CHAPTER ONE

Basic concepts and terms

Language tests have become a pervasive part of our education system and society. Scores from language tests are used to make inferences about individuals' language ability and to inform decisions we make about those individuals. For example, we use language tests to help us identify second or foreign language learners in schools, to select students for admission to universities, to place students into language programs, to screen potential immigrants and to select employees. Language tests thus have the potential for helping us collect useful information that will benefit a wide variety of individuals. However, to realize this potential, we need to be able to demonstrate that scores we obtain from language tests are reliable, and that the ways in which we interpret and use language test scores are valid. If the language tests we use do not provide reliable information, and if the uses we make of these test scores cannot be supported with credible evidence, then we risk making incorrect and unfair decisions that will be potentially harmful to the very individuals we hope to benefit. Thus, if we want to assure that we use language tests appropriately, we need to provide evidence that supports this use. An important kind of evidence that we collect to support test use is that which we derive from quantitative data – scores from test tasks and tests as a whole – and the appropriate statistical analyses of these data. An understanding of the nature of quantitative data and how to analyze these statistically is thus an essential part of language testing.

Much of the data we obtain from language assessment is quantitative, consisting of numbers, and **statistics** is a set of logical and mathematical procedures for analyzing quantitative data. In order to appropriately use

statistics as a tool for test development and use, we need to understand the two contexts upon which language assessment draws. The **applied linguistics context**, which includes the nature of language use, language learning, language ability and language use tasks, provides the basis for identifying and defining the abilities we want to measure. For example, when we want to use a language test we must define what we want to measure, whether this is some aspect of language ability, progress in language learning, or the use of language in real-world settings. Applied linguistic theory also guides the design of assessment tasks, as we attempt to develop test tasks that will reflect language use outside of the test itself and that will engage the abilities we want to assess. The applied linguistics context thus provides an essential basis for the development and use of language tests. This context is discussed extensively in a number of other general books on language testing, for example Bachman, 1990; Bachman & Palmer, 1996; McNamara, 1996. For specific areas of language testing, see the other volumes in this series: Alderson, 2000, for reading; Buck, 2001, for listening; Douglas, 2000, for language for specific purposes; Luoma, 2004, for speaking; Purpura, in press, for grammar; Read, 2000, for vocabulary; and Weigle, 2002, for writing – these will only be touched on here and there in this book, as needed. The **measurement context** is concerned with the relationship between the quantitative results of assessments (numbers) on the one hand and their meaning, interpretation and use on the other. An understanding of measurement theory will also inform the decisions we make about the appropriate uses of statistics. As with the applied linguistics context, the measurement context for language testing is dealt with in a large number of textbooks (e.g. Hopkins, 1998; Linn & Gronlund, 2000). However, since this context is probably less familiar to many language assessment practitioners than the applied linguistics one, the measurement context will be discussed more extensively in this book.

This chapter will cover some of the basic concepts and terms that are essential to the appropriate use of statistics in the development and use of language assessments. It will cover the following topics:

- Test usefulness
- The nature of language assessment
- The uses of language assessments
- The nature of quantitative data
- The limitations on measurement

- Frame of reference (norm-referenced and criterion-referenced approaches to measurement)
- Using statistics for understanding and interpreting test scores

Test usefulness

An overriding consideration in designing, developing and using language tests is that of **test usefulness**, which Bachman and Palmer (1996) define as comprising several qualities: reliability, construct validity, authenticity, interactiveness, impact and practicality. The usefulness of a given test depends to a great extent on how test takers perform on the test. This implies that the evaluation of test usefulness must include the empirical investigation of test performance. There are two aspects of test performance that we need to investigate in our evaluation of test usefulness: the processes or strategies test takers use in responding to specific test tasks and the product of those processes or strategies – individuals' responses to the test tasks and the scores that they obtain. In order to evaluate the usefulness of a given test, we need to investigate both aspects. While the investigation of the processes and strategies test takers employ provides important information for the evaluation of test usefulness, this book will focus on quantitative statistical procedures for investigating the products of test performance, focusing on the scores that test takers obtain, either from individual test tasks, parts of tests or from entire tests. These quantitative procedures are of primary relevance to two of the qualities of measurement, reliability and construct validity. Bachman and Palmer (1996) define these qualities as follows:

> RELIABILITY: consistency of measurement. A reliable test score will be consistent across different characteristics of the testing situation.
>
> CONSTRUCT VALIDITY: the meaningfulness and appropriateness of the *interpretations* that we make on the basis of test scores. Test scores are to be interpreted appropriately with respect to a specific *domain of generalization*, or set of tasks in a specific target language use domain. (Bachman & Palmer, 1996: 19, 21)

It is the responsibility of *test developers* to go beyond mere assertions of reliability and construct validity, and to provide evidence to test users that *demonstrates* that their tests have the qualities the developers claim. That is, test developers must provide evidence that supports the claims they make about how test scores are to be interpreted and used. Similarly,

it is the responsibility of *test users* to require test developers to provide such evidence, and to use this evidence appropriately and ethically in their own selection and use of language tests.

Test developers and test users can employ many different procedures and activities to collect the evidence for assessing the usefulness of tests for the particular purposes, test takers and situations for which they are intended. This evidence will ideally include both quantitative data, such as test scores, scores for items or tasks, or responses to questionnaires and self-ratings, and qualitative data, such as observations, verbal self-reports by test takers, or samples of language produced during the assessment, that provides information about the usefulness of a given test. This book will focus on the kinds of quantitative data that can be collected, and some of the statistical analyses that can be used to help us evaluate the usefulness of the tests we develop and use. The statistical procedures described in this book can be used with any quantitative data, and they are relevant to the investigation of the qualities of usefulness.

The nature of language assessment

Settings for language assessment

Language assessment takes place in a wide variety of situations, including educational programs and real-world settings. In educational programs, the results of assessments are most commonly used to describe both the processes and outcomes of learning for the purposes of diagnosis or evaluating achievement, or make decisions that will improve the quality of teaching and learning and of the program itself. In real-world settings, language assessment is often used to inform decisions about employment, professional certification and citizenship.

Assessment concepts and terms

Assessment

The term 'assessment' is commonly used with a variety of different meanings. Indeed, the term has come to be used so widely in so many different ways in the fields of language testing and educational measurement that there seems to be no consensus on what precisely it means. Furthermore,

a number of other terms are frequently used more or less synonymously to refer to assessment. For the purpose of this book, **assessment** can be thought of broadly as the process of collecting information about a given object of interest according to procedures that are systematic and substantively grounded. A product, or outcome of this process, such as a test score or a verbal description, is also referred to as an **assessment**.

The object of interest in a language assessment is most frequently some aspect of language ability. In some situations we may also be interested in gathering information about other qualities of individuals, such as their attitudes toward the test, or their background characteristics, such as age, native language, or level of education.

There are two requirements that distinguish assessment from informal observations and reports: that the assessment is systematic and substantively grounded. By **systematic** I mean that assessments are designed and implemented in a way that is clearly described and potentially replicable by other individuals. That is, assessment is carried out according to explicit procedures that are open to public scrutiny. These procedures provide the link between what we want to assess and our observations. Thus, although I might be able to describe in great detail the qualities of a particular person on the basis of my observations and a conversation at a party, this would not constitute an assessment. This is because I would probably not be able to describe the way I observed this person and the nature of our conversation with enough precision for me to replicate it and come up with the same description, or for another person to replicate my observations and conversation. This **systematicity requirement** in assessment is closely linked to reliability.

It is also essential for language assessments to be substantively grounded, because this provides the basis for interpreting the results of our assessment, whether these be quantitative or qualitative. By **substantively grounded**, I mean that the assessment must be based on a widely-accepted theory about the nature of language ability, language use or language learning, or prior research, or accepted and current practice in a particular field. Informal observations and reports, such as in the party example above, generally fail the substantive requirement of assessment, since most people, other than language testers, do not engage in such activities with the intent of assessing an individual's capacity for language use. That is, informal observations and conversations are generally not informed by an explicit theory of language use or a course syllabus. This **substantive requirement** in assessment is closely linked to the quality of validity.

Assessment can draw information from a wide range of elicitation, observation and data-collection procedures, including multiple-choice tests, extended responses, such as essays and portfolios, questionnaires, oral interviews, introspections and observations. The results of assessments can be reported both quantitatively, as numbers, such as test scores, ratings, or rankings, and qualitatively, as verbal descriptions, or as visual or audio images.

Measurement

Another term that is often associated with assessment is 'measurement', and I will adopt Bachman's (1990: 18) definition of this term as follows:

> **Measurement** is the process of quantifying the characteristics of an object of interest according to explicit rules and procedures.

A product, or outcome of this process is also referred to as a **measurement**, or a **measure**.

Measurement is one type of assessment that involves quantification, or the assigning of numbers, and this characteristic distinguishes measures from non-quantitative assessments such as verbal descriptions or visual images. We assign numbers not to people or groups, but to the *attributes* of people or groups. Furthermore, in language testing, the attributes we generally want to measure are not directly observable physical features, such as height or eye color, but are *unobservable* abilities or attributes, sometimes referred to as traits, such as grammatical knowledge, strategic competence or language aptitude. As with other types of assessment, measurement must be carried out according to explicit rules and procedures, such as are provided in test specifications, criteria and procedures for scoring, and directions for test administration. These specifications and procedures provide the link between the unobservable ability we want to measure and number we assign to observable performance.

Test

Another term that needs to be clarified is 'test', which Carroll (1968) defined as follows:

> . . . a **test** is a procedure designed to elicit certain behavior from which one can make inferences about certain characteristics of an individual. (Carroll, 1968: 46)

A test is a particular type of measurement that focuses on eliciting a specific sample of performance. The implication of this is that in designing and developing a test we construct specific tasks or sets of tasks that we believe will elicit performance from which we can make the inferences we want to make about the characteristics of individuals (see Alderson, Clapham, & Wall, 1995; and Bachman & Palmer, 1996, for discussions of designing and developing language test tasks).

Evaluation

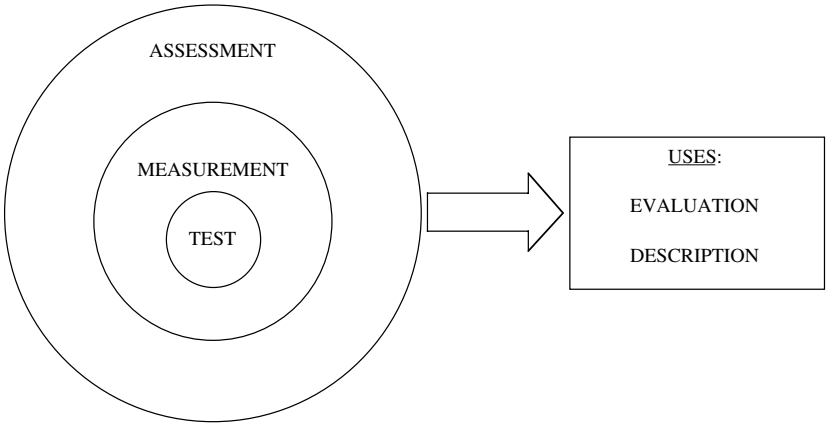
Another term that is often associated with assessment is 'evaluation'. **Evaluation**, which involves making value judgments and decisions, can best be understood as one possible *use* of assessment, although judgments and decisions are often made in the absence of information from assessment. The use of assessment for evaluation is particularly common in educational programs, where we often use information from assessment to make decisions about selection and placement and to assign grades or marks. In some situations the primary purpose of assessment is to provide a **description** of the attributes of individuals, that is, for making interpretations, or inferences, about individuals on the basis of the information that is collected in the assessment. This purpose is particularly common in applied linguistics research, where the focus is often on describing processes, individuals and groups, and the relationships among language use, the language use situation, and language ability.

The relationships among assessment, measurement, tests, and their uses are illustrated in Figure 1.1 overleaf.

The uses of language assessments

One use of assessments is to make inferences about abilities or attributes such as lexical knowledge, sociolinguistic awareness, language aptitude, or motivational orientation. Assessments can provide information about attributes of individuals such as their relative strengths and weaknesses, their achievement in a language course, or their levels of proficiency in a language. The descriptions, inferences, or interpretations we make on

Figure 1.1 Relationships among assessment, measurement, test, and their uses for description and evaluation in different settings



the basis of assessments provide input into the decisions we may need to make, both about individuals and about programs.

We also use assessments as a basis for making decisions. These decisions can be about either individuals or programs, which Bachman (1981) refers to as 'micro-evaluation' and 'macro-evaluation', respectively. Bachman (1990) describes in detail the various types of decisions that are made on the basis of assessment in educational programs, and these can be summarized as follows:

- Decisions about individuals, such as
  - selection for admission or employment
  - placement
  - diagnosis
  - grading/marking
  - certification
- Decisions about programs
  - formative, relating to making changes to improve an existing program;
  - summative, relating to continuing an existing program or implementing a new program.

Relative and absolute selection decisions

The decisions that we make on the basis of assessments are of two general kinds: relative and absolute. A **relative decision** is one in which we select or reward test takers based on their relative standing in a group on some ability or attribute. Relative decisions are typical of situations in which the places or resources available are limited and can be allocated to only a fixed number of individuals. In such situations, the decision maker generally wants to allocate these places to the individuals who are the highest among the group being considered. College admissions decisions, for example, are typically relative, since in most cases only a limited number of individuals can be admitted, and those who are admitted are generally at the top of the group who apply. Other examples of relative decisions would be 'grading on the curve', in which only the top five percent, say, of the students in a class receive As, and the hiring of the top person for a job, from a pool of many applicants.

An **absolute decision** is one in which we select or reward test takers on the basis of their level of knowledge or ability, according to some pre-determined criteria. Absolute decisions are typical of situations in which the places or resources available are unlimited and can be given to an unlimited number of individuals. In such situations, the decision maker selects or rewards those individuals who possess the knowledge or level of ability required. Certification decisions are absolute decisions, since only those individuals who achieve a certain pre-determined level of performance on an examination may be considered to be qualified in a given area. Examples of tests used for certification decisions include driving exams, bar exams for lawyers and medical exams for doctors. Other examples of absolute decisions would be awarding a grade of A to all students who demonstrated mastery of the course content, or hiring those individuals who meet certain minimum standards, irrespective of how many individuals this might be.

Relative importance of decisions

Not all of the decisions that are made on the basis of assessment results are equally important in terms of their effects on individuals and programs, and it is common to distinguish between high-stakes and low-stakes decisions. Any time we make a decision, there is a possibility that we will make the wrong decision, such as admitting an individual who will

eventually fail into a program, or not admitting someone who would succeed. These decision errors will involve certain costs. **High-stakes** decisions are major, life-affecting ones where decision errors are difficult to correct. Because of the importance of their effects, the costs associated with making the wrong decision are very high. In large-scale tests the potential effects of decision errors are of particular concern, since the lives of many individuals are affected. **Low-stakes** decisions, on the other hand, are relatively minor ones, where decision errors are relatively easy to correct. Because their effects are limited and errors are easy to correct, the costs associated with making the wrong decision are relatively low. These differences are illustrated in Table 1.1.

Table 1.1 *Relative importance of decisions*

High-Stakes	**Low-Stakes**
• *Major*, life-affecting decision	• *Minor* decision
• Decision errors *difficult* to correct	• Decision errors *easy* to correct
• *High* costs of making wrong decision	• *Low* costs of making wrong decision

Although I have described the relative importance of decisions as either high-stakes or low-stakes, in fact, as the above examples illustrate, there is a range of importance, from very high to very low. An example of a *very* high-stakes decision would be that of admission to universities in a country where this decision is based largely, if not entirely, on the results of a nationwide university entrance examination. In this case, the lives of individuals are very strongly affected, since if they are not admitted in a given year, they may have to wait another year to try again, or may never be admitted to a university at all. In a situation such as this any decision errors, that is not admitting applicants who would have succeeded, on the one hand, or admitting applicants who eventually fail, on the other, are very difficult to correct, because these errors may not become apparent for months, if not years. The costs of not admitting students who would succeed in an academic program are difficult to estimate, but can be thought of in terms of opportunity lost, to the person, to the program, and potentially to society. These costs are likely to be quite large, given the importance of education to the economic well-being of any country. Admitting a person who eventually fails costs time and effort, on the part of both the person and those who are involved with running the program, such as teachers and administrators, as well as resources. These costs are also likely to be very high, given the costs of higher education in most countries.

An example of a *relatively* high-stakes decision would be that of hiring individuals for a job, where the assessment is likely to involve a variety of approaches, including both tests and other forms of assessment, such as portfolios or interviews. Even if there is only one job, the decision is a high-stakes one for each applicant, since it may mean the difference between being able to adequately provide for the needs of a family and not being able to survive economically. As with the first example, correcting decision errors quickly may be difficult. Applicants who are not hired may subsequently seek jobs elsewhere, and it may be several months before the company can determine whether or not the person who is hired will become a productive employee. The cost to the company of hiring an individual who will not become a productive employee is quite high, as is the cost of not hiring someone who would have been able to contribute to the company.

An example of a *relatively* low-stakes decision would be a classroom teacher's decision to move on to the next lesson, based on the class's performance on a quiz. In this case, the decision is a relatively minor one, since relatively few individuals are affected, and a wrong decision can be quite easily corrected. If the teacher discovers, from the students' classroom performance, that they are not ready to proceed to the next lesson, he can go back and review the material from the previous lesson.

An example of a *very* low-stakes decision would be an individual's decision to study a foreign language, based on his self-assessment of his language aptitude, using a structured questionnaire. In this case only one individual is affected, and he can very quickly reverse his decision if he finds that he is not learning as quickly as he had expected and is not likely to achieve his desired level of proficiency, or eventually loses interest in studying.

The nature of quantitative data

In order to determine what statistical procedures are appropriate to use for analyzing the results of language tests, we need to understand the nature of the data we have collected. Although the quantitative data we analyze with statistics consists of numbers, these numbers come from many different types of assessments, and have different properties. Thus, in order to analyze quantitative data appropriately and meaningfully, we need to understand the specific assessment procedures or instruments we have used to collect the data, and the properties of the numbers these procedures provide.