Introduction

Reliability is one of the most important elements of test quality. It has to do with the consistency, or reproducibility, of an examinee's performance on the test. For example, if you were to administer a test with high reliability to an examinee on two occasions, you would be very likely to reach the same conclusions about the examinee's performance both times. A test with poor reliability, on the other hand, might result in very different scores for the examinee across the two test administrations. If a test yields inconsistent scores, it may be unethical to take any substantive actions on the basis of the test. There are several methods for computing test reliability including test-retest reliability, parallel forms reliability, decision consistency, internal consistency, and interrater reliability. For many criterion-referenced tests decision consistency is often an appropriate choice.

Types of Reliability

Test-Retest Reliability

To estimate test-retest reliability, you must administer a test form to a single group of examinees on two separate occasions. Typically, the two separate administrations are only a few days or a few weeks apart; the time should be short enough so that the examinees' skills in the area being assessed have not changed through additional learning. The relationship between the examinees' scores from the two different administrations is estimated, through statistical correlation, to determine how similar the scores are. This type of reliability demonstrates the extent to which a test is able to produce stable, consistent scores across time.

Parallel Forms Reliability

Many exam programs develop multiple, parallel forms of an exam to help provide test security. These parallel forms are all constructed to match the test blueprint, and the parallel test forms are constructed to be similar in average item difficulty. Parallel forms reliability is estimated by administering both forms of the exam to the same group of examinees. While the time between the two test administrations should be short, it does need to be long enough so that examinees' scores are not affected by fatigue. The examinees' scores on the two test forms are correlated in order to determine how similarly the two test forms function. This reliability estimate is a measure of how consistent examinees' scores can be expected to be across test forms.

Decision Consistency

In the descriptions of test-retest and parallel forms reliability given above, the consistency or dependability of the *test scores* was emphasized. For many criterion referenced tests (CRTs) a more useful way to think about reliability may be in terms of examinees' *classifications*. For example, a typical CRT will result in an examinee being classified as either a master or non-master; the examinee will either pass or fail the test. It is the reliability of this classification decision that is estimated in decision consistency reliability. If an examinee is classified as a master on both test administrations, or as a non-master on both occasions, the test is producing consistent decisions. This approach can be used either with parallel forms or with a single form administered twice in test-retest fashion.

Internal Consistency

The internal consistency measure of reliability is frequently used for norm referenced tests (NRTs). This method has the advantage of being able to be conducted using a single form given at a single administration. The internal consistency method estimates how well the set of items on a test correlate with one another; that is, how similar the items on a test form are to one another. Many test analysis software programs produce this reliability estimate automatically. However, two common differences between NRTs and CRTs make this method of reliability estimation less useful for CRTs. First, because CRTs are typically designed to have a much narrower range of item difficulty, and examinee scores, the value of the reliability estimate will tend to be lower. Additionally, CRTs are often designed to measure a broader range of content; this results in a set of items that are not necessarily closely related to each other. This aspect of CRT test design will also produce a lower reliability estimate than would be seen on a typical NRT.

Interrater Reliability

All of the methods for estimating reliability discussed thus far are intended to be used for objective tests. When a test includes performance tasks, or other items that need to be scored by human raters, then the reliability of those raters must be estimated. This reliability method asks the question, "If multiple raters scored a single examinee's performance, would the examinee receive the same score. Interrater reliability provides a measure of the dependability or consistency of scores that might be expected across raters.

Summary

Test reliability is the aspect of test quality concerned with whether or not a test produces consistent results. While there are several methods for estimating test reliability, for objective CRTs the most useful types are probably test-retest reliability, parallel forms reliability, and decision consistency. A type of reliability that is more useful for NRTs is internal consistency. For performance-based tests, and other tests that use human raters, interrater reliability is likely to be the most appropriate method.