Chapter 1

Understanding Language Tests and Testing Practices

All ESOL teachers come into contact with language assessment in one form or another, yet many find principles of assessment an aspect of professional knowledge that is difficult to update and apply effectively. This chapter, therefore, lays out the critical concepts required for an understanding of the language tests and testing practices outlined in subsequent chapters.

We begin with a brief discussion of the importance of assessment in the broader context of TESOL and education followed by an explanation of the problem that has developed because of the intellectual division between those concerned with language testing and those who teach. We argue that changes in educational practice, measurement theory, and language testing research necessitate bridging that division to meet current and future classroom needs. This chapter begins to build the bridge by deconstructing the dichotomy that narrows the knowledge and responsibility of the language teacher, and by replacing it with a more robust set of concepts for understanding a range of testing practices: test purpose, test method, and test justification. These overarching concepts form the basis for introducing key terms often used to describe language tests.

The Importance of Assessment

Teachers are involved in many forms of assessment and testing through their daily teaching and use of test scores. The significance of testing issues is evidenced in practitioner publications like TESOL Journal, which in the mid-1990s published one special issue and a number of articles that reported the
successful use of student-constructed assessments in ESOL classrooms (see, e.g., Gardner, 1996; Gottlieb, 1995; McNamara & Deane, 1995; Murphy, 1995; Smolen, Newman, Wathen, & Lee, 1995). For the most part, these assessments were implemented in contexts where teachers were making low-stakes decisions and had the freedom to select the forms of assessment. Even so, all classroom materials, whether they are for teaching, assessment, or both, need to be constructed and administered in a manner that ensures their appropriate use. Assessments that begin as onetime classroom innovations may draw the attention of colleagues, who may ask to borrow them or use the scores for other purposes.

Moreover, teachers are often involved in high-stakes, mandated assessment programs that prescribe forms of assessment, including standardized, traditional forms of assessment. As recently as academic year 1992–1993, 70% of the statewide educational testing programs in the United States were using multiple-choice assessments (Barton & Coley, 1994). The ever-increasing sense of mismatch between educational philosophy, instructional practices, and assessment techniques has some practitioners frustrated with many standardized tests (Herman, Aschbacher, & Winters, 1992) and interested in alternatives. Gottlieb (1995) echoes the sentiments of many ESOL teachers when she states that “rich, descriptive information about the processes and products of learning cannot be gathered by conventional teaching and testing methods” (p. 12). She maintains there has been a “rise of instructional and assessment practices that are holistic, student centered, integrated, and multidimensional” (p. 12).

The Division Between Teachers and Testers

Despite the effect of teacher-initiated practices on classroom assessments, a division continues to exist between language teachers and testers. Many teachers voice dissatisfaction with high-stakes, standardized tests but do not feel qualified to argue effectively against them and propose alternatives. In situations involving classroom assessments or programwide tests, teachers need to be empowered to address questions about the choice, development, and use of traditional tests and other forms of assessment.

Although teachers construct tests and test specialists may teach or have taught ESOL, the daily activities and roles of the two groups are generally different. The division of labor evident in language education is part of a broader phenomenon that began in the educational community in the first few decades of the 20th century and continued unabated for the next half-century. Teaching and testing became separated as the trend toward academic specialization accelerated and culminated in the emergence of a psychometric perspective that was dedicated to developing highly refined, standardized, objective measures (Stoynoff, 1996). Created by testing
specialists who worked in universities and research centers, and designed to measure human traits and abilities, these tests were scientifically developed and empirically tested; moreover, some of the more widely used tests have been continuously researched. During this period of language testing, which Spolsky (1978, 1995) referred to as psychometric-structuralist or modern, society in general—and teachers in particular—vested a great deal of power and authority in testing specialists.

In many respects, this authority was justly earned as the science of educational testing rapidly matured and advanced during the 20th century. Throughout the century, testing specialists extended research methods, improved their ability to develop and empirically evaluate tests (often by applying increasingly sophisticated statistical procedures and techniques to test development), and built more comprehensive theories to explain the abilities they sought to measure. This activity was supported by an academic culture that emphasized basic research, test development, empirical evaluation, and theory building. These activities, moreover, contributed to the refinement of important test constructs and produced sophisticated tests that accurately measured what they were designed to measure. But as the science of testing expanded, so did the gulf between what teachers knew and what testing specialists knew about testing. Educators who worked in schools were consumers of what was produced by those in academe, and the culture of schools emphasized test selection and administration, interpretation of results, and decision making based on test results. This division of labor permitted both cultures to focus on what they did best. Teachers taught, and test specialists developed standardized tests that schools used to evaluate students.

Writing about the use of standardized tests in U.S. schools, Stiggins (1997) observed that

The paradox is that, as a society (within and outside schools), we seem to have been operating on blind faith that these tests are sound, and that educators are using them appropriately. As a society, almost to a person, we actually know very little about standardized tests or the scores they produce. It has been so for decades. This blind faith has prevented us from understanding either the strengths or the important limitations of standardized tests. (p. 352)

Stiggins explained that the majority of standardized tests used to measure educational attainment, including language tests, are intended to offer general estimates of the learner’s ability or achievement in broad content domains or in certain kinds of reasoning. Administrators and teachers are to use the results to sort learners by ability or gauge their general achievement after a substantial amount of learning has occurred. As such, the results are unsuitable for assessing the daily progress of learners or their achievements at the end of a single course.

In other words, the division of labor has produced tests that are useful
for some purposes but not for others. But the more systemic result of the division is the partial and fragile knowledge that teachers have about how to collect information systematically for the purpose of making certain determinations about learners, which has led to the perception that testing and assessment are completely distinct educational processes. Ironically, this perception further dichotomizes the classroom assessment that teachers engage in and the testing that is the responsibility of researchers. Despite the value of a strong theory and practice of classroom-based assessment, maintaining the separation between testing and assessment keeps teachers from applying their knowledge of assessment to high-stakes testing.

**From Division to Unity**

Forging stronger links between teaching and assessment is essential if educators hope to optimize classroom learning. This section highlights some of the salient developments that occurred during the transition to the postmodern period and their effect on the relationship between teaching and testing.

**Educational Practices**

The educational landscape in the United States changed dramatically in the early 1980s. The U.S. educational reform movement was precipitated by the publication of *A Nation at Risk: The Imperative for Educational Reform* (National Commission on Excellence in Education, 1983). This 32-page report to the U.S. secretary of education documented a decline in the academic quality of U.S. educational institutions (public and private, from kindergarten through university), and it recommended five major reforms to correct the declines in achievement. Two recommendations with implications for testing and assessment were (a) to restore an academic core (called *new basics*) to the curriculum that should reflect a decidedly applied orientation and (b) to implement more rigorous and measurable standards for academic performance. Instructional practices and assessment procedures were modified to conform to the curriculum reforms that followed the release of the report (Linn, 1994). The reform movement gained momentum in the 1990s, when the federal government passed the Goals 2000: Educate America Act (1994), which created a structure and empowered a body to develop guidelines for national education standards and offer states exemplary standards and the assessments to use in achieving these new national standards. Barton and Coley (1994) underscored the shift in assessment as a result of these reforms.
The nation is entering an era of change in testing and assessment. Efforts at the national and state levels are now directed at greater use of performance assessment, constructed response questions, and portfolios based on actual student work. (as cited in Linn, 1994, p. 5)

With the enactment of Public Law 107-110, widely referred to as the No Child Left Behind Act of 2001 (2002), the federal government increased the pressure on states to adopt challenging academic achievement standards and improve the performance of low-achieving students, including those for whom English is a second language (L2). The legislation compels states to require that students pass state assessments a minimum of once between Grades 3–5, 6–9, and 10–12 in mathematics, reading, and language arts beginning in academic year 2005–2006 and in history and science by 2007–2008. Any nonnative speaker of English who has received 3 consecutive years of schooling in the United States will be expected to pass the same assessments without accommodations (pp. 1449–1451). However, Title III of the act stipulates that, beginning in academic year 2002–2003, schools must assess annually the English language proficiency of all limited-English-proficient students and determine the extent of their progress in acquiring English. Although schools are permitted to select or develop their assessments, measures must be approved by the state and need to assess the “child’s level of comprehension, speaking, listening, reading, and writing skills in English” (p. 1701). Moreover, the act indicates that assessments must be consistent with the standards established by the professional testing community and “enable itemized score analyses to be produced and reported” (pp. 1451–1452). Clearly, the educational reforms contained in this legislation will increase the assessment responsibilities of ESOL teachers and program administrators.

The shift toward standards-based education and the assessment of learner performance relative to a set of predetermined outcomes is not limited to the United States. In fact, there is evidence that the language teaching profession is in the midst of a global trend in establishing educational goals and holding teachers and programs accountable for monitoring and documenting learners’ achievements (Brindley, 1998). European reforms in L2 education have been especially ambitious and noteworthy over the past 30 years. The Council of Europe—a consortium of 40 nations founded in 1949—has been instrumental in promoting reforms in the teaching, learning, and evaluation of foreign language abilities across Europe’s languages by coordinating the activities of educational administrators, curriculum developers, teachers, and testing specialists (Davies et al., 1999; “Towards a European Framework,” 1994). Working with the Association of Language Testers in Europe, a professional network of European institutions that develop and administer language examinations, the Council of Europe (2001) created a comprehensive program that includes a common set of
objectives and procedures for monitoring learners' achievement and evaluating their ability in a foreign language. The European approach to reforming language learning and testing differs from the legislative approach taken in the United States. Although the Council of Europe encourages member states to adopt the new educational framework, states are not required to do so.

**Educational Measurement Theory**

Another impetus for change comes from developments in educational measurement theory. In the eyes of the public, government-initiated educational reform may have sparked changes in assessment in the 1990s, but, in fact, expressions of dissatisfaction with theory and practice had been stirring in the educational measurement community throughout the previous decade. Five important papers published in the 1980s and 1990s offer a glimpse into these developments. Fredrickson (1984) pointed out that the overuse of multiple-choice testing could have a negative effect on student learning because of the practice of teaching to the test. In view of this concern and other developments in measurement theory, Messick (1989) published his seminal paper proposing an expanded concept of validity. Validity, he suggested, should be seen as an argument concerning the extent to which test use can be justified for a particular purpose, and one aspect of the argument should include the effects of test use on instruction.

Linn, Baker, and Dunbar (1991) pointed out the importance of the criteria used in developing the validity argument and suggested criteria that would privilege assessments with complex constructed responses over multiple-choice tests. Moss (1992) took up the polemic issue of the relationship between reliability and validity, suggesting that the orthodox view of reliability as essential for validity precluded the acceptance of some forms of assessments, a point that is critical for language testing (Swain, 1993). Concerned with the scope of the theoretically sound validity argument versus the practical needs of test development and use, Shepard (1993) suggested that the use of a particular test needed to figure substantively in the development of its validity argument.

What is apparent from this summary is that educational measurement theory centers on how one defines validity. This makes sense if one considers that the types of evaluative questions test users ask about tests all concern validity (see Table 1.1). The change in how the educational measurement community views validity is reflected in the questions about tests implied by former and current conceptions of validity. In the past, validity was considered a characteristic of a test—the extent to which a test measures what it is supposed to measure—whereas today it is considered an argument concerning test interpretation and use—the extent to which test interpretations and uses can be justified. Reliability was seen as distinct from and a necessary condition for validity, but now reliability is more
typically seen as one type of validity evidence. In the past, validity was largely established through correlations of a test with other tests, but now validity is best argued on the basis of a number of types of rationales and evidence, including the consequences of testing (e.g., its effect on teaching). Construct validity was seen as one of three types of validity: content, criterion related, and construct. But today, validity is a unitary concept with construct validity as central; content and criterion-related evidence can be used as evidence about construct validity.

These changes are interesting and important for language test users particularly because of three themes that underlie them. First, the changes have resulted in a view of validity as a context-specific argument rather than a test characteristic that can be established in a universal way. As a consequence, a second theme is the view that justifying the validity of test use is the responsibility of all test users rather than a job solely within the purview of testing researchers who develop large-scale, high-stakes tests. A third theme is that one consideration of ESOL test users should be the effects of tests on the teaching and learning of English.

**Language Testing Research and Practice**

These three themes may have arisen from the U.S. educational measurement scene, but their influence has been substantial in the international community of language testing. This influence is embodied in a common journal (Language Testing), an electronic discussion list, an international organization (the International Language Testing Association), and an annual conference.
ESOL Tests and Testing: A Resource for Teachers and Administrators

(the Language Testing Research Colloquium) as well as in international EFL testing programs such as the Test of English as a Foreign Language (TOEFL) and the International English Language Testing System (IELTS).

As for the first theme of a situation-specific validity argument, recent work in language testing questions whether building generally accepted and valid models of language ability is practical given that language use and testing occur in such varied contexts (Chalhoub-Deville, 1997). Hence, some analysts have suggested that a more useful endeavor would be to develop what Chalhoub-Deville describes as operational models appropriate for particular test situations. This conclusion follows logically from the theory of communicative language ability as it has evolved over the past 25 years: Canale and Swain (1980) viewed communicative competence as including the strategies that would come into play during language use; Bachman’s (1990) and Bachman and Palmer’s (1996) concept of communicative language ability includes the context of language use in the overall schematic of their discussion of communicative language ability; Chapelle’s (1998) description of an interactionalist construct definition goes one step further, stating that part of construct definition is context definition; that is, a definition of language ability needs to include the range of contexts of language use. What follows from the situation-specific construct definition—operational or theoretical—is the purpose-specific nature of tests, meaning that the validity of test use clearly rests on the situation of use. The tension between situation-specific construct definition and validation, on the one hand, and the need for general theories and principles, on the other, is defining one focus of language testing inquiry in the postmodern period.

The second theme—that testing not be left solely to the language testing researcher—has been taken up to some degree by the alternative assessment movement, but another less apparent manifestation is the expansion of language testing research to include more than model-fitting studies concerned with issues associated with reliability. A variety of research methodologies have led to important insights into the ways in which test takers’ performance varies across test characteristics such as the type of tasks and content of the test (e.g., topics, instructions, genre, text types) and that this variability reflects the variability that exists in L2 performance. Such research requires expertise that extends beyond statistical matters. The ideal expressed in Messick’s (1989) reaching discussion of validity inquiry is being realized in language testing research that has explored test score meaning from theoretical perspectives and through qualitative and quantitative methods.

One of these relatively new approaches, the study of testing consequences, addresses the third theme (Alderson & Wall, 1996; Bailey, 1999). Concern for the effects of testing on learning is one aspect of the larger issue of ethical considerations in language testing, which is of growing importance to language testing specialists (Davies, 1997). The ethics of language testing refers to the responsibility of those who develop and
choose tests to see that they are used fairly. This discussion builds on the work of Canale (1987), who in the 1980s was an advocate for appreciating that test specialists and practitioners alike have a responsibility to “ensure that language tests are valuable experiences and yield positive consequences for all involved” (Douglas & Chapelle, 1993, p. 3). One of the ongoing issues of the postmodern period is to gain a greater understanding of how test fairness should come into play in the testing process (Kunnan, 1997).

Clearly, recent developments in educational practices, measurement theory, and language testing research offer compelling reasons for ESOL professionals to be assessment literate, which means being able to choose and use assessments for all of their purposes (Stiggins, 1997). At one time, the roles of language teachers and testing specialists were highly differentiated, leading many ESOL teachers and program administrators to become increasingly disconnected from the technical developments and practices associated with language tests and different types of assessments. However, the postmodern period is placing more responsibility for selecting, developing, and justifying assessments in the hands of practitioners, many of whom lack sufficient assessment literacy and confidence to fulfill these responsibilities.

### Understanding Assessment and Testing

How can language teachers and program administrators take more responsibility for choosing, developing, using, and interpreting all forms of assessments and tests? As mentioned above, one response has been to draw a distinction between traditional testing and alternative assessment, the former being the domain of researchers and the latter the responsibility of teachers. For example, Herman et al. (1992) distinguish between traditional, multiple-choice testing and alternative forms of assessment, which include interviews, essays with prompts and scoring criteria, documented observations, self-evaluation, and portfolios. Other testing specialists refer to the distinction between objectively scored, paper-and-pencil tests and alternative assessments, which include compositions, performance assessments such as demonstrations or portfolios, and communicative exchanges such as interviews or conferences (Stiggins, 1997). Brown and Hudson (1998) present a typology for classifying assessments that recognizes differences in the nature of the responses (i.e., selected, constructed, and personal) and assert that personal-response assessments such as conferences, portfolios, and self- or peer assessments should not be considered alternative assessments but rather as “alternatives in assessment” (p. 657).

On the surface, the assessment/testing dichotomy appears useful in defining a manageable domain for teachers, but in fact it is regressive in at least three ways. Most important, it attempts to reinforce the historically
instantiated division of labor between researchers and teachers, implying that researchers should continue to focus on large-scale testing and teachers should concern themselves with classroom assessments. Second, the fundamental principles of assessing language abilities are the same whether the process is termed testing or assessment. To compartmentalize the activities is to deny the relevance of teachers’ knowledge about assessment and researchers’ knowledge about testing to the other group’s practices. Third, in practice, it is impossible to draw any clear-cut distinction between testing and assessment. For example, Balliro (1993) reports that the use of the term alternative assessment has spread among those working in adult ESOL literacy programs in recent years, but she believes that “the simple distinction between standardized versus alternative assessment is of limited use” (p. 558). Balliro and others (e.g., Huerta-Macías, 1995) acknowledge the absence of a precise definition for alternative assessment but suggest that it represents an alternative perspective to the psychometric tradition, one that relies less on quantitative data and values multiple alternative sources of information in the learning environment. Attempts to apply these fuzzy distinctions, however, raise confusion rather than bring clarity.

To say that the testing/assessment distinction is not productive, however, is not to say that no differences exist among language tests and assessments. The problem is that the simple dichotomy fails to capture the many important differences among assessment possibilities to consider in selecting, constructing, using, and interpreting tests. The simple dichotomy needs to be replaced by a more complex view of the factors involved in addressing a question such as Why should I use this test for my particular purpose? These three factors are test purpose, test method, and justification for test use (see Figure 1.1).

**Understanding Test Purpose**

Test purpose can be defined through three dimensions that capture the important functions of the test. The first is the inferences to be made from test scores or, in other words, what the test is intended to measure. As
illustrated in Figure 1.2, the inference can be described very narrowly, as it pertains to what is taught and learned in a particular course; very generally as overall language proficiency; or at a number of points along a continuum. At the left end, a very specific inference about learners’ ability would be their ability to handle the language of greetings and introductions after they had worked with these functions in a language class. An example of a specific-purpose inference would be the ability to use the language of tourism to guide guests around a city. At the extreme right is general-purpose language ability. The IDEA Proficiency Test (IPT) I—Oral English is an example of a test with results that can be used to make inferences about test takers’ English language ability and readiness to enter mainstream classrooms where English is the medium of instruction. Inferences from language tests can be defined in a number of ways, including the areas of language knowledge (e.g., vocabulary) or skills (e.g., listening comprehension) one might infer on the basis of language test performance, and each of these areas can vary in terms of its context specificity.

The second dimension of test purpose is the use to be made of inferences (see Figure 1.3). Test uses refer to the types of decisions made on the basis of test scores or profiles, and such decisions are often described in terms of the stakes they hold for test takers. For example, at one extreme are the many self-tests learners can find on the Internet, which allow them to respond to a series of items and then receive a score. Such tests offer

---

**Figure 1.2. Types of Inferences That Can Be Drawn From Language Tests**

<table>
<thead>
<tr>
<th>Inference</th>
<th>Specific</th>
<th>General</th>
</tr>
</thead>
<tbody>
<tr>
<td>Specific</td>
<td>Ability connected to specific material taught</td>
<td>Specific-purpose language ability taught in class</td>
</tr>
<tr>
<td>Specific</td>
<td>Ability connected to class or program</td>
<td>Specific-purpose language ability</td>
</tr>
<tr>
<td>General</td>
<td>General-purpose language ability</td>
<td>General-purpose language ability</td>
</tr>
</tbody>
</table>

---

**Figure 1.3. Educational Uses for Language Tests**

<table>
<thead>
<tr>
<th>Use</th>
<th>Low stakes</th>
<th>High stakes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diagnosis</td>
<td>Achieve-ment in a class</td>
<td></td>
</tr>
<tr>
<td>Placement in a class or program</td>
<td>Admissions</td>
<td></td>
</tr>
<tr>
<td>Certification</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
learners an opportunity to determine how well they know a particular lexical distinction, for example. Based on their score, learners can decide for themselves whether or not they wish to study this point further. Near the other extreme would be a test such as the TOEFL, which is intended to help admissions officers decide whether or not applicants’ English is sufficient to undertake postsecondary studies in North American institutions. In other words, scores are used in a process involving great consequences for learners. Medium-stakes decisions include those made on the basis of test results within language classes and programs.

A test’s intended impacts refer to the effects that the test designer intends it to have on its users (see Figure 1.4). Entities potentially affected by a test, as Bachman and Palmer (1996) point out, include individuals (e.g., students and teachers), language classes and programs, and society. In the past, those who developed and chose tests might not have thought of positive impact as a concern, but in today’s postmodern period, a test’s impact should be considered along with its inference and use. For example, in developing the Basic English Skills Test, testers wanted to contribute to the improvement of adult education by providing a mechanism for accurate placement of students. One might argue that appropriate placement was intended to affect not only those involved in the ESOL programs but also the institutions and society in which the learners would be more likely to contribute positively as a result of achievement in ESOL classes.

Developing a test purpose statement should be the first step in choosing or developing a test. For example, a program seeking a test designed for selecting candidates for a training program on farming in the United States might develop their test purpose statement as follows:

The test is needed to measure candidates’ ability to speak about farming in English [inference] in order to select students for a short training program on farming in the United States [use] and to demonstrate to students and their sponsoring agency the level of their field-specific language ability to help focus training [impact].

---

**Figure 1.4. The Scope of Impact of Language Tests**

<table>
<thead>
<tr>
<th>Impact</th>
<th>Narrow</th>
<th>Broad</th>
</tr>
</thead>
<tbody>
<tr>
<td>On an individual student</td>
<td>On students and teachers</td>
<td>On students, teachers, classes, and programs</td>
</tr>
<tr>
<td>On students, teachers, classes, and programs</td>
<td>On students, teachers, classes, programs, and institutions</td>
<td>On students, teachers, classes, programs, institutions, and society</td>
</tr>
</tbody>
</table>
Understanding Test Method

A consideration of test purpose is possible only with a clear understanding of test method. It has been conceptualized in a number of different ways in work on language testing (e.g., Brown & Hudson, 1998; Cohen, 1994; Weir, 1990), but the most productive way of gaining new insight into test method is to draw from the perspective that treats tests as analyzable components or facets rather than as a menu of two or more types, such as alternative versus traditional or cloze versus multiple choice. This need was first addressed by Bachman (1990). In this book and a subsequent one (Bachman & Palmer, 1996), he outlined five facets of test method (also called test task characteristics) as a way of defining the important characteristics of language tests: the test setting; the testing rubric, which includes procedures for test taking expressed in the instructions as well as those for response evaluation; characteristics of the input to the learner; characteristics of the expected response; and the relationship between the input and the expected response. These facets of test method are useful for analyzing existing tests, describing the design of new tests, and envisioning possibilities for revising existing tests to better serve their purpose. In the interest of simplicity, however, in this chapter we describe existing tests by highlighting three aspects of the test method facets: the characteristics of the input; the characteristics of the expected response; and two aspects of the rubric, degree of examinee control over rubric and method for response evaluation.

Input to the Examinee

Input on a language test refers to the aurally and visually presented materials that are given to the examinee as part of the test tasks. For example, in the IELTS listening module, examinees listen to short monologues and conversations and respond to questions, often by filling in a diagram or gaps in a chart. The aural input is what examinees listen to, and the written input is what appears on the page. The aural input is for the most part linguistic whereas the written input is linguistic and nonlinguistic.

Bachman (1990) and Bachman and Palmer (1996) introduced a number of relevant categories for detailed analysis of the input, but in this simplified account, the characteristic of the input we consider is the length of any linguistic input that the test presents (see Figure 1.5). At one end of the continuum are tests composed of individual questions, such as one finds on the Combined English Language Skills Assessment in a Reading Context, a multiple-choice cloze test of grammatical knowledge and comprehension of language meaning in context. On the other end are tests that require the learner to comprehend and integrate ideas in the target language. The TOEFL reading subtest, with its reading passages and comprehension questions, is an example of a test near that end.
Examinees’ Responses

Messick (1994) cautioned against making a dichotomous distinction between multiple-choice items and open-ended performance tasks and argued that they represent “different degrees of response structure” (p. 15). Similarly, the dichotomy selected versus constructed (see Figure 1.6) is too clear-cut a distinction to describe response types meaningfully. Messick submitted that multiple-choice assessments constitute one end of a continuum whereas “student-constructed products or presentations” (p. 15) form the other.

Characteristics of the Rubric

The rubric includes all aspects of the procedures for administering and taking the test as well as the methods used to evaluate the examinees’ responses. Two aspects of the rubric are of concern here: the role of the examinee in structuring the response and the method of evaluating responses. The amount of responsibility the learner has for structuring responses can vary from no responsibility to full responsibility (see Figure 1.7). For example, examinees may respond to a restricted set of alternatives that have been structured for them. Other examples of less traditional test methods, nonetheless structured, include use of a fixed-response protocol, checklist—to assess either a product or a process—inventory, or scale. In other cases, learners construct responses or complete tasks that are partially structured for them, such as fill-in-the-blank, cloze procedure, short-answer, essay response to a prompt, and dictation or dictocomp. Partially structured forms of the open-ended protocol, checklist, inventory, or scale permit learners to respond to structured items and construct responses to open-ended items. Forms of assessment such as projects,

Figure 1.5. A Range of Possible Input in Language Tests

Length of input

Short

| Individual questions |

| | Lecture |

Extended

Figure 1.6. A Range of Possible Response Types in Language Tests

Response types

Selected

True-false questions Check-list Cloze Essay Project including essays

Constructed


demonstrations, interviews, conferences, reflection journals or learning logs, portfolios, and open-ended introspective assessments represent responses that are structured, largely or completely, by the learner.

The method of evaluation can vary in terms of three factors: who does the scoring, whether the result is a single value or a profile, and whether the score is obtained by counting the number correct, judging the level of performance, or identifying the difficulty of items that examinees can consistently answer correctly. Table 1.2 shows how these three scoring options are combined in various tests. An assessment can be scored by, for example, an independent assessor, a teacher, a peer, or the learner. Most

Table 1.2. Factors in Scoring Various Tests

<table>
<thead>
<tr>
<th>Factor: How is the test evaluated?</th>
<th>Factor: Who evaluates?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number correct</td>
<td>Independent assessor</td>
</tr>
<tr>
<td></td>
<td>Teacher</td>
</tr>
<tr>
<td></td>
<td>Peers</td>
</tr>
<tr>
<td></td>
<td>Learner</td>
</tr>
<tr>
<td>True-false questions</td>
<td>Test of English as a Foreign Language (TOEFL): Reading Test (SS)</td>
</tr>
<tr>
<td></td>
<td>Woodcock-Muñoz Language Survey: Reading/Writing (SS)</td>
</tr>
<tr>
<td>Judgment of level</td>
<td>Basic English Skills Test: Interview (PP)</td>
</tr>
<tr>
<td></td>
<td>Maculaitis Test of English Language Proficiency II: Speaking/Writing (SS)</td>
</tr>
<tr>
<td>Difficulty of items correct</td>
<td>TOEFL Computer-Based Test: Grammar (SS)</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. SS = single score; PP = profile of performance.
commercially produced forms of assessment call for structured responses scored by an independent assessor (often a machine) or a teacher. However, other forms of assessment can be scored by learners in ways that replicate the language and activities they engage in when they are learning or using the language in real or simulated ways. A third option for scoring concerns the actual result of the scoring process, that is, whether it results in a single score or a profile of performance.

**Understanding Justification of Testing Practice**

The third critical component of an assessment is the way in which the assessment is justified for its intended purpose. As described in the section Language Testing Research and Practice above, this justification refers to the validity argument that presents evidence for the appropriateness of test use in a particular situation. We gave some background for this approach to considering validity in the section Educational Measurement Theory above. The approach can be summarized as a set of principles, as shown in Table 1.3. These principles help guide the process of justifying test use, first by clarifying what validity is (i.e., an argument) and what it is not (i.e., a quality that is either present or absent in a test). The second principle asserts the primary authority of work in applied linguistics for developing a

**Table 1.3. Principles for Justifying Language Test Use Through a Validity Argument (Chapelle, 2001)**

<table>
<thead>
<tr>
<th>Principle</th>
<th>Implication</th>
</tr>
</thead>
<tbody>
<tr>
<td>Validity is an argument about the appropriateness of test use.</td>
<td>Tests are not valid or invalid; a test use is more or less valid depending on the evidence that supports the use.</td>
</tr>
<tr>
<td>Validation criteria for evaluating language tests should be based on work in applied linguistics.</td>
<td>The specific practices of language test evaluation are best guided by theory and research in language testing.</td>
</tr>
<tr>
<td>Validation criteria must be applied in view of test purpose.</td>
<td>The purpose of a test must be clearly specified in order to consider questions about validity.</td>
</tr>
<tr>
<td>Construct validity is central in test evaluation.</td>
<td>The construct that the test is intended to measure must be clearly defined, and other evaluative issues about a test should be secondary.</td>
</tr>
<tr>
<td>Tests need to be evaluated through logical and empirical analyses.</td>
<td>Methodologies for examining the test method and the performance are needed.</td>
</tr>
</tbody>
</table>
validity argument. This, of course, includes perspectives from the L2 classroom, such as the need to use tests that are consistent with instruction. The third principle links the validity argument to test purpose as defined above. Because the validity argument pertains to test use in a particular context, the purpose of the test in that context figures into the validity argument.

The fourth principle is related to the third. Construct validity refers to the extent to which evidence suggests that the test measures the construct it is intended to measure, in other words, that the inference specified as one facet of test purpose is justified. Construct validity is considered central because the test needs to measure what the user intends to measure if the test is to be used appropriately and have the intended impacts. The fifth principle refers to how validity arguments about test use are developed. It suggests that tests should be examined and that judgments should be made about test methods and test performance. Chapter 4 describes these methodologies in greater detail; here we explain important concepts associated with these analyses.

**Essential Vocabulary**

A full consideration of language tests and testing requires a working knowledge of the terminology used to express and develop knowledge of this domain. In this section we define the vocabulary used to discuss and evaluate language tests and test results. Of course, these terms are also associated with the three essential components of testing described above: test purpose, test method, and test justification.

**Test Purpose Vocabulary**

Because test purpose is central to all language testing, a number of specialized terms exist for talking specifically about it. As we have noted, test purpose consists of three components: the inference, the use, and the intended impact of the test. The inference that the test user wants to make about the examinee’s language ability (e.g., listening comprehension, grammatical competence) can be defined in several different ways, one of which is called a trait. A trait is an unobservable construct that is expected to be constant across different situations. A score awarded for grammatical competence on the Michigan English Language Assessment Battery (MELAB), for example, is thought to indicate something about the examinee’s general grammatical knowledge and ability to use that knowledge to perform certain language tasks, such as writing an academic essay or reading a journal article on the Web.

Another way of defining the inference to be made from a test is through performance, which refers to the language that the learner produces in a particular test. This use of the term performance is tied to the expression performance assessment, which refers to a test requiring learners to construct
extended responses. Like so many of the concepts discussed in this chapter, the inference can be defined as a continuum, with the trait-type definition at one end and the performance-type at the other.

What lies between the two is an interactionalist construct definition, that is, an ability as it is expected to come into play in a particular set of circumstances (Chapelle, 1998). This way of defining inferences is probably the most interesting and useful. After all, very few language teachers would expect learners to call on the same grammatical competence when they order a pizza as when they write a letter of application to a chemistry department. At the same time, the idea that a learner's performance on a language test is a single, never-to-be-repeated event is untenable. The assumption must be that one is teaching and testing for some language use that extends beyond the educational events in which learners engage. The essence of an understanding of language testing is the nature of the inference to be made based on the test performance, although considerable attention has been directed to test use as well.

As a consequence, the discussion of test use involves a number of terms, including those associated with test scores, their description, and their interpretation. A score represents a summary of an examinee's performance across one or more tasks on a test. A score may be a single number or a profile describing performance on various components of the test. A score is often expressed as a number, but some test developers consider this appearance of precision to be misleading, and, indeed, it is, because a score should be viewed as indicating a point within the range where the examinee's true ability is likely to fall. To reflect this idea more explicitly than a single score does, some test developers report performance summaries within a band or range. Such reports are called band scores.

The examination and interpretation of test results is important and is predicated on the ability to understand certain terms and concepts. Test results are typically discussed in terms of selected statistical characteristics of a group of test scores. We describe many of the most basic terms in the balance of this chapter; these terms are used to describe the tests reviewed in chapter 2 and inform the discussion of test manuals presented in chapter 3. A more complete set of statistical definitions may be found in introductory statistics books and in books on testing.

Perhaps the most widely used statistical expression is mean score, which refers to the arithmetic average of all the scores within a group. A high mean score indicates that the test was easy relative to the ability of the examinees who took it. Another way of describing a set of scores is through their dispersion, or variance, in other words, how spread out they are. If all test takers score within 1–2 points of the mean, the variance for that group would be small; however, if test takers' scores fell within a wide range, say, within 80 points of the mean, the variance would be large. The common statistical term used to express the variance is the standard deviation, which refers to the average distance of scores from the mean.
An elaborate science has developed around the investigation of the characteristics of groups of test scores. The most familiar aspect of this work is that associated with the normal distribution; this term refers to a natural tendency for scores to disperse in a particular pattern, with the greatest number of scores clustering near the middle (i.e., the mean) and fewer distributed toward the ends of the scale. In fact, the normal distribution is characterized by a certain percentage of scores that should fall near the mean (i.e., 68% within 1 standard deviation). A normal distribution of test scores is the desired outcome for most tests used to make decisions about admissions or placement because decision makers want a test that shows differences in examinees’ abilities. Imagine a director of an intensive English program who gives a placement test to 85 incoming students only to find that all learners have obtained a perfect score. This lack of distribution would indicate that the test did not identify any differences among learners’ abilities and was therefore not at all useful for dividing the group of learners into meaningful clusters that might be used to place them in classes.

Such a decision about examinees’ placement in classes is called a norm-referenced decision because it is made based on a comparison of an individual’s abilities with those of other individuals who took the test. When the test has been taken repeatedly by examinees for whom it was intended, an individual’s score can be compared with that of a larger norm group, a group from which typical performance statistics, or norms, have been obtained. Tests used in this fashion are often called norm-referenced tests. Another type of decision made on the basis of test scores is called a criterion-referenced decision. This term refers to a decision that evaluates an examinee’s performance relative to a predetermined criterion, such as a particular score or level of performance on a test. Unlike the norm-referenced decision, in which an examinee’s test score is compared with the scores of other examinees, for a criterion-referenced decision the score user sets a cutoff score. Examinees with scores above this cutoff are considered to have demonstrated a requisite level of ability.

Test impact refers to the effects of the test on those who use it. Since the 1980s, as language testers have become increasingly concerned with the broad scope of consequences of test use, the term backwash (or washback, which means exactly the same thing) has been coined to denote the effects of a test on test takers and particularly on teaching. Discussion of backwash has revealed that it is particularly potent (e.g., potentially dangerous) for high-stakes tests—tests that are used to make important decisions about examinees’ lives. Tests used for certification of proficiency for employment, for example, are considered high stakes because results from such tests determine examinees’ access to employment opportunities.

**Test Method Vocabulary**

In addition to the terms introduced in the section Understanding Test Purpose, another important term is authenticity, which, simply put, refers to
the degree to which the test tasks, including the language, resemble those that examinees will encounter beyond the test setting. Authenticity is typically argued to be a desirable quality for a test, as we explain in chapter 4. Two other terms that have become widely used in describing language test methods are discrete (point) and integrative. Discrete refers to test tasks that aim to measure a single aspect of language knowledge, whereas integrative refers to those that require examinees to call on multiple aspects of language knowledge simultaneously. Authenticity is not necessarily connected to discrete or integrative test methods.

**Test Justification Vocabulary**

Test justification, or validation, entails a number of rational and empirical procedures for analyzing the appropriateness of a test for its intended purpose. As a consequence, the process of test justification draws from a set of concepts and terms for describing the characteristics of tests. Chapter 4 explores this broader conceptualization in considerable detail. In this section we define several important types of analyses that are used to establish reliability and validity evidence during test development.

The term test item is often used as if its meaning were clear-cut and well known. Most test designers would agree that a single question on a multiple-choice test represents an item, but what about a question (prompt) to which the examinee must respond by composing an essay? Or a portfolio composed of several essays? Because different types of tests ask examinees to respond to a variety of problems, language test designers and researchers often refer to a unit of activity on a test as a task. In other words, the terms item and task on a test are functionally equivalent.

One way of investigating the quality of a test is to examine test takers' responses to each of the test tasks in a process called item analysis. This process can entail a variety of qualitative and quantitative procedures, one of which is calculation of item discrimination. An item discrimination calculation shows the relationship between examinees' performance on a single item and their performance on the test as a whole. A good item is one that the low-ability test takers tend to answer incorrectly and that the high-ability test takers answer correctly.

A correlation—one of the most widely used calculations in test analyses—is an estimate of the strength of the relationship among two or more sets of performance. A correlation coefficient represents the statistical summary of the relationship between two sets of performances and permits the analyst to determine how strongly related or how similar they are. The type of correlation calculated depends on the type of data used, and its interpretation depends on the purpose for calculating it. For example, a point-biserial correlation is a discrimination index for dichotomous items (i.e., items with responses of 0/1). The calculation estimates the relationship between the response to an item and the overall score on a test. Spearman rank-order correlations are used when one or both members in a set of data
are ordinal, such as scores derived as a level of judgment rather than as a total number correct, if the number of cases is very small, or for any other reason a normal distribution of scores cannot be expected. *Pearson product-moment correlations* are used for sets of interval data when a near-normal distribution can be expected.

Also in the family of correlational techniques is the *multiple regression analysis*, a statistical procedure used for looking at relationships among sets of scores. Unlike simple correlations, it can be used to determine which combination of variables can best predict performance. *Factor analysis* is another powerful statistical procedure that is used to reduce a large number of variables (e.g., test or questionnaire items) to a smaller number (thought to represent the underlying abilities the test developer is seeking to measure) of variables. To achieve this reduction, the test developer clusters highly correlated variables to form factors and then subjectively identifies these factors as representing specific abilities (e.g., grammatical ability or listening ability). The developers of the MELAB utilized a factor analysis procedure to provide construct validity evidence for the overall test by analyzing the similarity of the scores within two components of the test (i.e., grammar/vocabulary/reading and listening) and across two forms of the test.

*Reliability* (discussed further in chapter 4) as the term is used in testing manuals can be construed as the consistency of the test scores or the absence of error from a set of test scores. A test score is said to contain error if it reflects more than what the test developer wishes to assess; for example, in the case of a language test, error would be anything other than language ability. Measurement error can be attributed to a variety of sources: noise in the test environment, cheating, or the fact that examinees have jet lag, for example. The statistical index that expresses the amount of error estimated to be present in a set of scores is the *standard error of measurement*. This concept is a convenient way to account for the imprecision in a test, and it allows test users to estimate the range within which a test taker’s true score is likely to lie.

The opposite of error, *reliability*, is expressed as a coefficient between the values of 0 and 1. It can be calculated in several different ways, and each method of calculating reliability provides a different type of information about the reliability of the scores. *Internal consistency reliability* (e.g., using the *Kuder–Richardson [K-R] 20* statistical procedure) estimates the degree of consistency reflected in the test scores that is due to variation among the test tasks and other factors internal to the test. Because internal consistency is based on item variance, it is dependent on the number of test items and on the range of ability of the test population. An *intrarater reliability* analysis shows the degree of consistency between scores based on raters’ judgments. *Interrater reliability* indicates the degree of consistency among judgments made by the same raters on two different occasions. *Test-retest reliability* indicates the degree of consistency between test performance
at two different times. For this type of reliability to be calculated, the examinees have to take a test twice.

Some of the terms associated with the study of validity are changing as the shifts in ideas about validity mentioned above gain acceptance. Because the investigation of validity is really a process of considering evidence for and against test interpretations and uses, the terms used today generally refer to types of evidence rather than types of validity, as in the past. Many test manuals, however, continue to refer to types of validity rather than validity evidence.

*Concurrent validity*, or concurrent evidence, is established when strong positive correlations exist between the test of interest and another test or criterion of the same construct. *Concurrent* in the expression means that the other test scores have to be obtained at the same time as the score on the test of interest. *Criterion-related validity evidence* is similar to concurrent validity but is established by comparing performance on a specific test with performance on an external criterion (which may be another test or assessment, e.g., teacher judgments or course grades). High correlations between the test of interest and the specified criterion may be offered as *predictive evidence for validity*. Predictive validity is achieved by establishing how well performance on the test of interest predicts performance on some other test or criterion. *Content-related validity* is obtained by systematically collecting the judgments of experts who agree the test items are good indicators of what the test is intended to measure. This kind of validity evidence can refer to either or both of two conditions: the adequacy of the sample language being tested or the judgment of experts regarding whether the items assess what the test developer claims the items are intended to test. Evidence for *construct validity* can be drawn from any data that support the hypothesis that the test measures the construct as defined in the statement of test purpose (i.e., the inference). One of the many ways to find this evidence is through studies that demonstrate that particular examinees score systematically better than others for reasons other than the language ability tested. Such systematic error is referred to as *test bias*. Test bias can result from test methods, test takers’ characteristics, or other factors.

A term that still appears in the language testing literature but has little if any technical meaning is *face validity*. This term has been used to denote that test users and test takers feel that the test is a fair and reasonable test of what it is intended to measure. However, it is not clear how this quality should be documented or whether positive attitudes toward a test should be considered a form of validity evidence at all.
Conclusion

This chapter has laid the groundwork for examining current ESOL tests and assessments. We have reviewed the historical division that exists between language teachers and language testers as well as the changes over the past 20 years that make such a division untenable for both groups in the postmodern period of language testing. In order to reconceptualize testing and assessment in a more productive way, we have rejected the distinction between the two and introduced concepts and terms for understanding the notions within the domains of test purpose, test method, and test justification.