Statistical Analyses for Language Assessment

Lyle F. Bachman

CAMBRIDGE
UNIVERSITY PRESS

PUBLISHED BY THE PRESS SYNDICATE OF THE UNIVERSITY OF CAMBRIDGE
The Pitt Building, Trumpington Street, Cambridge, United Kingdom

CAMBRIDGE UNIVERSITY PRESS
The Edinburgh Building, Cambridge CB2 2RU, UK
40 West 20th Street, New York, NY 10011–4211, USA
477 Williamstown Road, Port Melbourne, VIC 3207, Australia
Ruiz de Alarcón 13, 28014 Madrid, Spain
Dock House, The Waterfront, Cape Town 8001, South Africa

http://www.cambridge.org

© Cambridge University Press 2004

This book is in copyright. Subject to statutory exception
and to the provisions of relevant collective licensing agreements,
no reproduction of any part may take place without the
written permission of Cambridge University Press.

First published 2004

Printed in the United Kingdom at the University Press, Cambridge

Typeface 9.5/13pt Utopia. *System* QuarkXPress™ [SE]

A catalogue record for this book is available from the British Library

Library of Congress Cataloguing in Publication data

ISBN 0 521 80277 6 hardback
ISBN 0 521 00328 8 paperback

Contents

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

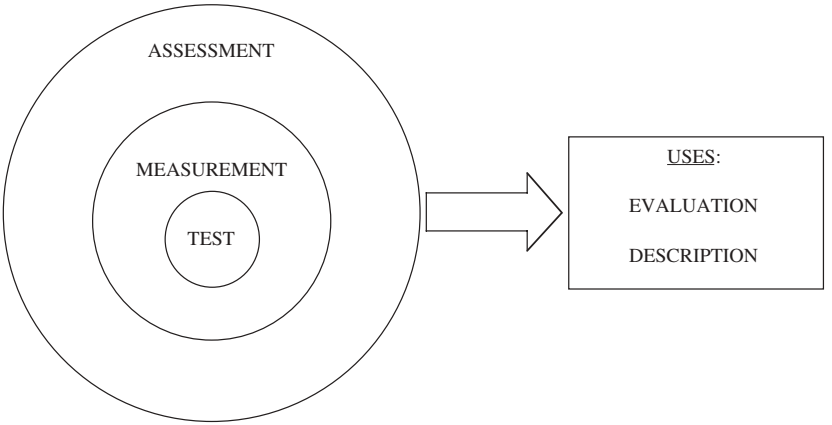
Another term that is often associated with assessment is 'evaluation'. **Evaluation**, which involves making value judgments and decisions, can best be understood as one possible *use* of assessment, although judgments and decisions are often made in the absence of information from assessment. The use 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 predetermined 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.

Steps in the measurement process

As indicated above, measurement is a process of assigning numbers to attributes of individuals or groups according to specific rules and procedures. This process consists of three logically ordered steps: (1) defining the construct conceptually, (2) defining the construct operationally, and (3) quantifying our observations (Bachman, 1990: 40–45). Unlike physical attributes, such as height, eye color and shoe size, the attributes we generally want to assess, such as cognitive style, pronunciation accuracy, knowledge of grammar, or type of planning, cannot be observed directly, and for this reason we need to define these in a way that will enable us to link our observations of performance, such as written responses to a questionnaire, spoken utterances, samples of writing, and amount of time on task, to these unobservable attributes. The steps in measurement provide a basis for this linkage, or for making inferences about unobservable attributes on the basis of observed performance.

Defining the construct conceptually

Although it is generally quite adequate, for purposes of general description or discussion, to use terms such as language ability, knowledge of grammar, or reading comprehension to describe the attributes of individuals without developing precise definitions for these, for the purpose of *measuring* attributes such as these, we must define them precisely enough to distinguish them from other attributes, and to understand their relationship to other, similar attributes. The term 'language ability', for example, is understood in many different ways, and even though it probably has a common core of general meaning for most people, quite different theoretical, or conceptual views of this attribute can be seen in the language testing literature (e.g. Bachman, 1990; Canale & Swain, 1981; Carroll, 1961a; Chapelle, 1998; Lado, 1961; Lowe, 1986; Oller, 1979). These different conceptual views have informed differing approaches to the measurement of this ability. Similarly, the term, 'language achievement' is likely to have quite different meanings for teachers in language programs that are based on different syllabi and that incorporate different learning objectives and different types of learning activities. Thus, if we want to develop a procedure for measuring an attribute or ability, we need to construct a precise definition of this. The following definition, for example, identifies 'organizational knowledge' as a component of lan-

guage knowledge, as opposed to topical knowledge or strategic competence, while also indicating that it consists of several subcomponents:

> Organizational knowledge is that component of language knowledge that is involved in controlling the formal structure of language for producing or comprehending grammatically acceptable utterances or sentences, and for organizing these to form texts, both oral and written. There are two areas of organizational knowledge: grammatical knowledge and textual knowledge.
>
> (Bachman & Palmer, 1996: 67–68)

We need to define this ability in a way that is appropriate to the specific testing situation, that is, the particular purpose for which the measure is intended and the particular individuals who will be tested. For example, for a particular testing situation, we may want to focus only on individuals' knowledge of vocabulary and cohesive markers, in which case we would define the construct – organizational knowledge – more narrowly than in the example above.

When we define an ability in this way, it becomes the construct about which we want to make inferences for this particular testing situation. A **construct**, then, is an attribute that has been defined in a specific way for the purpose of a particular measurement situation. Bachman and Palmer (1996) point out that **construct definitions** are generally based on either a theory of language ability, or proficiency, or on the content of an instructional syllabus.

Defining the construct operationally

The second step in measurement is to specify the procedures and conditions under which we will observe or elicit the performance that will enable us to make inferences about the construct we want to measure. These procedures and conditions are specified in the test specifications, or blueprint, which include detailed information about the types and numbers of test tasks to be included and how these will be ordered in the test, the amount of time to be allowed, and how responses to these test tasks will be scored (see Alderson et al., 1995, Ch. 2; Bachman & Palmer, 1996, Ch. 9; and Davidson & Lynch, 2002, for extensive discussions of test specifications). In specifying these procedures, we are defining the construct operationally, and these measurement procedures, that is, the test tasks and how they are to be scored, thus become the **operational definition** of the construct.

In language assessment, we define constructs operationally in many different ways, which is another way of saying that we use a wide variety of assessment procedures. In some situations, we may find scores from a paper-and-pencil or computer-based test most useful, while at other times we may feel it will be most useful to collect samples of natural speech, under as nearly natural conditions as possible, and provide a rich description of this. This is simply to illustrate the fact that our observations do not necessarily need to be test scores, or even numbers. It also illustrates the fact that what has been described up to this point, applies more generally to the process of assessment, since we could analyze and describe the natural speech samples that we have obtained qualitatively, with verbal descriptions and illustrative examples.

Quantifying our observations

The third step in measurement is to determine the specific procedures we will follow to quantify, or assign numbers to, our observations of performance, or variables. It is this step that distinguishes measurement from other forms of assessment. The particular measurement procedure we use will depend on the nature of the attribute we want to measure and the way in which we have obtained the performance to be measured.

When we use elicitation procedures, such as tests, questionnaires or interviews, to obtain performance, there are essentially two different ways in which we can assign numbers: (1) judge the quality, or level, of the performance according to a rating scale with defined levels, or (2) count the scores or marks for the individual tasks or items. For example, numbers might be assigned to writing samples obtained from an essay examination by asking expert judges to rate these for the knowledge of appropriate features for marking cohesion, on a scale such as the following:

0 Zero *No evidence of knowledge* of textual cohesion
Range: zero
Accuracy: not relevant
1 Limited *Limited knowledge* of textual cohesion
Range: few markers of cohesion
Accuracy: relationships between sentences frequently confusing
2 Moderate *Moderate knowledge* of textual cohesion
Range: moderate range of explicit devices

Accuracy: relationships between sentences generally clear but could often be more explicitly marked

3 Extensive *Extensive knowledge* of textual cohesion
Range: wide range of explicit cohesive devices including complex subordination
Accuracy: highly accurate with only occasional errors in cohesion

4 Complete *Evidence of complete knowledge* of cohesion
Range: evidence of complete range of cohesive devices
Accuracy: evidence of complete accuracy of use

(Bachman & Palmer, 1996: 278–9)

Individual tasks can be scored or marked as right or wrong (1, 0), in which case the maximum score is 1, or as partial credit, in which case the maximum score may be more than 1, depending on how many points or marks the task is worth. Another example of the counting approach is in questionnaires that include items to which individuals respond on a multi-point scale (sometimes called a 'Likert scale', pronounced to rhyme with 'lick'). These two different approaches to scoring – judging and counting – are appropriate for different types of tasks. The counting approach is used most typically with tasks in which individuals select a response from among several choices that are given, respond along points on a scale, or with items that require completion or short-answers. The judging approach is used typically with tasks that require test takers to produce an extended sample of language, such as in a composition test or an oral interview (see Alderson *et al.*, 1995, Ch. 5, and Bachman & Palmer, 1996, Ch. 11, for discussions of scoring, or marking, procedures; see Alderson, 1991, and Pollitt, 1991, for discussions of counting and rating as scoring procedures).

When we collect our data by observing, rather than by eliciting responses to tasks, we can assign numbers in two ways. One way is to assign numbers to members of groups that have different attributes, such as native language, occupation or academic major, in order to indicate the categories of attributes to which they belong. For example, if the individuals we observe belong to mutually exclusive groups on some attribute, such as native language, we could use numbers to represent the different values of this attribute. Thus, we might use a '1' to represent native speakers of Amharic, a '2' to represent native speakers of Arabic, and so on. Another way to assign numbers to observations is to count the number of occurrences of a particular attribute. Thus, we could count how many individuals are native speakers of Amharic, Arabic, and so on. Or, if we wanted to measure an individual's mastery of a particular

cohesive marker, we might count how many times that marker is used appropriately and inappropriately in a sample of language, either written or spoken. Counts such as these can be reported either as frequencies of occurrence, or as proportions or percentages of all the different individuals or occurrences in the study. For example, we could report the percentages of all the individuals in the study who were native speakers of different languages. Or, we could report the percentages of all occurrences of a particular request form that were appropriate and inappropriate.

The way we quantify our observations will depend on a number of factors, such as the purpose of the measurement, the way we have defined the construct, and the procedures we use to elicit or observe performance. These different ways of quantifying our observations yield numbers with different measurement properties, or that provide different kinds and amounts of information. It is particularly important to keep this in mind when we decide how to quantify our observations, as this will affect, to some extent, the kinds of statistical analyses we use; but more importantly, it will affect the way we interpret the results of these statistical analyses. The different measurement properties of numbers are discussed below, under 'Measurement scales'.

Variables and constructs

The score that we obtain from a measurement procedure is called a **variable**, which is a term for something that can have different values, or which can vary. For example, we might design a multiple-choice test to measure the construct, organizational knowledge, as defined on page 15 above. If different individuals were to complete this test, they most likely would not perform in exactly the same way, so that their scores would vary. In this case, the variable – test score – would have different values, or would vary, from one individual to the next. If we interviewed these same individuals and assigned a rating for organizational knowledge to their speech samples, then we would have a different variable – interview rating – as an indicator of this construct. If we gave the test of organizational knowledge to individuals at the beginning of a language course and again at the end of the course, we would have separate measures, pre-test score and post-test score, or two variables for each individual, as indicators of this construct.

The distinction between constructs and the variables that represent

them is a critical one, because we must always keep in mind that we cannot observe constructs themselves directly. We can only make inferences about these constructs on the basis of our observations of performance. The operational definition provides the essential link between our construct definition, on the one hand and our numbers, or variables, on the other. In other words, it is the operational definition that provides the logical basis for interpreting numbers, or variables, as indicators of the constructs we want to measure.

Why the measurement process is essential for the statistical analyses of test scores

The statistical analyses that we use with test scores can be applied to any set of numbers we might come up with. However, unless these numbers are consistent indicators and can be clearly linked to underlying constructs or attributes, the results of our statistical analysis will be meaningless. The steps in measurement provide the basis for investigating and demonstrating the reliability of our test scores and the construct validity of our interpretations. The reliability of our test scores depends, to a large extent, on how well we have implemented the second step of measurement – carefully specifying the measurement procedures to be used and following these specifications, in both designing and administering these procedures. Reliability will also depend on the third step – how we quantify our observations – since, as will be seen in Chapter 5, the way we estimate reliability statistically depends, in part, on the level of measurement (discussed in the next section) of our test scores. The construct validity of our score interpretations depends on the clarity with which we have defined the construct (first step) and the appropriateness of the specific procedures we have used to obtain our test scores (second step).

Measurement scales

When we use measurement to assign numbers to our observations, these numbers are variables that represent the attribute we intend to measure. The information these numbers contain will depend on how we have defined the attribute as a construct, and the rules and procedures we have used to measure it. From the examples above, it is clear that we can use a variety of procedures for quantifying our observations, and these

different measurement procedures produce sets of numbers, or **measurement scales**, that contain different kinds and amounts of information. We can identify four different measurement scales: nominal, ordinal, interval and ratio. Because these different scales provide increasing amounts of information, from nominal scales up to ratio scales, they are sometimes called **levels of measurement**.

Nominal scales

A **nominal scale** consists of numbers that are used to name, or stand for different, mutually exclusive groups or categories of individuals, in terms of a particular attribute, such as native language, academic discipline, country of residence, or occupation. In nominal measurement, each individual is classified into one and only one category that represents the unique group that has a particular attribute, such as native speaker of Zulu, student of criminal psychology, resident of Tahiti, or swimming pool technician. This measurement procedure of assigning different numbers to different groups of individuals will produce a nominal scale. The numbers we use to represent the attribute are arbitrary, since any number can be assigned to any group, as long as the categories are mutually exclusive and each category is assigned a unique number. Counts of entities, such as the number of individuals in a particular native language group or the number of appropriate uses of a particular speech act, whether these are reported as frequencies, proportions or percentages, also constitute nominal scales. One particular type of nominal scale, in which there are only two categories, is called a **dichotomous scale**. A dichotomous scale that is of particular interest in language assessment is the scale we obtain when we score responses to individual test tasks as either right or wrong, assigning scores of '1' and '0', respectively. These scores, sometimes called 'item scores', constitute a nominal scale, while the total score that we obtain by adding up these item scores is generally treated as an interval scale (see below).

Nominal scales provide information only about the *distinctiveness* of individuals on the attribute, and it is this property of numbers that enables us to differentiate among values for a given attribute. To put it another way, nominal data answer the question, 'Are they different?'

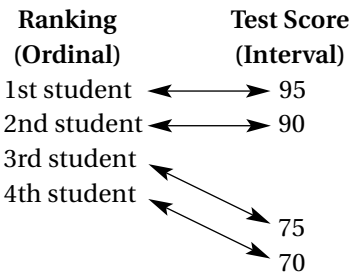
Ordinal scales

Numbers can also be *ordered*, so that any given number will be larger than some numbers, and smaller than others. If the attribute we want to measure varies in amount, so that individuals have more or less of it, or are at different levels on this attribute, then we might assign numbers in a way that will capture this information. When we have used a measurement procedure that yields numbers that indicate differing levels of an attribute, we obtain an **ordinal scale**, in which the numbers are not only distinct from each other, but are also ordered with respect to each other. A common example of an ordinal scale would be a teacher's ranking of his students in terms of achievement. Scores obtained by the judging scoring method discussed above, in which the quality or level of performance is judged according to an ordered set of descriptions, or rating scales, may also constitute an ordinal scale.

Ordinal scales provide information about both the distinctiveness and the *ordering* of individuals on the attribute. In other words, ordinal data answer two questions: 'Are they different?' and 'Which is larger?' Because they provide additional information, ordinal scales are considered to be a higher level of measurement than nominal scales.

Interval scales

An additional type of information, that we are frequently interested in obtaining from our measures, is that of *how large* the difference is between one number and another. That is, we often want to know not only whether one number is larger than another, but also how much larger. For example, we might rank four students as the highest achiever, second highest, third highest and fourth, in terms of their classroom performance. For many assessment purposes, this information would be sufficient. Suppose, however, that we wanted to know how much more of the course content the highest student had mastered than the second had, or which of these students had mastered a sufficient amount of the course content to be promoted to the next level or grade. For this purpose, a ranking – an ordinal scale – would not be sufficient, because it only provides information about the relative ordering among the four students. For either of these purposes, we might decide to give these students a test of their achievement, based on a representative sampling of the course objectives, from which we might obtain the following scores for the four students who ranked first through fourth:



From these scores we can see that differences between the scores of the first and second students and between the third and fourth students, which are 5 points, are much smaller than that between the second and third highest students, which is 15 points. In addition, if we had set a score of 80 as the criterion for mastery, we see that only the highest two students achieved scores that were considered high enough to indicate mastery of the course.

In order to interpret the scores from this test in this way, we would need to assume, or demonstrate through research, using some of the statistical procedures described in this book, that the test scores constitute an **interval scale**, in which the differences, or intervals, between the different points on the score scale are equal. Because the differences between the different points on an interval scale are equal, the addition and subtraction of such numbers yield results that are meaningful. In the above example, if we can assume the test scores constitute an interval scale, it would be meaningful to say that the top student scored 5 points more than the second student, and 25 more than the fourth student. Likewise, it would be meaningful to say that the difference of 5 points between the first and second students is the same as that between the third and fourth students. Such comparisons are not possible with ordinal-scaled data.

In addition to information about distinctiveness and ordering, interval scales provide information about the *amount of difference* between different scores in the measurement scale. Interval scales thus answer three questions: 'Are they different?', 'Which is larger?' and 'How much larger?' Because they provide additional information, interval scales are considered to be a higher level of measurement than either nominal scales or ordinal scales.