2 Measurement

Introduction

In developing language tests, we must take into account considerations and follow procedures that are characteristic of tests and measurement in the social sciences in general. Likewise, our interpretation and use of the results of language tests are subject to the same general limitations that characterize measurement in the social sciences. The purpose of this chapter is to introduce the fundamental concepts of measurement, an understanding of which is essential to the development and use of language tests. These include the terms 'measurement', 'test', and 'evaluation', and how these are distinct from each other, different types of measurement scales and their properties, the essential qualities of measures – reliability and validity, and the characteristics of measures that limit our interpretations of test results. The process of measurement is described as a set of steps which, if followed in test development, will provide the basis for both reliable test scores and valid test use.

Definition of terms: measurement, test, evaluation

The terms 'measurement', 'test', and 'evaluation' are often used synonymously; indeed they may, in practice, refer to the same activity.[1] When we ask for an evaluation of an individual's language proficiency, for example, we are frequently given a test score. This attention to the superficial similarities among these terms, however, tends to obscure the distinctive characteristics of each, and I believe that an understanding of the distinctions among the terms is vital to the proper development and use of language tests.

Measurement

Measurement in the social sciences is the process of quantifying the characteristics of persons according to explicit procedures and rules.[2]

This definition includes three distinguishing features: quantification, characteristics, and explicit rules and procedures.

Quantification

Quantification involves the assigning of numbers, and this distinguishes measures from qualitative descriptions such as verbal accounts or nonverbal, visual representations. Non-numerical categories or rankings such as letter grades ('A, B, C . . .'), or labels (for example, 'excellent, good, average . . .') may have the characteristics of measurement, and these are discussed below under 'properties of measurement scales' (pp. 26–30). However, when we actually use categories or rankings such as these, we frequently assign numbers to them in order to analyze and interpret them, and technically, it is not until we do this that they constitute measurement.

Characteristics

We can assign numbers to both physical and mental characteristics of persons. Physical attributes such as height and weight can be observed directly. In testing, however, we are almost always interested in quantifying mental attributes and abilities, sometimes called traits or constructs, which can only be observed indirectly. These mental attributes include characteristics such as aptitude, intelligence, motivation, field dependence/independence, attitude, native language, fluency in speaking, and achievement in reading comprehension.

The precise definition of 'ability' is a complex undertaking. In a very general sense, 'ability' refers to being able to do something, but the circularity of this general definition provides little help for measurement unless we can clarify what the 'something' is. John B. Carroll (1983c, 1987a) has proposed defining an ability with respect to a particular class of cognitive or mental tasks that an individual is required to perform, and 'mental ability' thus refers to performance on a set of mental tasks (Carroll 1987a: 268). We generally assume that there are degrees of ability and that these are associated with tasks or performances of increasing difficulty or complexity (Carroll 1980, 1987a). Thus, individuals with higher degrees of a given ability could be expected to have a higher probability of correct performance on tasks of lower difficulty or complexity, and a lower probability of correct performance on tasks of greater difficulty or complexity.

Whatever attributes or abilities we measure, it is important to understand that it is these attributes or abilities and *not* the persons themselves that we are measuring. That is, we are far from being able to claim that a single measure or even a battery of measures can adequately characterize individual human beings in all their complexity.

Rules and procedures

The third distinguishing characteristic of measurement is that quantification must be done according to explicit rules and procedures. That is, the 'blind' or haphazard assignment of numbers to characteristics of individuals cannot be regarded as measurement. In order to be considered a measure, an observation of an attribute must be replicable, for other observers, in other contexts and with other individuals. Practically anyone can rate another person's speaking ability, for example. But while one rater may focus on pronunciation accuracy, another may find vocabulary to be the most salient feature. Or one rater may assign a rating as a percentage, while another might rate on a scale from zero to five. Ratings such as these can hardly be considered anything more than numerical summaries of the raters' personal conceptualizations of the individual's speaking ability. This is because the different raters in this case did not follow the same criteria or procedures for arriving at their ratings. Measures, then, are distinguished from such 'pseudo-measures' by the explicit procedures and rules upon which they are based. There are many different types of measures in the social sciences, including rankings, rating scales, and tests.[3]

Test

Carroll (1968) provides the following definition of a test:

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

From this definition, it follows that a test is a measurement instrument designed to elicit a specific sample of an individual's behavior. As one type of measurement, a test necessarily quantifies characteristics of individuals according to explicit procedures. What distinguishes a test from other types of measurement is that it is designed to obtain a specific sample of behavior. Consider the following example. The Interagency Language Roundtable (ILR) oral interview (Lowe 1982), is a test of speaking consisting of (1) a set of elicitation procedures, including a sequence of activities and sets of question types and topics; and (2) a measurement scale of language proficiency ranging from a low level of '0' to a high level of '5', on which samples of oral language obtained via the elicitation procedures are rated. Each of the six scale levels is carefully defined by an extensive verbal description. A qualified ILR interviewer might be able to rate an individual's oral proficiency in a given language according to the ILR rating scale, on the basis of several years' informal contact with that individual, and this could constitute a measure of that individual's oral proficiency. This measure could not be considered a test, however, because the rater did not follow the procedures prescribed by the ILR oral interview, and consequently may not have based her ratings on the kinds of specific language performance that are obtained in conducting an ILR oral interview.

I believe this distinction is an important one, since it reflects the primary justification for the use of language tests and has implications for how we design, develop, and use them. If we could count on being able to measure a given aspect of language ability on the basis of *any* sample of language use, however obtained, there would be no need to design language tests. However, it is precisely because any given sample of language will not necessarily enable the test user to make inferences about a given ability that we need language tests. That is, the inferences and uses we make of language test scores depend upon the sample of language use obtained. Language tests can thus provide the means for more carefully focusing on the specific language abilities that are of interest. As such, they could be viewed as supplemental to other methods of measurement. Given the limitations on measurement discussed below (pp. 30–40), and the potentially large effect of elicitation procedures on test performance, however, language tests can more appropriately be viewed as the best means of assuring that the sample of language obtained is sufficient for the intended measurement purposes, even if we are interested in very general or global abilities. That is, carefully designed elicitation procedures such as those of the ILR oral interview, those for measuring writing ability described by Jacobs *et al.* (1981), or those of multiple-choice tests such as the *Test of English as a Foreign Language* (TOEFL), provide the best assurance that scores from language tests will be reliable, meaningful, and useful.[4]

While measurement is frequently based on the naturalistic observation of behavior over a period of time, such as in teacher rankings or grades, such naturalistic observations might not include samples of behavior that manifest specific abilities or attributes. Thus a rating based on a collection of personal letters, for example, might not provide any indication of an individual's ability to write effective argumentative editorials for a news magazine. Likewise, a teacher's rating of a student's language ability based on informal interactive social language use may not be a very good indicator of how well that student can use language to perform various 'cognitive/academic' language functions (Cummins 1980a). This is not to imply that other measures are less valuable than tests, but to make the point that the value of tests lies in their capability for eliciting the specific kinds of behavior that the test user can interpret as evidence of the attributes or abilities which are of interest.

Evaluation

Evaluation can be defined as the systematic gathering of information for the purpose of making decisions (Weiss 1972).[5] The probability of making the correct decision in any given situation is a function not only of the ability of the decision maker, but also of the quality of the information upon which the decision is based. Everything else being equal, the more reliable and relevant the information, the better the likelihood of making the correct decision. Few of us, for example, would base educational decisions on hearsay or rumor, since we would not generally consider these to be reliable sources of information. Similarly, we frequently attempt to screen out information, such as sex and ethnicity, that we believe to be irrelevant to a particular decision. One aspect of evaluation, therefore, is the collection of reliable and relevant information. This information need not be, indeed seldom is, exclusively quantitative. Verbal descriptions, ranging from performance profiles to letters of reference, as well as overall impressions, can provide important information for evaluating individuals, as can measures, such as ratings and test scores.

Evaluation, therefore, does not necessarily entail testing. By the same token, tests in and of themselves are not evaluative. Tests are often used for pedagogical purposes, either as a means of motivating students to study, or as a means of reviewing material taught, in which case no evaluative decision is made on the basis of the test results. Tests may also be used for purely descriptive purposes. It is

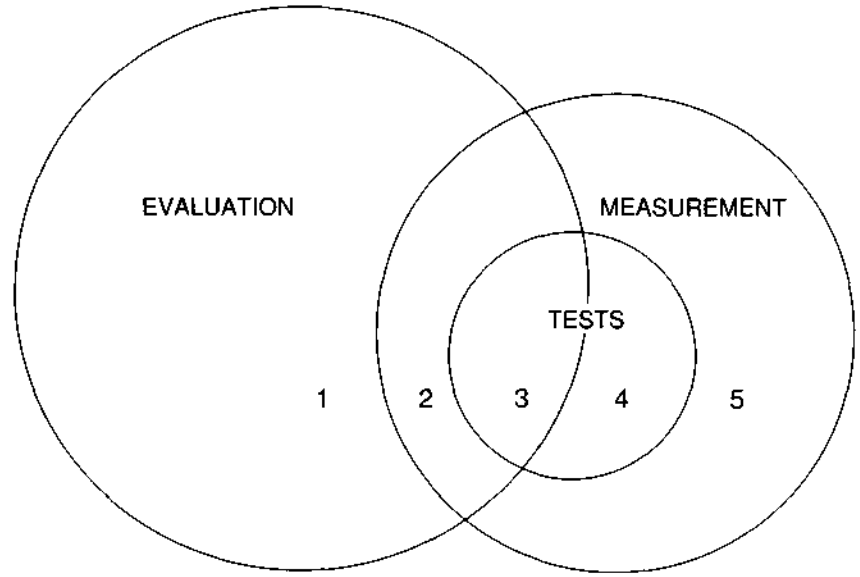

Figure 2.1 Relationships among measurement, tests, and evaluation

only when the results of tests are used as a basis for making a decision that evaluation is involved. Again, this may seem a moot point, but it places the burden for much of the stigma that surrounds testing squarely upon the test user, rather than on the test itself. Since by far the majority of tests are used for the purpose of making decisions about individuals, I believe it is important to distinguish the information-providing function of measurement from the decision-making function of evaluation.

The relationships among measurement, tests, and evaluation are illustrated in Figure 2.1. An example of evaluation that does not involve either tests or measures (area '1') is the use of qualitative descriptions of student performance for diagnosing learning problems. An example of a *non-test measure* for evaluation (area '2') is a teacher ranking used for assigning grades, while an example of a *test* used for purposes of evaluation (area '3') is the use of an achievement test to determine student progress. The most common non-evaluative uses of tests and measures are for research purposes. An example of tests that are not used for evaluation (area '4') is the use of a proficiency test as a criterion in second language acquisition research. Finally, assigning code numbers to subjects in second language research according to native language is an example of a *non-test*

measure that is not used for evaluation (area '5'). In summary, then, not all measures are tests, not all tests are evaluative, and not all evaluation involves either measurement or tests.

Essential measurement qualities

If we are to interpret the score on a given test as an indicator of an individual's ability, that score must be both reliable and valid. These qualities are thus essential to the interpretation and use of measures of language abilities, and they are the primary qualities to be considered in developing and using tests.

Reliability

Reliability is a quality of test *scores*, and a perfectly reliable score, or measure, would be one which is free from errors of measurement (American Psychological Association 1985). There are many factors other than the ability being measured that can affect performance on tests, and that constitute sources of measurement error. Individuals' performance may be affected by differences in testing conditions, fatigue, and anxiety, and they may thus obtain scores that are inconsistent from one occasion to the next. If, for example, a student receives a low score on a test one day and a high score on the same test two days later, the test does not yield consistent results, and the scores cannot be considered reliable indicators of the individual's ability. Or suppose two raters gave widely different ratings to the same writing sample. In the absence of any other information, we have no basis for deciding which rating to use, and consequently may regard both as unreliable. Reliability thus has to do with the consistency of measures across different times, test forms, raters, and other characteristics of the measurement context.

In any testing situation, there are likely to be several different sources of measurement error, so that the primary concerns in examining the reliability of test scores are first, to identify the different sources of error, and then to use the appropriate empirical procedures for estimating the effect of these sources of error on test scores. The identification of potential sources of error involves making judgments based on an adequate theory of sources of error. Determining how much these sources of error affect test scores, on the other hand, is a matter of empirical research. The different approaches to defining and empirically investigating reliability will be discussed in detail in Chapter 6.

Validity

The most important quality of test interpretation or use is validity, or the extent to which the inferences or decisions we make on the basis of test scores are *meaningful*, *appropriate*, and *useful* (American Psychological Association 1985). In order for a test score to be a meaningful indicator of a particular individual's ability, we must be sure it measures that ability and very little else. Thus, in examining the meaningfulness of test scores, we are concerned with demonstrating that they are not unduly affected by factors other than the ability being tested. If test scores are strongly affected by errors of measurement, they will not be meaningful, and cannot, therefore, provide the basis for valid interpretation or use. A test score that is not reliable, therefore, cannot be valid. If test scores are affected by abilities other than the one we want to measure, they will not be meaningful indicators of that particular ability. If, for example, we ask students to listen to a lecture and then to write a short essay based on that lecture, the essays they write will be affected by both their writing ability and their ability to comprehend the lecture. Ratings of their essays, therefore, might not be valid measures of their writing ability.

In examining validity, we must also be concerned with the appropriateness and usefulness of the test score for a given purpose. A score derived from a test developed to measure the language abilities of monolingual elementary school children, for example, might not be appropriate for determining the second language proficiency of bilingual children of the same ages and grade levels. To use such a test for this latter purpose, therefore, would be highly questionable (and potentially illegal). Similarly, scores from a test designed to provide information about an individual's vocabulary knowledge might not be particularly useful for placing students in a writing program.

While reliability is a quality of test scores themselves, validity is a quality of test interpretation and use. As with reliability, the investigation of validity is both a matter of judgment and of empirical research, and involves gathering evidence and appraising the values and social consequences that justify specific interpretations or uses of test scores. There are many types of evidence that can be presented to support the validity of a given test interpretation or use, and hence many ways of investigating validity. Different types of evidence that are relevant to the investigation of validity and approaches to collecting this evidence are discussed in Chapter 7.

Reliability and validity are both essential to the use of tests.

Neither, however, is a quality of tests themselves; reliability is a quality of test scores, while validity is a quality of the interpretations or uses that are made of test scores. Furthermore, neither is absolute, in that we can never attain perfectly error-free measures in actual practice, and the appropriateness of a particular use of a test score will depend upon many factors outside the test itself. Determining what degree of relative reliability or validity is required for a particular test context thus involves a value judgment on the part of the test user.

Properties of measurement scales

If we want to measure an attribute or ability of an individual, we need to determine what set of numbers will provide the best measurement. When we measure the loudness of someone's voice, for example, we use decibels, but when we measure temperature, we use degrees Centigrade or Fahrenheit. The sets of numbers used for measurement must be appropriate to the ability or attribute measured, and the different ways of organizing these sets of numbers constitute *scales of measurement*.

Unlike physical attributes, such as height, weight, voice pitch, and temperature, we cannot directly observe intrinsic attributes or abilities, and we therefore must establish our measurement scales by definition, rather than by direct comparison. The scales we define can be distinguished in terms of four properties. A measure has the property of *distinctiveness* if different numbers are assigned to persons with different values on the attribute, and is *ordered in magnitude* if larger numbers indicate larger amounts of the attribute. If equal differences between ability levels are indicated by equal differences in numbers, the measure has *equal intervals*, and if a value of zero indicates the absence of the attribute, the measure has an *absolute zero point*.

Ideally, we would like the scales we use to have all these properties, since each property represents a different type of information, and the more information our scale includes, the more useful it will be for measurement. However, because of the nature of the abilities we wish to measure, as well as the limitations on defining and observing the behavior that we believe to be indicative of those abilities, we are not able to use scales that possess all four properties for measuring every ability. That is, not every attribute we want to measure, or quantify, fits on the same scale, and not every procedure we use for observing and quantifying behavior yields the same scale, so that it is

necessary to use different scales of measurement, according to the characteristics of the attribute we wish to measure and the type of measurement procedure we use. Ratings, for example, might be considered the most appropriate way to quantify observations of speech from an oral interview, while we might believe that the number of items answered correctly on a multiple-choice test is the best way to measure knowledge of grammar. These abilities are different, as are the measurement procedures used, and consequently, the scales they yield have different properties. The way we interpret and use scores from our measures is determined, to a large extent, by the properties that characterize the measurement scales we use, and it is thus essential for both the development and the use of language tests to understand these properties and the different measurement scales they define. Measurement specialists have defined four types of measurement scales – *nominal*, *ordinal*, *interval*, and *ratio* – according to how many of these four properties they possess.[6]

Nominal scale

As its name suggests, a nominal scale comprises numbers that are used to 'name' the classes or categories of a given attribute. That is, we can use numbers as a shorthand code for identifying different categories. If we quantified the attribute 'native language', for example, we would have a nominal scale. We could assign different code numbers to individuals with different native language backgrounds, (for example, Amharic = 1, Arabic = 2, Bengali = 3, Chinese = 4, etc.) and thus create a nominal scale for this attribute. The numbers we assign are arbitrary, since it makes no difference what number we assign to what category, so long as each category has a unique number. The distinguishing characteristic of a nominal scale is that while the categories to which we assign numbers are distinct, they are *not ordered* with respect to each other. In the example above, although '1' (Amharic) is *not equal to* '2' (Arabic), it is neither greater than nor less than '2'. Nominal scales thus possess the property of distinctiveness. Because they quantify categories, nominal scales are also sometimes referred to as 'categorical' scales. A special case of a nominal scale is a *dichotomous scale*, in which the attribute has only two categories, such as 'sex' (male and female), or 'status of answer' (right and wrong) on some types of tests.