An Instructor's Guide to Understanding Test Reliability

Craig S. Wells
James A. Wollack

Testing & Evaluation Services
University of Wisconsin
1025 W. Johnson St., #373
Madison, WI 53706

November, 2003

An Instructor's Guide to Understanding Test Reliability

Test reliability refers to the consistency of scores students would receive on alternate forms of the same test. Due to differences in the exact content being assessed on the alternate forms, environmental variables such as fatigue or lighting, or student error in responding, no two tests will consistently produce identical results. This is true regardless of how similar the two tests are. In fact, even the same test administered to the same group of students a day later will result in two sets of scores that do not perfectly coincide. Obviously, when we administer two tests covering similar material, we prefer students' scores be similar. The more comparable the scores are, the more reliable the test scores are.

It is important to be concerned with a test's reliability for two reasons. First, reliability provides a measure of the extent to which an examinee's score reflects random measurement error. Measurement errors are caused by one of three factors: (a) examinee-specific factors such as motivation, concentration, fatigue, boredom, momentary lapses of memory, carelessness in marking answers, and luck in guessing, (b) test-specific factors such as the specific set of questions selected for a test, ambiguous or tricky items, and poor directions, and (c) scoring-specific factors such as nonuniform scoring guidelines, carelessness, and counting or computational errors. These errors are random in that their effect on a student's test score is unpredictable – sometimes they help students answer items correctly while other times they cause students to answer incorrectly. In an unreliable test, students' scores consist largely of measurement error. An unreliable test offers no advantage over randomly assigning test scores to students.

Therefore, it is desirable to use tests with good measures of reliability, so as to ensure that the test scores reflect more than just random error.

The second reason to be concerned with reliability is that it is a precursor to test validity. That is, if test scores cannot be assigned consistently, it is impossible to conclude that the scores accurately measure the domain of interest. Validity refers to the extent to which the inferences made from a test (i.e., that the student knows the material of interest or not) is justified and accurate. Ultimately, validity is the psychometric property about which we are most concerned. However, formally assessing the validity of a specific use of a test can be a laborious and time-consuming process. Therefore, reliability analysis is often viewed as a first-step in the test validation process. If the test is unreliable, one needn't spend the time investigating whether it is valid–it will not be. If the test has adequate reliability, however, then a validation study would be worthwhile.

There are several ways to collect reliability data, many of which depend on the exact nature of the measurement. This paper will address reliability for teacher-made exams consisting of multiple-choice items that are scored as either correct or incorrect. Other types of reliability analyses will be discussed in future papers.

The most common scenario for classroom exams involves administering one test to all students at one time point. Methods used to estimate reliability under this circumstance are referred to as measures of *internal consistency*. In this case, a single score is used to indicate a student's level of understanding on a particular topic. However, the purpose of the exam is not simply to determine how many items students answered correctly on a particular test, but to measure how well they know the content area. To achieve this goal, the particular items on the test must be sampled in a way as to be representative of the entire domain of interest. It is expected that students mastering

the domain will perform well and those who have not mastered the domain will perform less well, regardless of the particular sample of items used on the exam. Furthermore, because all items on that test tap some aspect of a common domain of interest, it is expected that students will perform similarly across different items within the test.

Reliability Coefficient for Internal Consistency

There are several statistical indexes that may be used to measure the amount of internal consistency for an exam. The most popular index (and the one reported in Testing & Evaluation's item analysis) is referred to as Cronbach's alpha. Cronbach's alpha provides a measure of the extent to which the items on a test, each of which could be thought of as a mini-test, provide consistent information with regard to students' mastery of the domain. In this way, Cronbach's alpha is often considered a measure of item homogeneity; i.e., large alpha values indicate that the items are tapping a common domain. The formula for Cronbach's alpha is as follows:

$$\hat{\alpha} = \frac{k}{k-1}\left(1 - \frac{\sum_{i=1}^{k} p_i(1-p_i)}{\hat{\sigma}_X^2}\right).$$

k is the number of items on the exam; p_i, referred to as the item difficulty, is the proportion of examinees who answered item i correctly; and $\hat{\sigma}_X^2$ is the sample variance for the total score. To illustrate, suppose that a five-item multiple-choice exam was administered with the following percentages of correct response: $p_1 = .4$, $p_2 = .5$, $p_3 = .6$, $p_4 = .75$, $p_5 = .85$, and $\hat{\sigma}_X^2 = 1.84$. Cronbach's alpha would be calculated as follows:

$$\hat{\alpha} = \frac{5}{5-1}\left(1 - \frac{1.045}{1.840}\right) = .54.$$

Cronbach's alpha ranges from 0 to 1.00, with values close to 1.00 indicating high consistency. Professionally developed high-stakes standardized tests should have internal consistency coefficients of at least .90. Lower-stakes standardized tests should have internal consistencies of at least .80 or .85. For a classroom exam, it is desirable to have a reliability coefficient of .70 or higher. High reliability coefficients are required for standardized tests because they are administered only once and the score on that one test is used to draw conclusions about each student's level on the trait of interest. It is acceptable for classroom exams to have lower reliabilities because a student's score on any one exam does not constitute that student's entire grade in the course. Usually grades are based on several measures, including multiple tests, homework, papers and projects, labs, presentations, and/or participation.

Suggestions for Improving Reliability

There are primarily two factors at an instructor's disposal for improving reliability: increasing test length and improving item quality.

Test Length

In general, longer tests produce higher reliabilities. This may be seen in the old carpenter's adage, "measure twice, cut once." Intuitively, this also makes a great deal of sense. Most instructors would feel uncomfortable basing midterm grades on students' responses to a single multiple-choice item, but are perfectly comfortable basing midterm grades on a test of 50 multiple-choice items. This is because, for any given item, measurement error represents a large percentage of students' scores. The percentage of measurement error decreases as test length increases. Even very low achieving students can answer a single item correctly, even through guessing; however it is much less likely that low achieving students can correctly answer all items on a 20-item test.

Although reliability does increase with test length, the reward is more evident with short tests than with long ones. Increasing test length by 5 items may improve the reliability substantially if the original test was 5 items, but might have only a minimal impact if the original test was 50 items. The Spearman-Brown prophecy formula (shown below) can be used to predict the anticipated reliability of a longer (or shorter) test given a value of Cronbach's alpha for an existing test.

$$\alpha^{new} = \frac{m\alpha^{old}}{1+(m-1)\alpha^{old}}$$

α^{new} is the new reliability estimate after lengthening (or shortening) the test; α^{old} is the reliability estimate of the current test; and *m* equals the new test length divided by the old test length. For example, if the test is increased from 5 to 10 items, *m* is 10 / 5 = 2. Consider the reliability estimate for the five-item test used previously ($\hat{\alpha} = .54$). If the test is doubled to include 10 items, the new reliability estimate would be

$$\alpha^{new} = \frac{2(.54)}{1+(2-1)*.54} = .70,$$

a substantial increase. Note, however, that increasing a 50-item test (with the same reliability) by 5 items, will result in a new test with a reliability of just .56.

It is important to note that in order for the Spearman-Brown formula to be used appropriately, the items being added to lengthen a test must be of a similar quality as the items that already make-up the test. In addition, before lengthening a test, it is important to consider practical constraints such as time limit and examinee fatigue. As a general guideline, it is wise to use as many items as possible while still allowing most students to finish the exam within a specified time limit.

Item Quality

Item quality has a large impact on reliability in that poor items tend to reduce reliability while good items tend to increase reliability. How does one know if an item is of low or high quality? The answer lies primarily in the item's discrimination. Items that discriminate between students with different degrees of mastery based on the course content are desirable and will improve reliability. An item is considered to be discriminating if the "better" students tend to answer the item correctly while the "poorer" students tend to respond incorrectly. Item discrimination can be measured with a correlation coefficient known as the point-biserial correlation (r_{pbi}). r_{pbi} is the correlation between students' scores on a particular item (1 if the student gets the item correct and 0 if the student answers incorrectly) and students' overall total score on the test. A large, positive r_{pbi} indicates that students with a higher test score tended to answer the item correctly while students with a lower test score tended to respond incorrectly. Items with small, positive r_{pbi}'s will not improve reliability much and may even reduce reliability in some cases. Items with negative r_{pbi}'s will reduce reliability. For a classroom exam, it is preferable that an item's r_{pbi} be 0.20 or higher for all items. Note that the item analysis provided by Testing and Evaluation Services reports the r_{pbi} for each item.

Regarding item difficulty, it is best to avoid using too many items that nearly all of the students answer correctly or incorrectly. Such items do not discriminate well and tend to have very low r_{pbi}'s. In general, 3-, 4-, and 5-alternative multiple-choice items that are answered correctly by about 60% of the students tend to produce the best r_{pbi}'s. For 2-alternative items, the target item difficulty is 75% correct.