Oxford Handbooks
for Language Teachers

# Designing and Analyzing Language Tests

Nathan T. Carr

CD-ROM INCLUDED

OXFORD

# CONTENTS

**PART I**

**PART II**

# PREFACE

This book would never have been started, much less finished, without the help and encouragement of a number of people. In particular, I am grateful to Julia Sallabank, who originally encouraged the project and oversaw its initiation at OUP; Julia Bell, the editor who very supportively oversaw the process of turning a manuscript into a finished product; two anonymous reviewers, who provided a number of helpful suggestions; Lynda Taylor, my content editor, who challenged me on points large and small, and really helped make this book much stronger and clearer; James Greenan, the copy editor, who helped polish the text, providing many valuable suggestions on points throughout, in a way that only an informed nonspecialist could; and Publishing Manager Catherine Kneafsey and Editor Natasha Forrest, who were also of assistance during the writing process. Of course, the blame for any remaining errors, omissions, and infelicities rests with me.

I owe a great debt of thanks to many others, for encouragement, advice, error spotting, and many other things. In particular, these included (in alphabetical order): Ben Bangs, Kyle Crocco, Chad Davison, Jan Eyring, Juan Carlos Gallego, Peter Groves, Jinkyung Stephanie Kim, Antony Kunnan, Nancy Lay, Ashley Kyu-Youn Lee, Lorena Llosa, Michelle Luster, Mary Ann Lyman-Hager, Teruo Masuda, Gary Ockey, Rene Palafox, Amparo Pedroza, Vanessa Russell, M. Trevor Shanklin, Simeon Slovacek, Hyun-Jung Suk, Debbie Thiercof, Xiaoming Xi, and Cheryl Zimmerman. Thanks also to the other students in my language testing course at California State University, Fullerton, and the participants in two summer workshops at the San Diego State University Language Acquisition Resource Center, who provided the impetus for developing the materials that developed into this book, and gave extensive feedback in class as things evolved.

I also wish to extend my appreciation to the American Council on the Teaching of Foreign Languages, the UCLA ESL Service Courses Program, and the University of Cambridge Press for their gracious permission to reproduce material here.

Thanks are also due to Lyle Bachman, not only for teaching me about language assessment, but training me in how to think about and approach it as well; and to Dilin Liu, for advising me to specialize in testing, rather than some other area of applied linguistics.

Finally, I wish to thank my family for their encouragement and support throughout this process. In particular, my wife Eva has waited while her husband practically disappeared to sit at the computer writing. She has endured far too many days, nights, and weekends without my company, as our dog Odo would wander back and forth to check on each of us, and as the writing and editing process dragged on. And on. And then on some more. This book would not have been possible without her love, support, patience, and understanding.

# LIST OF ACRONYMS AND ABBREVIATIONS

| | |
|---|---|
| $\alpha$ | Cronbach's alpha |
| $\lambda$ | cut score |
| $\Phi$ | index of dependability |
| $\phi$ | item phi; phi correlation coefficient |
| $\rho$ | Spearman rho |
| $\Phi(\lambda)$ | Phi(lambda) |
| AYL | assessing young learners |
| B, B-index | Brennan's B-index |
| B* | B-index for polytomous data |
| CAS | computer-automated scoring |
| CAT | computer-adaptive testing |
| CBT | computer-based testing |
| $\mathrm{CI}_{\mathrm{CRT}}$ | criterion-referenced confidence interval |
| CRT CI | criterion-referenced confidence interval |
| CRT | criterion-referenced testing; criterion-referenced test |
| DI | difference index |
| DI* | difference index for polytomous data |
| G theory | generalizability theory |
| $\mathrm{ID}_{\mathrm{UL}}$ | item discrimination |
| $\mathrm{ID}^{*}_{\mathrm{UL}}$ | item discrimination index for polytomous data |
| IF | item facility |
| IF* | item facility for polytomous data |
| IRT | item response theory |
| M | mean |
| Mdn | median |
| Mo | mode |
| NRT | norm-referenced testing; norm-referenced test |
| $p$ | item facility |
| P&P testing | paper-and-pencil testing |

| | |
|---|---|
| $pb(r)$ | point-biserial item discrimination index or correlation coefficient |
| $Q$ | semi-interquartile range |
| $r$ | Pearson product-moment correlation coefficient (usually refered to as "Pearson *r*") |
| $r_{\text{item-total}}$ | correlation between a polytomous item and total score (a Pearson *r*) |
| $r_{\text{item-mastery}}$ | correlation between a polytomous item and test mastery/non-mastery |
| $r(pb)$ | point-biserial item discrimination index or correlation coefficient |
| $r_{\text{p-bis}}$ | point-biserial item discrimination index or correlation coefficient |
| $r_{\text{pbis}}$ | point-biserial item discrimination index or correlation coefficient |
| $s$ | standard deviation |
| $s^2$ | variance |
| SD | standard deviation |
| SEM | standard error of measurement |
| VPA | verbal protocol analysis |
| WBT | web-based testing |

# INTRODUCTION

Language teachers often have to develop or choose tests or other assessments, whether for their whole school or their own classes, but often they do so with little or no training in what to do. Certainly, this was true for me the first time I was put in charge of testing for a language program, and I remember having little or no idea where to start. This book is intended to give teachers some idea of where to start, and to help them make better-informed decisions. The tests that they use should then be ones they can feel happier about using—there are few things as disheartening as working hard on a project, and doing the best job one can, but at the same time being sure there were important things done wrong or left undone, and having no idea what they were.

## Who This Book Is For

The book is intended for two groups: students in introductory language testing courses (and naturally their instructors), and language teachers who need to create tests for themselves or their school or program, or who have test results that need analyzing. I have tried to make it something that will be useful for both groups, a textbook for students like the ones I teach in my own testing course, and a handbook or guide for teachers who may have no training in language assessment but who nonetheless have to make tests or make sense of testing data.

## The Aims of This Book

The goal of the book is to help current and future teachers learn to develop and analyze their own tests. It should also be useful for helping teachers to choose from among commercially available tests, to decide to develop their own if nothing on the market is suitable, or to adapt existing tests to better meet their needs. It is *not* intended as a guide to important language tests in use around the world, although it should help readers to make better sense of the information test publishers provide, or help them realize when there is something off-base, incomplete, or suspicious-sounding about what is being claimed. By the end of the book, readers will be better placed to make informed decisions about assessment programs and the use of testing at their institution.  Expertise in test writing and test analysis is generally achieved through years of practice and experience, but this book will provide pre-service and in-service teachers with a valuable introduction to the tools necessary for these tasks, as well as useful advice for those interested in learning more on various topics.

## The Organization of This Book

This book is organized into two parts, conveniently labeled "Part I" and "Part II." Part I addresses fundamental concepts of language assessment, explains how to plan, write, and administer tests, provides a conceptual overview of how to score and analyze test results, and gives an overview of test validation procedures, as well as a basic orientation to several more advanced topics in language testing.

Part II explains in detail the quantitative procedures that should be used to describe test results, identify problematic test questions, estimate measurement error, and so forth. Each of these chapters begins with an explanation of the concepts, and is followed by one or two practice worksheets. Each worksheet is accompanied by detailed directions on how to complete it using Microsoft Excel. The DVD inside the back cover of this book contains the Excel files used in the worksheets, a completed version of each file showing the correct formulas, and video tutorials demonstrating how to do the worksheets. The tutorials use Flash, and should be viewable in any Flash-enabled web browser, although they have only been tested in Mozilla Firefox, Google Chrome, and Microsoft Internet Explorer. Each tutorial consists of a recording of my computer screen as I work through a worksheet, and includes my step-by-step narration of the process.

For a detailed listing of the topics covered, please see the Table of Contents.

## How to Use This Book

My practice when using this book with my own class, as part of a 15-week, 45-hour introductory course on language testing, is to cover two chapters a week, one from Part I and another from Part II. I spend two weeks apiece on Chapters 15 and 18, however, since each has two worksheets. Doing this gives the students time to move through the statistical portion of the book at a pace of one lesson per week, rather than two per week concentrated at the end of the course. Given that many language teachers have a certain amount of math anxiety, spreading the quantitative portions out, rather than compressing them, strikes me as less stressful for both students and instructors. This approach of interleaving the two parts also allows midterm exams, if they are used in a class, to assess both conceptual and quantitative aspects of the text, and it gives those students who are having initial difficulty with the math more time to catch up.

## Why Excel?

Some readers may wonder why this book uses Excel for the quantitative chapters. In some cases, the question is asked in the sense of "Why do we need to do math when we are testing language?" To that, I would answer that if we want to say how well students did on a test, we will not be able to speak in more than vague generalities—unless we can use descriptive statistics (Chapter 12), or graph their performance (Chapter 13). Furthermore, we will not know which test questions are doing a good job ("good job" being defined later in the book; see Chapters 5, 15, and 16), which incorrect answers to multiple-choice questions are working and which are not (Chapter 17), or how consistent scores are on a test (Chapters 6, 18, and 19). We will not be able to say how closely related scores were on two different parts of a test, or on two different tests (Chapter 14). And of course, if we want to average our students' grades at the end of a course (Chapter 11), the process is far neater, faster, and less painful with Excel than with paper, pencil, and a calculator or abacus. I would also point out that Excel is not a program for people who enjoy doing math, but rather a program for people who need to do some math but would rather tell a computer to do it for them.

Other readers will answer "Yes, yes, but why not something more powerful, such as" (insert the name of your favorite statistics package here). For more advanced statistical procedures, I certainly would not recommend Excel, since it is not really a statistics program. For the analyses covered in this book, however, it is quite adequate, and even preferable in some regards. First and foremost, nearly everyone around the world with access to a computer has access to it, and it is far cheaper than most statistics programs. Furthermore, most statistics packages cannot do certain types of item analyses (particularly calculating the upper-lower item discrimination, B-index, or distractor point-biserials) without programming or other complex procedures. As a program for entering data, Excel is at least as flexible as anything else. Language teachers will probably find Excel more approachable than most statistics packages, and I would further argue that entering the formulas for themselves will help readers better understand what they are doing. And finally, figures generated in Excel can be reformatted rather easily, and are easily pasted into other applications such as Microsoft Word or Powerpoint.

## Using the Glossary

When any of us first begin studying language testing, there is a great deal of new terminology to be learned. Even when the concepts are simple, the number of new terms being bandied about can be confusing at times. I have therefore included a glossary in this book, to make it faster and easier to look up terms that have been forgotten from earlier chapters, and to make studying easier for those readers who are using this as the textbook in a class on language assessment. At over 275 entries, including abbreviations, acronyms, and Greek letters, it is rather comprehensive, and covers all the terms introduced in bold throughout the book.

## Notes on Terminology

Finally, for the sake of clarification, I should point out that I will often treat the terms *test* and *assessment* as almost interchangeable in this book. "Assessment" is a more superordinate term that includes tests, as well as any other tool (e.g., portfolios) used to make decisions about students' levels of language ability. Tests, on the other hand, are assessments that tend to be somewhat formal, and are often used for making high-stakes decisions—in terms of the grades assigned at the end of courses, if nothing else. Many people have additional assumptions about what tests are and what they look like, but as this book will show, many of those assumptions are based on what people have encountered in the past, and should not necessarily be taken as guidelines for good testing practice. Similarly, I will use the terms *examinee* and *test taker* interchangeably. I will sometimes also use *student*, particularly when the context is clearly related to assessment in the context of a class, language program, or school.

# PART I

# INTRODUCTION

Part I of this book deals with fundamental testing concepts, including consideration of the reasons for testing, how we look at tests and test scores, and some of the more common types of test tasks, before moving on to a discussion of the process of planning and creating tests. It presents conceptual overviews of the quantitative topics of analyzing items (test questions) and the consistency of scoring, but postpones the relevant calculations until Part II of the book. It then discusses the procedures for creating scoring rubrics or rating scales for speaking and listening tests, procedures for test validation and administration, and concludes with brief overviews of several additional topics that, while important in language testing research and practice, require greater coverage than is possible in an introductory textbook.

Each chapter in Part I includes recommendations for further reading on the topics covered, as well as discussion questions asking readers to relate the content of the chapter to teaching and testing situations with which they are familiar—whether as teachers or learners of a language.

# 1 WHAT ARE WE TESTING AND WHY?

## Introduction

A key assumption made in this book is that tests are tools. We will begin by examining how this affects the ways in which we view tests, which leads to consideration of some common test purposes and types, followed by several contrasting paradigms, or ways of looking at tests. The question of what we want tests to tell us comes next, along with a discussion of how well a test needs to perform its appointed task and how we can judge this. The chapter concludes with an explanation of why graphs and descriptive statistics provide a crucial means of looking at test results.

## Tests as Tools

One of the most important things to keep in mind when making or using language tests is that tests and other assessments are tools. We want to use a test or assessment for a particular reason, to do a certain job, not "just because." We should have in mind what that reason is, and who is likely to be taking the test, before we start planning the test—let alone before we start writing it. Almost without fail, the reason for giving the test will have something to do with making decisions about students, or other people (for example, prospective students, prospective employees, or people wanting to have their language ability certified for some purpose). These decisions, naturally, should inform the way that we design our tests (Mislevy 2007).Keeping in mind that a test is a tool can do a lot to clarify our thinking about how to use it. A particular tool is better for some tasks than for others, as anyone who has ever used pliers to remove a screw can understand. Similarly, a certain test might work quite well for one purpose, but not so well for something else. Some tools are poorly made, and are not useful for much of anything; so are some tests, particularly those that are random collections of questions thrown together without any planning. Likewise, some tools are well made, but are highly specialized; in the same way, a given test might be intended for a particular purpose, such as assessing the English-speaking ability of air traffic controllers, and it might do a wonderful job performing that task, but it might not be a good indicator of a doctor's ability to converse with nurses and patients.

Often, there may be several options available for a tool, some high-priced and some cheap, but one of the cheaper alternatives may do the job quite well enough, and while the more expensive options might work even better, they may not be better *enough* to justify the extra expense. Finally, to draw the tool analogy to a close, we should always keep in mind that nobody asks whether someone has a good tool that they can borrow. If someone needs a hammer, they ask for one, not for a screwdriver or wrench! In spite of this, though, it is an all-too-common occurrence for a teacher to ask colleagues if they know any good tests that can be used.

Keeping this firmly in mind, we will next consider some of the purposes we use tests for, and some of the ways we look at test results.

## Test Purposes and Types

As Brown (1995) points out, language tests are normally used to help make decisions, and there are a number of types of decisions that they can be used for. We generally refer to tests, in fact, by the type of decision they are used to make. I think it is useful to divide these test and decision types into two broad categories: those that are closely related to a teaching or learning curriculum, and those that are not. I use this distinction because curriculum-related tests all have a specific domain—the curriculum—to which we can refer when planning and writing these tests. In contrast, when a test is not based on a particular curriculum, we have the burden or freedom (depending on one's point of view) of deciding what specifically it *should* be based on.

These types of tests are summarized in Table 1.1. Brief consideration, of course, will show that many tests are used for more than one purpose; I will refer to several common types of overlap in the following discussion. This is not necessarily problematic.

Furthermore, as will become evident shortly, the dividing line between one type of test and another is not always as clear and sharp as we might pretend. Nevertheless, there are several clearly identifiable types of decisions that are informed by testing, for which some sort of classification system is useful. Because the actual use of a test may change from what was originally planned, it is important to think in terms of types of *decisions* more so than types of tests *per se*; however, it is common in actual usage to *refer* to types of tests as a convenient shorthand.

| **Curriculum-related decisions** | **Other decision types** |
|---|---|
| Admission (sometimes including screening) | Proficiency |
| Placement | Screening |
| Diagnostic | |
| Progress | |
| Achievement | |

*Table 1.1 Test Purposes and Types*

## *Curriculum-Related Tests*

The first type of curriculum-related test that a new student might encounter is an **admission test**, which is used to decide whether a student should be admitted to the program at all; this could of course be viewed as a screening test for a language program (see below), illustrating that, as noted earlier, the lines between categories can often be rather fuzzy. A related type of test is a **placement test**, which is used to decide at which level in the language program a student should study. The student then gets "placed" into that level—hence the name. In many cases, a single test might be used for both purposes: to decide whether a student's language ability is adequate for even the lowest level in the program (admission decisions), and if they pass that threshold, to decide which level is most appropriate for them (placement decisions).

**Diagnostic tests** are used to identify learners' areas of strength and weakness. Sometimes diagnostic information is obtained from placement (or admissions) tests, but sometimes diagnostic tests are administered separately once students have already been placed into the appropriate levels. Some language programs also use diagnostic tests to confirm that students were placed accurately. This can be a good idea, especially if a program is not highly confident in its placement procedures, but it is debatable whether this is actually a diagnostic purpose *per se*. Diagnostic information can be used to help teachers plan what points to cover in class, to help them identify what areas a student may need extra help with, or to help students know which areas they need to focus on in their learning.

Once students are placed appropriately, teachers may wish to find out whether or how well their students are learning what is being taught. **Progress tests** assess how well students are doing in terms of mastering course content and meeting course objectives. This is done from the point of view that the learning is still ongoing—that is, that students are not expected to have mastered the material yet. Many progress decisions in the classroom do not involve testing, however, but are made informally, in the midst of teaching (see, for example, Leung 2004). This is often referred to as monitoring, or "just paying attention," and is assumed to be a fundamental part of teaching, but this does not make it any less a form of assessment. More formally, we often refer to smaller progress assessments as quizzes. However, to the extent that we are using these assessments—quizzes, tests, or whatever—to grade students, we are assessing something other than progress. **Achievement tests** are those that are used to identify how well students have met course objectives or mastered course content. To a large extent, the question of whether a particular test or quiz is an achievement or progress test depends upon how it is being used. To the extent that the test is used to make decisions about what or how fast to teach, it is a progress test, and to the extent that it is used to make decisions about how individual students have learned what they were supposed to, it is an achievement test.

For example, imagine that a test is given in the middle of a course. It is used to assign grades for how well students have learned the material in the first half of the course, but it is also used by the teacher to decide whether any of those points need to be reviewed in class. In such a case, the test is both a progress and an achievement test. As a second example, consider a test given at the very end of a course. This test is used to assign grades to students—to make decisions about how much learning they have achieved in the course—so it is purely an achievement test. In considering whether a test is actually serving as an assessment of progress, achievement, or both—regardless of what it is being *called* by a teacher or program—the key is to think in terms of the type(s) of decisions being made. This is especially important when the actual use of a test has changed from what was intended when it was originally designed.

Moving beyond the level of an individual course, achievement tests can also be used at the level of the school or language program for decisions about whether to promote students to the next level or tier of levels, or for program exit or graduation decisions. Often, of course, practicality dictates that achievement testing for such purposes be combined with end-of-course achievement testing.

Finally, there are two additional types of test-based decisions that closely relate to language curricula and programs, but which do not involve their "own" types of tests. The first involves program evaluation—one source of evidence to use when evaluating a program's effectiveness is tests. While we may want to consider the results of placement tests—and how good a job of placing students they seem to be doing—we may also want to examine achievement test results. In particular, if achievement tests are used at the end of a course, or for graduation, and if these tests are clearly tied to the goals and objectives (Brown 1995) of the course or program, then student performance on those tests should tell us something about how well the program is working.

A second type of curriculum-related decision that tests can help with involves the curriculum planning process. When trying to identify the needs of learners, or—in the case of a new program—prospective learners, we may wish to give a test to the students or potential students to see what they already know and what they still need to learn. Any of the types of tests just discussed might be used for this, although diagnostic, placement, and achievement tests are probably the most likely. One other type of test that might be used, however, is proficiency testing, which—unlike the types just described—is *not* generally tied to a particular language program.

## *Other Types of Tests*

There are two closely related test types that are not usually associated with any of the aspects of a language curriculum, but with which teachers should be familiar nonetheless. The first and most important of these is **proficiency tests**, which assess an examinee's level of language ability, but do so without respect to a particular curriculum. Typically, this involves assessing more than one narrow aspect of language ability; for example, one well-known proficiency test, the Test of English for International Communication (TOEIC; Educational Testing Service 2009b) has long included an assessment of both listening and reading ability, and now includes an optional test of speaking and writing. Similarly, the five Cambridge Main Suite exams all include assessments of reading, writing, speaking, and listening, as well as "use of English" at the more advanced levels (University of Cambridge ESOL Examinations 2008).

When proficiency tests are used to make selection decisions—most commonly about whether someone is sufficiently proficient in the target language to be qualified for a particular job—they are called **screening tests**. Technically, the admission tests discussed above are a type of screening test, but in an academic context. We will limit the use of the term, however, to those tests that involve non-academic selection decisions.

## Ways of Looking at Tests

Besides classifying tests by the types of decisions they are used to inform, we can also view them in several other ways. These involve considering tests in terms of frameworks for interpreting results, the things that examinees have to do during the test, and the ways that the tests are scored.

### *Norm-Referenced and Criterion-Referenced Testing*

One major way in which test results can be interpreted from different perspectives involves the distinction between norm- and criterion-referenced testing, two different frames of reference that we can use to interpret test scores. As Thorndike and Hagen (1969) point out, a test score, especially just the number of questions answered correctly, "taken by itself, has no meaning. It gets meaning only by comparison with some reference" (Thorndike and Hagen: 241). That comparison may be with other students, or it might be with some pre-established standard or criterion, and the difference between norm- and criterion-referenced tests derives from which of these types of criterion is being used.

**Norm-referenced tests (NRTs)** are tests on which an examinee's results are interpreted by comparing them to how well others did on the test. NRT scores are often reported in terms of test takers' **percentile scores**, that is, the percentage of other examinees who scored below them. (Naturally, percentiles are most commonly used in large-scale testing; otherwise, it does not make much sense to divide test takers into 100 groups!). Those others may be all the other examinees who took the test, or, in the context of large-scale testing, they may be the **norming sample**—a representative group that took the test before it entered operational use, and whose scores were used for purposes such as estimating item (i.e. test question) difficulty and establishing the correspondence between test scores and percentiles. The norming sample needs to be large enough to ensure that the results are not due to chance—for example, if we administer a test to only 10 people, that is too few for us to make any kind of trustworthy generalizations about test difficulty. In practical terms, this means that most norm-referenced tests have norming samples of several hundred or even several thousand; the number depends in part on how many people are likely to take the test after it becomes operational.

The major drawback of norm-referenced tests is that they tell test users how a particular examinee performed with respect to other examinees, *not* how well that person did in absolute terms. In other words, we do not know how much ability or knowledge they demonstrated, except that it was more or less than a certain percentage of other test takers. That limitation is why criterion-referenced tests are so important, because we usually want to know more about students than that. "About average," "a little below average," and "better than most of the others" by themselves do not tell teachers much about a learner's ability *per se*. On the other hand, **criterion-referenced tests (CRTs)** assess language ability in terms of how much learners know in "absolute" terms, that is, in relation to one or more standards, objectives, or other criteria, and not with respect to how much other learners know. When students take a CRT, we are interested in how much ability or knowledge they are demonstrating with reference to an external standard of performance, rather than with reference to how anyone else performed. CRT scores are generally reported in terms of the percentage correct, *not* percentile. Thus, it is possible for all of the examinees taking a test to pass it on a CRT; in fact, this is generally desirable in criterion-referenced achievement tests, since most teachers hope that all their students have mastered the course content.

Note also that besides being reported in terms of percentage correct, scores may also be reported in terms of a **scoring rubric** or a **rating scale**, particularly in the case of speaking or writing tests. When this is done with a CRT, however, the score bands are not defined in terms of below or above "average" or "most students," but rather in terms of how well the student performed—that is, how much ability he or she demonstrated. A rubric that defined score bands in terms of the "average," "usual," or "most students," for example, would be norm-referenced.

One important feature of CRTs is that they normally include the use of a **cut score**; that is, a particular score is established as a standard for meeting the criterion, for passing or failing, or for being considered to have demonstrated mastery or non-mastery of the material. Frequently, there will be more than one cut score on a given CRT. This is the case, for example, in placement testing: A cut score divides each pair of adjacent levels (for example, a five-level curriculum will require four cut scores). In classroom grading, there are also multiple cut scores; in the American system, for example, the traditional cut scores are 90%, 80%, 70%, and 60%, for assigning the grades of A, B, C, D, and F. In cases where there is plus/minus grading, there are 12 cut scores: A+/A, A/A-, and so on.

Generally speaking, CRTs are most commonly tied to language curricula, and report how much of the curriculum students have mastered, what portions they still need to master, or what level of the curriculum would be most appropriate for them. On the other hand, one common use for NRTs is for proficiency testing, and thus such tests will report test takers' ability in relation to that of others who have taken the test. It is important to keep in mind, however, that while proficiency tests do not *need* to be norm-referenced, it can be easier to develop this type of test, which merely ranks or orders test takers, than a CRT proficiency test, which requires that results be reported in terms of a framework of language ability.

One noteworthy example of a CRT proficiency test is the American Council on the Teaching of Foreign Languages Oral Proficiency Interview (Swender, Breiner-Sanders, Laughlin, Lowe, and Miles 1999), which rates examinees in terms of the ACTFL Speaking Proficiency Guidelines (American Council on the Teaching of Foreign Languages 1999). Rather than numerical scores, the OPI classifies test takers as being at various levels: Novice-Low, Novice-Mid, Novice-High, Intermediate-Low, Intermediate-Mid, and so on. Another example of a CRT proficiency test is the IELTS (International English Language Testing System; University of Cambridge ESOL Examinations 2009), which assigns scores from 1("non user") to 9 ("expert user"), along with brief descriptions of the language ability of test takers at each level.

Similarly, most tests which integrate multiple skills (for example, reading and writing; see below for a discussion of integrated tests) tend to be CRTs, but that is not an automatic thing. In fact, it is impossible to tell merely by looking at a test whether it is a CRT or NRT. The key is not the purpose, or the types of test tasks, but the way in which scores are interpreted—in terms of other people's performance, in terms of the overall amount of knowledge or ability demonstrated, or perhaps even in some combination of the two, since it may be that a test combines features of both a norm-referenced and criterion-referenced approach. When students' performance is interpreted in terms of the class or group average, or in comparison to "the average student," those are norm-referenced interpretations. In particular, when teachers decide that an average score is a C, and that a certain number of points above average is a B, and so on, that test is being interpreted as an NRT, even if the teacher uses the *term* criterion (for example, "I have set the criterion for earning an A at two **standard deviations** above the mean, so only the top two percent of students will receive this grade"). When we make a criterion-referenced interpretation, we are concerned with how well a student did without reference to anyone else's score. This does *not* mean that averages and other **descriptive statistics** are not useful in CRTs, but it means we do not use them in the same way that we would in NRTs. (See Chapter 12 for further discussion of descriptive statistics, why they are used, and how they are interpreted.)

## Summative vs. Formative Assessment

Another way of looking at and using tests and other assessments also involves two **interpretive frameworks**, but these frameworks have more to do with when tests are administered, and the purposes the results are used for. **Summative assessments** are typically given at the end of a unit, course, program, etc., and they provide information about how much students learned. They are therefore closely related to achievement tests, and in fact, most achievement testing is largely summative, and summative testing usually aims to assess learner achievement. On the other hand, **formative assessments** take place while students are still in the process of learning something and are used to monitor how well that learning is progressing (see, for example, Leung 2004). They are therefore closely related to progress assessment, and to the extent that the results of an assessment are used to guide the subsequent teaching and learning process, such assessment is formative. One way to keep the distinction between these two perspectives clear is to remember that summative assessments are used to sum up how well someone did, and formative assessments are used to shape or form what is being taught.

These two assessment types are usually viewed as being a dichotomy, but they are probably better thought of as being the endpoints of a continuum. For example, if a teacher informally assesses how well his students are mastering something being taught that day, without assigning any grades, this is clearly formative, and not summative at all. Similarly, if a student is required to take a test in order to graduate from a program, that is clearly a summative assessment. On the other hand, however, if a teacher gives a quiz over material that has been covered recently, and uses the results to both assign grades and decide whether to review the material further or move on to something else, the quiz is clearly performing both summative and formative functions, respectively. As an additional, learner-centered example, if a course includes a quiz at the end of every lesson, it could be put to multiple uses. If learners use the quiz to decide whether they are satisfied with how much they have learned on the topic, that would be a summative function, while if they use it to decide whether they need to review the material further before moving on, that would be a formative use of the quiz.

As a final note on this topic, a similar distinction is often made in the area of educational evaluation, where programs, courses, and so forth can be evaluated from a summative or formative perspective.

## *Objective vs. Subjective Testing*

Several other ways of looking at tests are fairly common, but nevertheless offer mistaken or inaccurate perspectives. One of these is the false distinction between objective and subjective testing. A so-called **objective test** is one that can be scored objectively, and therefore uses selected-response questions (particularly multiple-choice questions, but sometimes true-false or matching questions as well). A **subjective test**, on the other hand, is one that involves human judgment to score, as in most tests of writing or speaking. As testing researcher Lyle Bachman (Bachman 1990) has pointed out, however, there are several problems with these terms. First of all, we should consider where the questions on an "objective" test come from. Even the most principled expert planning decisions about what to assess and how to do it are somewhat subjective, as are any decisions made in the course of writing the actual test.

For example, imagine the case of teachers creating a final exam for a course on academic reading. Before they start writing the test, they make decisions as to what topics to use for reading passages, how long passages should be, how many passages and questions to include on the test, what types of questions they need (for example, main idea, reading for details, scanning, and inference questions), and how many of each type they want. Then, when they start writing the test, they make decisions about which reading passages to copy, or make decisions throughout the course of writing their own passages. Every question that they write is the product of multiple decisions about its content, intended purpose, and what choices they want to include. No matter how appropriate these decisions are, every one of them is clearly subjective to one degree or another.

On the other hand, "subjective" tests are not necessarily as subjective as their label might suggest. With the use of clearly written scoring rubrics (also referred to as rating scales), and rater training and **norming** using example responses, much of the subjectivity in scoring can be reduced. If appropriate record keeping and statistics are used, it can be monitored and reduced even further. As an example, if a language program uses a placement test that includes a writing section, just having teachers read the writing samples and assign a score based on their individual judgment would be highly subjective. In contrast, if the program had a clear scoring rubric, trained teachers in how to apply it, and kept track of how consistent scoring was for individual teachers and for all the teachers as a group, much of the subjectivity would be taken out of the process. It could still be *argued* that the scoring rubric was subjectively derived, but that argument could probably be countered fairly successfully. Thus, the terms objective and subjective should at most be used to refer to the scoring method used, not the overall test itself, and even then, the scoring is not quite as sharp or fuzzy as those two labels might imply.

The important question is therefore not whether a test is "objective" or "subjective," but where subjectivity and measurement error will come into play, and how that subjectivity and error can best be reduced. Generally speaking, this requires that any decisions about the test—including decisions about its planning, creation, administration, and scoring—be made consciously, that is, in a carefully considered manner. Reflective decisions are more likely to have some principled basis and thus be more defensible than those that are made implicitly or reflexively, without consideration or forethought. Obviously, making all decisions *conscious* decisions requires planning, which will be discussed in greater detail in Chapters 3 and 4. The need to make defensible decisions, and to show that assessments are appropriate for their intended uses, will be discussed later in this chapter, and in greater detail in Chapter 8. Once a test has been constructed, there are ways to estimate the level of measurement error, and to identify test questions that are likely causing a disproportionate share of that error; these topics are the focus of the quantitative chapters of this book and Chapters 5 and 6. Methods of scoring "subjective" (i.e. writing and speaking) tests more consistently are the subject of Chapter 7.

## *Direct vs. Indirect Testing*

A second problem revolves around the distinction between **direct** and **indirect tests**, which is less a false distinction than a misnomer. So-called "direct" tests are those that require examinees to use the ability that is supposed to be being assessed; for example, a writing test that requires test takers to write something, or a speaking test that requires examinees to speak. In contrast, indirect tests are those that attempt to assess one of the so-called (see Savignon 2001) "productive skills" through related tasks that do not require any speaking or writing. Instead, they rely upon tasks that will be easier and/or faster to grade; for example, an indirect test of writing might include a multiple-choice test of grammatical knowledge, error detection, knowledge of the rules of rhetorical organization, and so on. Similarly, an indirect test of speaking might include a test of listening comprehension and/or the ability to select the response that best completes a short dialog. In other words, rather than attempting to assess the ability itself, these tests assess related abilities, in the hope that this will provide an accurate enough estimate of an examinee's ability. Naturally, there is the potential in doing this that the resulting test will be convenient to administer and score, but will provide little or no useful information about someone's writing or speaking ability.

What is problematic about this distinction is that the so-called "direct" tests are themselves indirect (Bachman 1990). Note that it is not the "direct" tests themselves that are at issue, but rather the label they are being given. The problem relates to the distinction between competence and performance; that is, if an assessment requires students to do something, and the resulting performance is then scored, or otherwise evaluated, we can only observe that performance. Fortunately for those of us in the testing business, that performance can generally be taken as an *indication* of the underlying competence, but the performance is not the competence itself. We generally assume that good performance on a speaking or writing test results from high ability levels, but if someone does poorly, the reason(s) may be less clear: Perhaps it is a question of weak language ability, but it may also be a matter of unfamiliarity with the task, or even nervousness, among other things. Other factors could also lead a high-ability test taker to receive a poor score on a performance test. Some examples include unfamiliarity with the subject matter being talked or written about, emotional distress, misunderstanding what sort of response was expected, and a negative personal reaction to or by the person scoring the test. Thus, it is probably more accurate to refer to such tests as being more authentic or more communicative, and as perhaps tending to have greater **construct validity**. Despite the point I am raising here, though, the terms *direct* and *indirect* are still widely used.

Finally, it should be noted that the term **semi-direct tests** is generally used for speaking tests that require the test takers to record their speech rather than talk directly to a human interlocutor. These tests are generally tape-mediated or computer-mediated, as in the case of the TOEFL iBT speaking test (Educational Testing Service 2009a), Computerized Oral Proficiency Instrument (COPI; Malabonga, Kenyon, and Carpenter 2005), and Computer Assisted Screening Tool (CAST; Language Acquisition Resource Center at San Diego State University 2009; Malone 2007). This type of test is clearly quite similar to direct testing, in that it obviously assesses speaking ability by having examinees speak, rather than listen or read. It is the lack of a live interlocutor with whom the test taker can interact reciprocally, though, that distinguishes *semi-direct* from *direct tests*.

## *Discrete-Point vs. Integrated Tests*

Another important distinction between types of tests is the one between discrete-point and integrated assessments. Both approaches have strengths and weaknesses, which means that test designers must give careful thought to the trade-offs involved in choosing one, the other, or both. A **discrete-point** test is one that uses a series of separate, unrelated tasks (usually test questions) to assess one "bit" of language ability at a time. This is typically done with multiple-choice questions, and was long the format used for standardized language tests of reading, listening, grammar, and vocabulary. Although this hardly makes for lifelike language use, this approach to test design does in fact have several redeeming features. For one thing, having each question or task unrelated to the others, aside from the fact that they all assess the same ability, satisfies an important statistical assumption underlying **item response theory** (**IRT**; see, for example, Hambleton, Swaminathan, and Rogers 1991; Embretson and Reise 2000). IRT is a powerful statistical methodology commonly employed in large-scale standardized testing, and is very useful for—among other things—controlling item and test difficulty and estimating examinees' ability levels.

Another main attraction to discrete-point testing is that if a student gets an item wrong, it is presumably because of a lack of ability in a specific area, and not interference from some other thing that is being simultaneously tested—for example, getting a reading question wrong on a discrete-point test cannot stem from a lack of writing ability. Furthermore, because each question is so brief, discrete-point tests also allow the coverage of a large number of points, whether these are topics, situations, vocabulary items, or grammatical structures. They can also be used to assess a wide variety of communicative functions or language use tasks, although how *well* they might assess the ability to perform a particular function or task is open to debate.

Finally, discrete-point tests are useful for testing very specific areas of language, such as the grammar points that have been covered in a course. To continue that example, students taking a discrete-point grammar test will not be penalized for a lack of knowledge of other points that were not taught. Similarly, they cannot use their knowledge of other points to compensate for their lack of mastery of what has been taught; that is, they cannot "write around" or "talk around" their gaps.

Unfortunately, although discrete-point tests offer these benefits, this comes at the price of authentic language use. Very seldom in real life does anyone use language one discrete point at a time—outside language tests and highly structured classroom practice, language use tasks tend to involve the integration of multiple skills, and language users can often use strength in one area of language ability to compensate for weakness in another. Thus, discrete-point tests provide an incomplete picture of what learners can actually *do* with the language. This is the reason for the development of **integrated tests**, which require examinees to use multiple aspects of language ability, typically to perform more life-like tasks. Examples might include taking notes over a listening passage and then writing a summary, or writing something about one or more texts read during the test. Such tests more closely resemble real-life language use tasks, and thus require more communicative language use.

Since authenticity and communicative language use are things that are fundamental to the communicative approach to language teaching, one might wonder why all the language tests used in programs claiming to follow that approach are not integrated. The reason, as it turns out, is that integrated tests create their own set of problems. The first, and perhaps most easily addressed reason, is that integrated tests are more difficult to score than discrete-point multiple-choice questions. This challenge can be dealt with by establishing clear scoring rubrics, and training raters in how to use them. (The development of scoring rubrics will be dealt with in greater detail in Chapter 6.)

A second problem raised by the use of integrated tests is that it is often more difficult to interpret scores that result from them. For example, if test takers score highly on an integrated reading and writing task, we can probably assume that they both read and write well. If they do poorly on the test, though, is it because they are poor readers, poor writers, or both? Without some additional measure of reading and/or writing, we cannot be sure.

Another issue has to do with how broad an integrated test can be in terms of what it is covering. When we give a test, we want it to tell us something about how students will perform *outside* the test environment. In a single test, we cannot possibly cover every topic, situation, vocabulary item, rhetorical mode, literary genre, notion, communicative function, language use task, grammatical structure, and so on, that we might find important; as a result, we must be selective and use a representative *sample* from all the areas about which we wish to make claims. Because integrated tasks typically take up as much time as a larger number of discrete-point tasks, they reduce the number of points that can be sampled in a given test.

For example, a teacher might have covered four grammatical structures and vocabulary associated with four topics, and now wishes to assess their students' ability to comprehend them in reading and use them accurately in writing. This teacher plans to use integrated tasks, but only has time to include three tasks on the test. Each task will target one structure and one topic covered in the class. This plan is probably reasonable. On the other hand, if the teacher had covered 10 topics and 10 structures, and only wanted to include one task on the test, that would be rather problematic. Unfortunately, there is no hard-and-fast rule for determining what constitutes an adequate sample on a test, so teachers must have a clear rationale to support any decisions they make, a point that will be discussed later in this chapter, and then in greater detail in Chapter 8.

In deciding between discrete-point and integrated tests, besides the factors already discussed, there is also the matter of what effect the test might have on the teaching and learning process. If teachers are encouraged to use communicative language practice activities both inside and outside the classroom, but then use decontextualized tasks to assess their reading, listening, grammar, and vocabulary separately, this sends a mixed message about the importance of achieving communicative competence in the target language. Furthermore, if discrete-point tests are imposed from above, they tend to send an *un*mixed message that communicative practice does not matter. This decreases the desire and motivation of students to achieve communicative competence, since they will be assessed on the basis of something else. It also puts pressure on teachers to focus on developing discrete skills in isolation so as to better prepare students for their tests. On the other hand, the best way to prepare students for integrated tests would probably be to include extensive communicative language practice, both in class activities and as part of homework and other assignments.

As a final point on this topic, it is worth mentioning the idea of using **independent speaking and writing tasks**, which is probably the most common approach to assessing speaking and writing. In these tasks, test takers react to a prompt or interlocutor, but do not have to process significant amounts of written or spoken text—that is, they do not have to comprehend a reading or listening passage in order to respond. This should not be considered a discrete-point approach to testing speaking or writing, however, particularly since multiple aspects of speaking or writing could be assessed in these tasks (for example, vocabulary use, grammatical accuracy, fluency, or pronunciation in a speaking test). It is probably also worth pointing out that there is not always a clear line between integrated and independent speaking and writing tasks—that is, it is probably better to view the two types as the two ends of a continuum, rather than as discrete categories.

## *Performance Assessments: Focus on Task Completion vs. Focus on Language Use*

McNamara (1996: 6) perhaps best explains **performance assessments**, describing them as assessments that require "actual performances of relevant tasks ... rather than more abstract demonstration of knowledge, often by means of paper-and-pencil tests." He further distinguishes two ways of viewing second language performance assessments: the *strong sense* and the *weak sense*. The difference between the two perspectives lies in the criteria used to evaluate the performance. The **strong sense of language performance assessment** is concerned with how well a task is performed, using real-world criteria; the level of linguistic accuracy displayed only matters to the extent that it interferes with or enables task performance, making "adequate second language proficiency ... a necessary but not a sufficient condition for success on the performance task" (McNamara 1996: 43).

In contrast, the **weak sense of language performance assessment** is concerned with the level of the language used in performing the task. The purpose of the task is to elicit a sample of language to be evaluated; performance of the task, as such, is secondary, and if task completion or fulfillment is considered in the scoring, it is typically done with reference to language. In language teaching and assessment, we are generally more concerned with the weak sense of performance assessment; the strong sense is more likely to come up in vocational testing contexts.

## What we Want Tests to Tell us

Now that we have examined decisions that tests can help us make, and several ways of looking at tests and test results, it seems appropriate to discuss what it is that we hope to learn when we administer a test. First and foremost, we assume that a test or other assessment is providing information about one or more **constructs.** A construct is the ability that we want to assess, but which we cannot directly observe—for example, we cannot judge a student's level of reading ability just by looking at them, or by opening up their head and looking inside. We therefore have examinees *do* things, such as answering questions on tests, which provides us with an indirect indication of how much of a particular construct they possess. That is, based on their performance, we make inferences about how much of the construct they possess. In testing in a school or language program, the construct is probably based on the curriculum or syllabus being used, which in turn is (presumably) based on some theoretical model of language ability and its acquisition (see, for example, Bachman and Palmer 1996; Canale and Swain 1980; Purpura 2004). On the other hand, when tests—most notably proficiency tests—are not based on a particular syllabus or curriculum, the construct will be based directly on a theoretical model.

We also expect that what people do when they take a test is going to tell us something about how well they will use language *outside* the test. Our concern may be with how they would perform "out there" in the real world, or it may be with how they would use language in a course or program—although presumably what goes on in the classroom is tied somehow to real-world language use. Bachman and Palmer (1996) have coined the term **target language use (TLU) domain** to refer to the contexts outside the test, whether in the real world or in the classroom, where test takers will use the language. When students take tests, we make generalizations about how they will perform in these contexts—in the TLU domain—based on their performance and scores. Put another way, students take tests and receive scores based on how they did. We then make certain inferences or predictions about their ability to use language outside the test, in the TLU domain. Based on those scores, we make decisions: what level a new student should study in, whether the student has learned enough of the material we have taught, and so on. The process for how this takes place is summarized in Figure 1.1.

Students perform tasks on a test. The tasks will, presumably, provide information about students' ability (i.e. the construct(s) of interest)

↓

Students get scores for how well they perform those tasks

↓

Based on those scores, we make inferences about each student's ability to use the target language (inferences about constructs, as contextualized in the TLU domain)

↓

Based on these beliefs about their ability levels, we make decisions

*Figure 1.1 How Tests Are Used to Make Decisions*

As noted earlier, we can view test performance as indicating something about test takers' language ability—that is, how much of the construct they possess. We must also keep in mind, though, that a performance is a *contextualized* use of language, and therefore also depends on the features of that context (Chapelle 1998). Thus, how well language learners do on a particular test task (for example, answering a reading comprehension question, writing an essay, or taking part in a role-play) is a result of two things: their language ability and other attributes (for example, background knowledge, personal assertiveness, level of concern over detail, and tolerance for ambiguity), and the characteristics of the task (Bachman 2002b). It therefore follows that if we view the test as sampling tasks from the language use context(s) of interest—that is, the ones to which we wish to generalize on the basis of test scores—we must take care to sample broadly enough. Otherwise, any claims that performance on the test can be generalized to performance "out there" in the real world or in the classroom will not hold up to examination. The same thing is true about claims that performance on an achievement test is an indication of how much students have learned in a course. Without adequate sampling of the topics, situations, genres, rhetorical modes, functions, notions, structures, and tasks that were covered in class, it is not possible to claim, in fairness, that the test provides a clear picture of what students have or have not achieved.

As mentioned earlier, there is no simple rule for what level of sampling is "adequate." Obviously, it will not be possible to include examples of every relevant TLU task on a single test (see Figure 1.1). How much is "good enough" depends on the extent to which a test designer can make a convincing case that the coverage is broad enough and representative enough of the TLU domain(s) of interest (i.e. of the non-test language use contexts), and provide any necessary support for that argument. One way to ensure this is by systematically analyzing and modeling the TLU domain, and using the results of this process as the basis for planning and writing the test (Bachman and Palmer 1996; Mislevy and Haertel 2007).

## How Good Is Good Enough?

Obviously, no test is going to be perfect. What matters most is that it should be useful for its intended purpose (Bachman and Palmer 1996), and that it should do its job with fairness (Kunnan 2004). Evaluating the usefulness of a test is best done on a systematic basis. Bachman and Palmer propose doing this through the consideration of several **qualities of usefulness**, which are summarized in Table 1.2. The various qualities are all important to making sure a test is useful for its intended purpose. One or more will often be prioritized in a given situation, but not to the extent that the others are ignored. Test developers need to decide how important each quality is, and set minimally acceptable levels for it. In other words, besides prioritizing, it is important to decide what the lowest level or worst outcome is for each one that one could accept before deciding that the test could not do its job adequately. Chapter 3 will discuss the qualities in greater detail, as well as how to set minimally acceptable levels for each of them. Later on, Chapter

8 will address the question of how to make the argument that the test really does perform its job to a satisfactory degree.

| **Quality** | **Definition** |
|---|---|
| Reliability | consistency of scoring, estimated statistically |
| Authenticity | the degree to which test tasks resemble TLU tasks |
| Construct Validity | the degree to which it is appropriate to interpret a test score as an indicator of the construct (i.e. ability) of interest |
| Impact | effects of the test on people and institutions, including (but not limited to) **washback**—the effect of a test on teaching and learning |
| Practicality | the degree to which there are enough resources to develop and use the test |

*Table 1.2 Bachman and Palmer's Qualities of Usefulness*

Before concluding this discussion of test usefulness, however, it seems appropriate to raise an issue related to the quality of **reliability**, the notion of measurement error. While reliability and measurement error will be the focus of Chapters 6, 18, and 19, it is worth noting here that no test score is a perfect indicator of language ability, as tempting as it may be to pretend otherwise. Rather, it is only an *estimate* of examinee ability, and like any estimate, is subject to a certain margin of error. Reliability, or the consistency of scoring, involves determining how much effect error has on test scores. Error can be caused by a variety of factors involving the test itself, the people taking the test and conditions in which they take it, and the way in which it is scored. While some error is inevitable, the goal of much of this book is to teach readers how to plan, write, and administer tests in such a way as to help minimize this error.

## Why Graphs and Descriptive Statistics Matter

We conclude this chapter with a discussion of a particular way of looking at test results, and why teachers should concern themselves with it: the use of descriptive statistics and graphs. The best place to start might be with the question that many readers are subconsciously (or even consciously!) asking: *Why* do we need to bother with all that? Many teachers, especially in their capacity as novice test developers, may wonder. Experts, professionals, and people involved in large-scale testing need to, of course, but why do classroom teachers need to bother? After all, a fondness for math was probably not the reason most of them chose language teaching as a career!

Carr (2008b) notes a number of reasons why it is worth teachers' while to use graphs (see Chapter 13) and **descriptive statistics** (see Chapter 12). The first reason is probably the least convincing to the statistically skeptical: descriptive statistics are important for helping us decide whether certain statistical tests are appropriate. There are a number of such tests, and although they fall beyond the scope of this book, they are still very important. One common example (known as the *t*-test) is used when we have two sets of test scores (for example, two tests taken by the same class) and we want to know whether the difference between them is small enough to be the result of chance.

Another reason is that these tests can help us choose the correct **correlation coefficient**. Correlation coefficients are used when we want to learn exactly how closely related two sets of numbers are (see Chapter 14). Which correlation coefficient is appropriate will depend on the nature of our data, but we do not *know* the nature of the data without first calculating descriptive statistics.

Descriptive statistics are also important because the same formulas are used in calculating other useful things. In particular, they are useful as part of calculating the reliability of a test (see Chapters 18 and 19), or in determining which test questions have done a good job (see Chapters 15 and 16).

At a more fundamental level, descriptive statistics and visual representations of data are important because they give us basic information about how examinees did when they took a test. They tell us how well most students performed, but also provide information about the rest of the group as well. In particular, they tell us whether test scores were distributed the way we expected, wanted, or even needed. For example, at the end of a unit or course, we expect most of the students to have mastered the material covered. We do not know whether this is the case, though, until we administer some sort of assessment and look at the overall results. The "looking at" process is done with descriptives and graphs, unless one's class is so small (i.e. only a handful of students) that one can "eyeball" the scores to see that they are more-or-less what was expected.

Finally, this also involves a matter of testing ethics, as the International Language Testing Association Code of Ethics (2000) states that information about tests must be communicated both accurately and "in as meaningful a way as possible." Since it is impossible to discuss patterns of test performance in any meaningful way without using numbers, or graphs illustrating numbers, we are rather stuck. For example, graphs can help us see at a glance how many students were assigned to each level on a program's placement test, or how many students received an A, B, C, D, or F on a final exam. Similarly, descriptive statistics can help us make exact comparisons between two groups, as when we want to compare the scores of two classes that took the same test.

The point of all this is not to persuade you that you should necessarily *enjoy* the process of creating graphs and calculating statistics, but that it really does matter. As for whether the average language teacher can really ever learn to do these things properly, the answer is *yes*, something that I hope the second part of this book will prove to you.

## Summary

This chapter began by pointing out that tests are tools, to be used for specific purposes, and then described the most common of those: placement, admission, diagnosis, progress, achievement, proficiency, and screening. It then explored perspectives from which we can view tests, including norm- vs. criterion-referencing; summative vs. formative purposes; so-called objective vs. subjective testing; the arguably misnamed direct, indirect, and semi-direct tests; and discrete-point vs. integrated tests. The chapter subsequently introduced the notion of constructs, and how they are contextualized within a target language use (TLU) domain, before taking up the qualities of test usefulness and the related issue of measurement error. It then addressed the importance of graphs and descriptive statistics as an additional, crucial way of viewing test results.

### *Further Reading*

**American Council on the Teaching of Foreign Languages.** 1999. 'ACTFL proficiency guidelines—speaking' (revised 1999). Retrieved March 29 2008, from http://www.actfl.org/files/public/Guidelinesspeak.pdf.

**Bachman, L. F.** 1990. Chapter 3 'Uses of language tests'. *Fundamental Considerations in Language Testing.* Oxford: Oxford University Press.

**Bachman, L. F.** and **A. S. Palmer.** 1996. Chapter 1 'Objectives and expectations', Chapter 2 'Test usefulness: Qualities of language tests'. *Language Testing in Practice.* Oxford: Oxford University Press.

**Brown, J. D.** 1995. Chapter 4 'Testing'. *The Elements of Language Curriculum: A Systematic Approach to Program Development.* Boston: Heinle and Heinle Publishers.

**Canale, M.** and **M. Swain.** 1980. 'Theoretical bases of communicative approaches to second language teaching and testing'. *Applied Linguistics* 1: 1–47.

**Carr, N. T.** 2008b. 'Using Microsoft Excel to calculate descriptive statistics and create graphs'. *Language Assessment Quarterly* 5 (1): 43–62.

**International Language Testing Association.** 2000. *Code of ethics for ILTA.* Retrieved July 10 2009, from http://www.iltaonline.com/index.php?option=com_content&view=article&id=57&Itemid=47.

**Kunnan, A. J. 2004.** 'Test fairness' in M. Milanovic. and C. J. Weir (eds.). *European Language Testing in a Global Context: Proceedings of the ALTE Barcelona Conference, July 2001: 27–48.* Cambridge: UCLES/Cambridge University Press.

**Leung, C.** 2004. 'Developing formative teacher assessment: Knowledge, practice, and change'. *Language Assessment Quarterly* 1 (1): 19–41.

**McNamara, T.** 1996. Chapter 1 'Second language performance assessment'. *Measuring Second Language Performance.* London: Addison Wesley Longman.

**Purpura, J. E.** 2004. Chapter 3 'The role of grammar in models of communicative language ability', Chapter 4 'Towards a definition of grammatical ability'. *Assessing Grammar.* Cambridge: Cambridge University Press.

University of Cambridge ESOL Examinations. 2009. *Information for Candidates.* Retrieved June 6 2010, from
http://www.ielts.org/PDF/Information_for_Candidates_2009.pdf.

## Discussion Questions

**1** Has most of the language testing you have been part of (either taking or giving tests) been discrete-point or integrated? How has each type affected your language teaching or learning?

**2** Besides the test uses listed here, can you think of any other uses that language tests are put to?

**3** Think of a language test with which you are familiar.
- **a** What was its purpose?
- **b** Was it norm- or criterion-referenced?
- **c** Was it summative or formative?
- **d** Was it discrete-point or integrated?

**4** Consider a language program in which you have studied or taught, and think of a purpose for which a performance assessment would be appropriate in that context. Which would you feel more appropriate for that test: the strong or weak sense of language performance assessment?

**5** Imagine that you are planning a language proficiency test. What construct(s) would you include on your test? What target language use domain(s) would you want to generalize to—i.e. make claims about—on the basis of this test?

**6** Consider the qualities of usefulness listed in Table 1.2. How important would you view each of them as being in the following contexts, and why?
- **a** A progress assessment in a language class
- **b** An end-of-course achievement test in a university language course
- **c** A language proficiency test used by employers to make hiring decisions
- **d** The placement test for a university-level language program