best foot forward during a test. Through strategies-based test-taking, they can avoid miscues due to the format of the test alone. They should also be able to demonstrate their competence through an optimal level of performance, or what Swain (1984) referred to as "bias for best." Consider the before-, during-, and after-test options (Table 22.3).

Table 22.3. Before-, during-, and after-test options.

**Before the Test**

1. Give students all the information you can about the test. Exactly what will the test cover? Which topics will be the most important? What kind of items will be included? How long will it be?
2. Encourage students to do a systematic review of material. For example: skim the textbook and other material, outline major points, write down examples, etc.
3. Give them practice tests or exercises, if available.
4. Facilitate formation of a study group, if possible.
5. Caution students to get a good night's rest before the test.
6. Remind students to get to the classroom early.

**During the Test**

1. As soon as the test is distributed, tell students to quickly look over the whole test in order to get a good grasp of its different parts.
2. Remind them to mentally figure out how much time they will need for each part.
3. Advise them to concentrate as carefully as possible.
4. Alert students a few minutes before the end of the class period so that they can proofread their answers, catch careless errors, and still finish on time.

**After the Test**

1. When you return the test, include feedback on specific things the student did well, what he or she did not do well, and if possible, the reasons for such a judgment on your part.
2. Advise the student to pay careful attention in class to whatever you say about the test results.
3. Encourage questions from students.
4. Advise students to make a plan to pay special attention in the future to points that they are weak on.

**2. Face validity**

Sometimes students don't know what is being tested when they tackle a test. Sometimes they feel, for a variety of possible reasons, that a test isn't testing what it is "supposed" to test. Face validity, as we saw in Chapter 21, means that in the students' perception, the test is valid. You can help to foster that perception with



- a carefully constructed, well-thought-out format,
- a test that is clearly doable within the allotted time limit,
- items that are clear and uncomplicated,
- directions that are crystal clear,
- tasks that are familiar and relate to their course work, and
- a difficulty level that is appropriate for your students.

### 3. Authenticity

Make sure that the language in your test is as natural and authentic as possible. Also, try to give language some context so that items aren't just a string of unrelated language samples. Thematic organization of items may help in this regard. Or consider a storyline that may run through your items.

Also, the tasks themselves need to be tasks in a form that students have practiced and feel comfortable with. A classroom test is not the time to introduce brand-new tasks because you won't know if student difficulty is a factor of the task itself or of the language you are testing.

### 4. Washback

**Washback**, mentioned in the previous chapter, is the benefit that tests offer to learning. When students take a test, they should be able, within a reasonably short period of time, to utilize the information about their competence that test feedback offers. Formal tests must therefore be learning devices through which students can receive a diagnosis of areas of strength and weakness. Their incorrect responses can become windows of insight about further work. Your prompt return of written tests with your feedback is therefore very important to intrinsic motivation.

One way to enhance washback is to provide a generous number of specific comments on test performance. Many teachers, in our overworked (and underpaid!) lives, are in the habit of returning tests to students with a letter grade or number score on them, and considering our job done. In reality, letter grades and a score showing the number right or wrong give absolutely no information of intrinsic interest to the student. Grades and scores reduce a mountain of linguistic and cognitive performance data to an absurd minimum. At best they give a relative indication of a formulaic judgment of performance as compared to others in the class—which fosters competitive, not cooperative, learning.

So, when you return a written test, or even a data sheet from an oral production test, consider giving more than a number or grade or phrase as your feedback. Even if your evaluation is not a neat little paragraph, at least you can respond to as many details in the test as time permits. Give praise for strengths—the “good stuff”—as well as constructive criticism of weaknesses. Give strategic hints on how a student might improve certain elements of performance. In other words, take some time to make the test performance an intrinsically motivating experience through which a student will feel a sense of accomplishment and challenge.



Finally, washback also implies that students have ready access to you to discuss the feedback and evaluation you have given. I'm sure you have known teachers with whom you wouldn't dare argue about a grade. Such a tyrannical atmosphere is out of place in an interactive, cooperative, intrinsically motivating classroom. For learning to continue, learners need to have a chance to feed back on your feedback, to seek clarification of any fuzzy issues, and to set new appropriate goals for themselves for the days and weeks ahead.

## **SOME PRACTICAL STEPS TO TEST CONSTRUCTION**

If you haven't already had an occasion to create and administer a classroom test, your time is coming soon! Now that you have read about testing issues in this chapter and considered the guidelines for more effective classroom tests, you may be thinking that you must now go out there and create a wonderfully innovative instrument that will garner the accolades of your colleagues and the admiration of your students. But don't worry! First, traditional testing techniques can, with a little tinkering, be altered to adhere to the spirit of an interactive, communicative language curriculum. Second, entirely new, innovative testing formats take a lot of effort to design and a long time to refine through the process of trial and error. Your best tack as a new teacher is to work within the guidelines of accepted, known, traditional testing techniques to give an intrinsically motivating, interactive flavor to your tests. Slowly, with experience, you can get bolder in your attempts.

In that spirit, here are some practical steps to take in constructing classroom tests.

### **1. Test toward clear, unambiguous objectives.**

You need to know as specifically as possible what it is you want to test. Sometimes teachers give tests simply because it's Friday or it's the third week of the course; after hasty glances at the chapter(s) covered during the period, they dash off some test items so the students will have something to do during the class period. This is no way to approach a test. Instead, carefully list everything that you think your students should "know" or be able to "do," based on the material the students are responsible for.

Your "objectives" can, for testing purposes, be as simple as the following list of grammatical structures and communicative skills in a unit that, let's say, you have recently taught:

#### **Grammar:**

Tag questions

Simple past tense in negative statements and information questions

Irregular past tense verbs

*Who* as subject  
*Anyone, someone, and no one*  
 Conjunctions *so* and *because*

**Communication skills:**

Guessing what happened  
 Finding out who did something  
 Talking about family and friends  
 Talking about famous people and events  
 Giving reasons  
 Asking for confirmation

**2. From your objectives, draw up test specifications.**

Now, this sounds like you're supposed to be some sort of psychometrician with a Ph.D. in statistics. Wrong. Test specifications for classroom use can be a simple and practical outline of your test.\* Let's say you are testing the above unit. Your specifications will indicate how you will divide up the 45-minute test period, what skills you will test, and what the items will look like. Your "specs" may look something like this:

**Listening** (15 minutes)

Part 1: Minimal sentence pairs (choose the sentence that you think you hear) [10 pairs, 2 themes]

Cover: tag questions  
 negative statements  
 guessing what happened  
 finding out who did something

Part 2: Conversation (choose the correct answer) [5 items]

Cover: information questions  
 talking about family and friends

**Multiple Choice** (10 minutes) [15 items in a storyline (cloze) format]

Cover : simple past tense  
 past irregular verbs  
*anyone, someone, and no one*

**Writing production** (15 minutes) [topic: Why I liked/didn't like a recent movie]

Cover : affirmative and negative statements  
 conjunctions *so* and *because*  
 giving reasons

---

\* Note that for standardized, large-scale tests that are intended to be widely distributed and therefore widely generalized, test specifications are much more formal and detailed.

These informal classroom-oriented specifications give you an indication of (a) which of the topics (objectives) you will cover, (b) what the item types will be, (c) how many items will be in each section, and (d) how much time is allocated for each. Notice that a couple of communication skills and one grammatical structure are not tested—this may be a decision based on the time you devoted to these objectives, or only on the finite number of minutes available to administer the test. Notice, too, that this course quite likely has a good deal of oral production in it, but for reasons of practicality (perhaps oral testing was done separately?), oral production is also not included on this test.

### 3. Draft your test.

A first draft will give you a good idea of what the test will look like, how students will perceive it (face validity), the extent to which authentic language and contexts are present, the length of the listening stimuli, how well a storyline comes across, how things like the cloze testing format will work, and other practicalities. Your items may look like these:

#### Listening, Part 1 (theme: last night's party)

1. Teacher says: We sure made a mess last night, didn't we?  
 Student reads: (a) We sure made no mess last night, did we?  
 (b) We sure made a mess last night, didn't we?

#### Listening, Part 2 (theme: still at the party)

2. Teacher says:\* A. Mary, who was that gorgeous man I saw you  
 with at the party?  
 B. Oh, Nancy, that was my brother!  
 Student reads: (a) Mary's brother is George.  
 (b) Nancy saw Mary's brother at the party.  
 (c) Nancy's brother is gorgeous.

#### Multiple choice (theme: still at the party)

- Student reads: Then we 3 the loudest thunder you have  
 ever heard! And of course right away lightning  
4 right outside the house!  
 3. (a) heared (b) did hear (c) heard  
 4. (a) struck (b) stricken (c) strack

---

\* Ideally, for the sake of authenticity, you should enlist the aid of a colleague and make a tape in which each of you reads a different part so that students will readily perceive that two people are speaking. If time, equipment, and colleagues don't permit this, make sure that when you read the two parts, you differentiate clearly (with voice and also by bodily facing in two different directions) between the two characters.

As you can see, these items are quite traditional. In fact, you could justifiably object to them on the grounds that they ask students to rely on short-term memory and on spelling conventions. But the thematic format of the sections, the authentic language, and the contextualization add face validity, interest, and intrinsic motivation to what might otherwise be a mundane test. And the essay section adds some creative production to help compensate for the lack of an oral production component.

#### **4. Revise your test.**

At this stage, you will work through all the items you have devised and ask a number of important questions:

1. Are the directions to each section absolutely clear?
2. Is there an example item for each section?
3. Does each item measure a specified objective?
4. Is each item stated in clear, simple language?
5. Does each multiple-choice item have appropriate distracters, that is, are the wrong items clearly wrong and yet sufficiently "alluring" that they aren't ridiculously easy?
6. Does the difficulty of each item seem to be appropriate for your students?
7. Do the sum of the items and test as a whole adequately reflect the learning objectives?

#### **5. Final-edit and type the test.**

In an ideal situation, you would try out all your tests on some students before actually administering them. In our daily classroom teaching, the tryout phase is virtually impossible, and so you must do what you can to bring to your students an instrument that is, to the best of your ability, practical, reliable, and valid. So, after careful completion of the drafting phase, a final edit is in order.

In your final editing of the test before typing it for presentation to your class, imagine that you are one of your students. Go through each set of directions and all items slowly and deliberately, timing yourself as you do so. Often we underestimate the time students will need to complete a test. If the test needs to be shortened or lengthened, make the necessary adjustments. Then make sure your test is neat and uncluttered on the page, reflecting all the care and precision you have put into its construction. If your test has a listening component, make sure your script is clear and that the audio equipment you will use is in working order.

#### **6. Utilize your feedback after administering the test.**

After you give the test, you will have some information about how easy or difficult it was, about the time limits, and about your students' affective reaction to it and their general performance. Take note of these forms of feedback and use them for making your next test.



## 7. Work for washback.

As you evaluate the test and return it to your students, your feedback should reflect the principles of washback discussed earlier. Use the information from the test performance as a springboard for review and/or for moving on to the next unit.

## ALTERNATIVE ASSESSMENT OPTIONS

So far in this chapter, the focus has been on the administration of formal tests in the classroom. It was noted earlier that "assessment" is a broad term covering any conscious effort on the part of a teacher or student to draw some conclusions on the basis of performance. Tests are a special subset of the range of possibilities within assessment; of course they constitute a very salient subset, but not all assessment consists of tests.

In recent years language teachers have stepped up efforts to develop non-test assessment options that are nevertheless carefully designed and that adhere to the criteria for adequate assessment. Sometimes such innovations are referred to as **alternative assessment**, if only to distinguish them from *traditional* formal tests. Several alternative assessment options will be briefly discussed here: self- and peer-assessments, journals, conferences, portfolios, and cooperative test construction.

### 1. Self- and peer-assessments

A conventional view of language pedagogy might consider self- and peer-assessment to be an absurd reversal of the teaching-learning process. After all, how could learners who are still in the process of acquisition, especially the early processes, be capable of rendering an accurate assessment of their own performance? But a closer look at the acquisition of any skill reveals the importance, if not the necessity, of self-assessment and the benefit of peer-assessment. What successful learner has not developed the ability to monitor his or her own performance and to use the data gathered for adjustments and corrections? Successful learners extend the learning process well beyond the classroom and the presence of a teacher or tutor, autonomously mastering the art of self-assessment. And where peers are available to render assessments, why not take advantage of such additional input?

Research has shown (Brown & Hudson 1998) a number of advantages of self- and peer-assessment: speed, direct involvement of students, the encouragement of autonomy, and increased motivation because of self-involvement in the process of learning. Of course, the disadvantage of subjectivity looms large, and must be considered whenever you propose to involve students in self- and peer-assessment.

Following are some ways in which self- and peer-assessment can be implemented in language classrooms.

- **Oral production:** student self-checklists; peer checklists; offering and receiving a holistic rating of an oral presentation; listening to tape-recorded oral production to detect pronunciation or grammar errors; in natural

~~that entices the high group but is "over the head" of the low group, and therefore the latter students don't even consider it.~~

~~The other two distractors (A and B) seem to be fulfilling their function of attracting some attention from lower-ability students.~~

## SCORING, GRADING, AND GIVING FEEDBACK

### Scoring

As you design a classroom test, you must consider how the test will be scored and graded. Your scoring plan reflects the relative weight that you place on each section and items in each section. The integrated-skills class that we have been using as an example focuses on listening and speaking skills with some attention to reading and writing. Three of your nine objectives target reading and writing skills. How do you assign scoring to the various components of this test?

Because oral production is a driving force in your overall objectives, you decide to place more weight on the speaking (oral interview) section than on the other three sections. Five minutes is actually a long time to spend in a one-on-one situation with a student, and some significant information can be extracted from such a session. You therefore designate 40 percent of the grade to the oral interview. You consider the listening and reading sections to be equally important, but each of them, especially in this multiple-choice format, is of less consequence than the oral interview. So you give each of them a 20 percent weight. That leaves 20 percent for the writing section, which seems about right to you given the time and focus on writing in this unit of the course.

Your next task is to assign scoring for each item. This may take a little numerical common sense, but it doesn't require a degree in math. To make matters simple, you decide to have a 100-point test in which

- the listening and reading items are each worth 2 points.
- the oral interview will yield four scores ranging from 5 to 1, reflecting fluency, prosodic features, accuracy of the target grammatical objectives, and discourse appropriateness. To weight these scores appropriately, you will double each individual score and then add them together for a possible total score of 40. (Chapters 4 and 7 will deal more extensively with scoring and assessing oral production performance.)
- the writing sample has two scores: one for grammar/mechanics (including the correct use of *so* and *because*) and one for overall effectiveness of the message, each ranging from 5 to 1. Again, to achieve the correct weight for writing, you will double each score and add them, so the possible total is 20 points. (Chapters 4 and 9 will deal in depth with scoring and assessing writing performance.)

Here are your decisions about scoring your test:

	Percent of Total Grade	Possible Total Correct
Oral Interview	40%	4 scores, 5 to 1 range $\times 2 = 40$
Listening	20%	10 items @ 2 points each = 20
Reading	20%	10 items @ 2 points each = 20
Writing	20%	2 scores, 5 to 1 range $\times 2 = 20$
Total		100

At this point you may wonder if the interview should carry less weight or the written essay more, but your intuition tells you that these weights are plausible representations of the relative emphases in this unit of the course.

After administering the test once, you may decide to shift some of these weights or to make other changes. You will then have valuable information about how easy or difficult the test was, about whether the time limit was reasonable, about your students' affective reaction to it, and about their general performance. Finally, you will have an intuitive judgment about whether this test correctly assessed your students. Take note of these impressions, however nonempirical they may be, and use them for revising the test in another term.

## Grading

Your first thought might be that assigning grades to student performance on this test would be easy: just give an "A" for 90–100 percent, a "B" for 80–89 percent, and so on. Not so fast! Grading is such a thorny issue that all of Chapter 11 is devoted to the topic. How you assign letter grades to this test is a product of

- the country, culture, and context of this English classroom,
- institutional expectations (most of them unwritten),
- explicit and implicit definitions of grades that you have set forth,
- the relationship you have established with this class, and
- student expectations that have been engendered in previous tests and quizzes in this class.

For the time being, then, we will set aside issues that deal with grading this test in particular, in favor of the comprehensive treatment of grading in Chapter 11.

## Giving Feedback

A section on scoring and grading would not be complete without some consideration of the forms in which you will offer feedback to your students, feedback that you want to become beneficial washback. In the example test that we have been referring to here—which is not unusual in the universe of possible formats for periodic



classroom tests—consider the multitude of options. You might choose to return the test to the student with one of, or a combination of, any of the possibilities below:

1. a letter grade
2. a total score
3. four subscores (speaking, listening, reading, writing)
4. for the listening and reading sections
  - a. an indication of correct/incorrect responses
  - b. marginal comments
5. for the oral interview
  - a. scores for each element being rated
  - b. a checklist of areas needing work
  - c. oral feedback after the interview
  - d. a post-interview conference to go over the results
6. on the essay
  - a. scores for each element being rated
  - b. a checklist of areas needing work
  - c. marginal and end-of-essay comments, suggestions
  - d. a post-test conference to go over work
  - e. a self-assessment
7. on all or selected parts of the test, peer checking of results
8. a whole-class discussion of results of the test
9. individual conferences with each student to review the whole test

Obviously, options 1 and 2 give virtually no feedback. They offer the student only a modest sense of where that student stands and a vague idea of overall performance, but the feedback they present does not become washback. Washback is achieved when students can, through the testing experience, identify their areas of success and challenge. When a test becomes a learning experience, it achieves washback.

Option 3 gives a student a chance to see the relative strength of each skill area and so becomes minimally useful. Options 4, 5, and 6 represent the kind of response a teacher can give (including stimulating a student self-assessment) that approaches maximum washback. Students are provided with individualized feedback that has good potential for “washing back” into their subsequent performance. Of course, time and the logistics of large classes may not permit 5d and 6d, which for many teachers may be going above and beyond expectations for a test like this. Likewise option 9 may be impractical. Options 6 and 7, however, are clearly viable possibilities that solve some of the practicality issues that are so important in teachers' busy schedules.

\$   \$   \$   \$   \$

~~In this chapter, guidelines and tools were provided to enable you to address the five questions posed at the outset: (1) how to determine the purpose or criterion of the test, (2) how to state objectives, (3) how to design specifications, (4) how to~~