Part One

Background: The Fundamentals

Chapter One

Test Theory

What Is Testing?

There are four related terms that can be somewhat confusing at first: evaluation, assessment, measurement, and testing. These terms are sometimes used interchangeably; however, we think it is useful to make the following distinctions among them:

- *Testing* is the collection of quantitative (numerical) information about the degree to which a competence or ability is present in the test-taker. There are right and wrong answers to the items on a test, whether it be a test comprised of written questions or a performance test requiring the demonstration of a skill. A typical test question might be: "List the six steps in the selling process."
- *Measurement* is the collection of quantitative data to determine the degree of whatever is being measured. There may or may not be right and wrong answers. A measurement inventory such as the *Decision-Making Style Inventory* might be used to determine a preference for using a Systematic style versus a Spontaneous one in making a sale. One style is not "right" and the other "wrong"; the two styles are simply different.
- *Assessment* is systematic information gathering without necessarily making judgments of worth. It may involve the collection of

quantitative or qualitative (narrative) information. For example, by using a series of personality inventories and through interviewing, one might build a profile of "the aggressive salesperson." (Many companies use Assessment Centers as part of their management training and selection process. However, as the results from these centers are usually used to make judgments of worth, they are more properly classed as evaluation devices.)

- *Evaluation* is the process of making judgments regarding the appropriateness of some person, program, process, or product for a specific purpose. Evaluation may or may not involve testing, measurement, or assessment. Most informed judgments of worth, however, would likely require one or more of these data gathering processes. Evaluation decisions may be based on either quantitative or qualitative data; the type of data that is most useful depends entirely on the nature of the evaluation question. An example of an evaluation issue might be, "Does our training department serve the needs of the company?"

Practice

Here are some statements related to these four concepts. See whether you can classify them as issues related to Testing, Measurement, Assessment, or Evaluation:

1. "She was able to install the air conditioner without error during the allotted time."
2. "Personality inventories indicate that our programmers tend to have higher extroversion scores than introversion."
3. "Does the pilot test process we use really tell us anything about how well our instruction works?"
4. "What types of tasks characterize the typical day of a submarine officer?"

Feedback

1. Testing
2. Measurement
3. Evaluation
4. Assessment

What Does a Test Score Mean?

Suppose you had to take an important test. In fact, this test was so important that you had studied intensively for five weeks. Suppose then that, when you went to take the test, the temperature in the room was 45 degrees. After 20 minutes, all you could think of was getting out of the room, never mind taking the test. On the other hand, suppose you had to take a test for which you never studied. By chance a friend dropped by the morning of the test and showed you the answer key. In both situations, the score you receive on the test probably doesn't accurately reflect what you actually know. In the first instance, you may have known more than the test score showed, but the environment was so uncomfortable that you couldn't attend to the test. In the second instance, you probably knew less than the test score showed due now to another type of "environmental" influence.

In either instance, the score you received on the test (your observed score) was a combination of what you really knew (your true score) and those factors that modified your true score (error). The relationship of these score components is the basis for all test theory and is usually expressed by a simple equation:

$$X_o = X_t + X_e$$

where X_o is the observed score, X_t the true score and X_e the error component. It is very important to remember that in test theory "error" doesn't mean a wrong answer. It means the factor that accounts for any mismatch between a test-taker's actual level of knowledge (the true score) and the test score the person receives. Error can make a score higher (as we saw when your friend dropped by) or lower (when it got too cold to concentrate).

The primary purpose of a systematic approach to test design is to reduce the error component so that the observed score and the true score are as nearly identical as possible. All the procedures we will discuss and recommend in this book will be tied to a simple assumption: the primary purpose of test development is the reduction of error. We think of the results of test development like this:

$$X_o = X_t + x_e$$

where error has been reduced to the lowest possible level.

Realistically, there will always be some error in a test score, but careful attention to the principles of test development and administration will help reduce the error component.

Practice

See if you can list at least three situations that could inflate a test-taker's score and three that could reduce the score:

Inflation Factors	Reduction Factors
1. Sees answer key	1. Room too cold
2. ————————	2. ———————-
3. ————————	3. ———————-
4. ————————	4. ———————-

Feedback

Inflation Factors	Reduction Factors
1. Sees answer key	1. Room too cold
2. Looks at someone's answers	2. Test scheduled too early
3. Unauthorized job aid used	3. Noisy heating system in room
4. Answers are cued in test	4. Can't read test directions

Reliability and Validity: A Primer

Reliability and validity are the two most important characteristics of a test. Later on we will explore these topics and provide you with specific statistical techniques for determining these qualities in your tests. For now, we want to provide an overview so that you will see how these ideas serve as standards for our attempts to reduce error in testing.

Reliability

Reliability is the consistency of test scores. There is no such thing as validity without reliability, so we want to begin with this idea.

There are three kinds of reliability that are typically considered in CRT construction:

- equivalence reliability
- test-retest reliability
- inter-rater reliability

Equivalence reliability is consistency of test scores between or among forms. There are several reasons why parallel forms of a test (different questions that measure the same competencies) might be desirable, for example, pretest/posttest comparisons. Equivalence reliability is a measure of the extent to which test-takers receive approximately the same scores on Form B of the test as they did on Form A. Forms that measure the same competencies and yield approximately the same scores are said to be "parallel." If each of your test-takers has the same score on Form B as he or she had on Form A, then you have perfect reliability. If there is no relationship between the test scores on the two forms, then you have a reliability estimate of zero.

Test-retest reliability is the consistency of test scores over time. In other words, did the test-takers receive approximately the same scores on the second administration of the test as they did on the first (assuming no practice or instruction occurred between the two administrations and the administrations were relatively close together)? If your test-takers have the same scores the second time they take the test as they had the first, then you have perfect reliability. Again, if there is no relationship between the test scores, then you have a reliability estimate of zero.

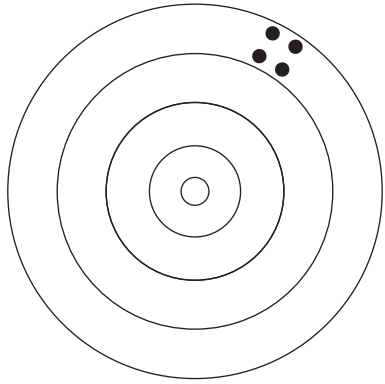
Inter-rater reliability is the measure of consistency among judges' ratings of a performance. If you have determined that a performance test is required, then you need to be sure that your judges (raters) are consistent in their assessments. In Olympic competition we expect that the judges' scores should not deviate significantly from each other. The degree to which they agree is the measure of inter-rater reliability. This agreement will also vary between perfect and zero.

Validity

Validity has to do with whether or not a test measures what it is supposed to measure. A test can be consistent (reliable) but measure the wrong thing. For example, assume that we have designed a course to teach employees how to install a new telephone switchboard. We could devise an end-of-course test that asks learners to list all the steps for installing the new equipment. We might find that the learners can consistently list these steps, but that they can't install the switchboard, which was the intended goal of the course. Hence, our test is reliable, but not a valid measure for the installation task.

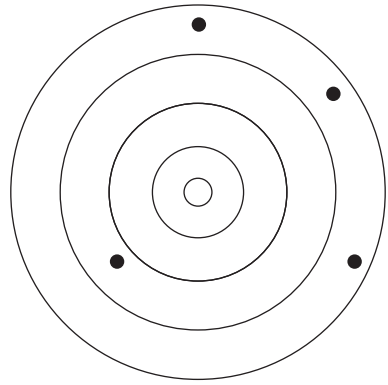
Figure 1.1 illustrates the relationship between reliability and validity. Let's consider that a marksman's job is to hit the center of a shooting target, i.e., the bulls-eye. In Figure 1.1a, the marksman has fired all of her shots in a tight group. Her shooting might be termed "reliable" because the shots are all in the same place, but her shooting isn't valid since she missed the bulls-eye.

Figure 1.1a. Reliable, But Not Valid.



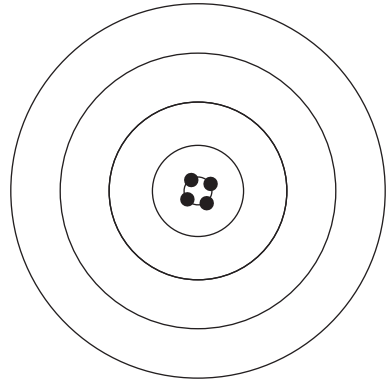
The marksman who produces Figure 1.1b is neither reliable, nor valid.

Figure 1.1b. Neither Reliable Nor Valid.



In Figure 1.1c the marksman's shots are both reliable and valid (she consistently hit the bulls-eye). Notice that it is not possible for the marksman's shots to be valid without also being reliable. Validity requires reliability. Hence, the truism that a test cannot be valid if it is not reliable.

Figure 1.1c. Reliable and Valid.



Practice

1. "Bob, I don't know if this test should be considered a reliable measure of performance. What do you think?"

Person	Week 1 Score	Week 2 Score
Sid	89	90
Indrani	92	90
Atena	75	79
Pui Yi	65	68

2. "Lorie, here's the test you wanted to see. We selected the items to match the job descriptions for our participants. The test scores are highly reliable from one test administration to the next. Do you think this will work?"

Feedback

1. The test appears to be reliable. The scores are very close between each administration. The time lapse of one week is probably a good choice. Waiting too long encourages forgetting or additional learning of the content; not waiting long enough allows pure memorization of the test items.
2. The test may well be valid. The items are linked to the job descriptions, which should increase the likelihood that the items are valid measures of expected performance. Furthermore, the test has demonstrated reliability, a prerequisite for validity. However, it would be impossible to know for sure whether the test were valid without running a job content study as described in Chapter 5.

As mentioned above, test reliability is a necessary but not sufficient condition for test validity. Establishing reliability assures consistency; establishing validity assures that the test consistently measures what it is supposed to measure. And while there are several measures of reliability (which we will discuss in Chapters 14 and 15), it is more important as you begin the CRTD process that you have a basic understanding of four types of validity:

- face validity
- content validity
- concurrent validity
- predictive validity

Of these four, only the latter three are typically assessed formally.

Face Validity. The concept of face validity is best understood from the perspective of the test-taker. A test has face validity if it *appears* to test-takers to measure what it is supposed to measure. For the purposes of defining face validity, the test-takers are not assumed to be content experts. The legitimate purpose of face validity is to win acceptance of the test among test-takers. This is not an unimportant consideration, especially for tests with significant and highly visible consequences for the test-taker. Test-takers who do not do well on tests that lack face validity may be more litigation prone than if the test appeared more valid.

In reality, criterion-referenced tests developed in accordance with the guidelines suggested in this book are not likely to lack face validity. If the objectives for the test are taken from the job or task analysis, and if the test items are then written to maximize their fidelity with the objectives, the test will almost surely have strong face validity. Norm-referenced tests that use test items selected primarily for their ability to separate test-takers, rather than items grounded in competency statements, are much more likely to have face validity problems.

It is important to note that, while face validity is a desirable test quality, it is not adequate to establish the test's true ability to measure what it is intended to measure. The other three types of validity are more substantive for that purpose.

Content Validity. A test possesses content validity when a group of recognized content experts or subject-matter experts has verified that the test measures what it is supposed to measure. Note the distinction between face validity and content validity; content validity is formally determined and reflects the judgments of experts in the content or competencies assessed by the test, whereas face validity is an impression of the test held among non-experts. *Content validity is the cornerstone of the CRTD process and is probably the most important form of validity in a legal defense.* Content validity is not determined through statistical procedures but through logical analysis of the job requirements and the direct mapping of those skills to a test. The detailed procedures for establishing content validity are found in Chapters 5, 6, and 9.

Concurrent Validity. Concurrent validity refers to the ability of a test to correctly classify masters and non-masters. This is, of course, what you *hope* every criterion-referenced test will do; however, face validation and even content validation do not actually demonstrate

the test's ability to classify correctly. Concurrent validation is the technical process that allows you to evaluate the test's ability to distinguish between masters and non-masters of the assessed competencies. The process requires that subject-matter experts identify known masters and non-masters. The test is then administered to each group, and a statistic is calculated to determine that the test can separate these performers of known competence. Concurrent validity procedures are often difficult to apply in the corporate world, although we have seen them used relatively easily in the right circumstances. Chapter 14 lists the steps of this process.

Predictive Validity. Predictive validity is frequently confused with concurrent validity. There is an important conceptual distinction between the two and the procedures for calculating them. Whereas concurrent validity means that a test can correctly classify test-takers of currently known competence, predictive validity means that a test can accurately predict future competence. Predictive validity is important for many personnel selection devices that are used to choose persons for specific job responsibilities. Tests used to help persons select careers also require high predictive validity. In both of these cases, the test is taken first, while the demonstration of competence—job performance or successful career achievement—comes later; hence the term *predictive* validity. The procedures for calculating a test's predictive validity are also found in Chapter 14.

Concluding Comment

As you begin the CRTD process, bear in mind the following observation and let it guide your choices: *An invalid test is not worth anything, to anybody, at any time, for any purpose.*