NCME Instructional Module on

Reliability of Scores From Teacher-Made Tests

David A. Frisbie
The University of Iowa

Reliability is the property of a set of test scores that indicates the amount of measurement error associated with the scores. Teachers need to know about reliability so that they can use test scores to make appropriate decisions about their students. The level of consistency of a set of scores can be estimated by using the methods of internal analysis to compute a reliability coefficient. This coefficient, which can range between 0.0 and +1.0, usually has values around 0.50 for teacher-made tests and around 0.90 for commercially prepared standardized tests. Its magnitude can be affected by such factors as test length, test-item difficulty and discrimination, time limits, and certain characteristics of the group—extent of their testwiseness, level of student motivation, and homogeneity in the ability measured by the test.

Reliability is the name given to one of the properties of a set of test scores—the property that describes how consistent or error-free the measurements are. We know that some tests can be fairly precise measuring tools, but we also realize that sometimes the scores they yield are not so dependable; students can obtain scores that are either higher or lower than they really ought to be. Consequently, it is important for teachers to determine how consistent the scores from their tests are so that those scores can be used wisely to make instructional decisions about students.

Scores from teacher-made tests are used by teachers and students for a variety of purposes. For example, the scores for a class indicate whether learning has been complete and whether instruction has been effective for the class. For individual students, areas of deficiency might be identified so that remediation can be planned. The scores indicate to students whether or not they are prepared adequately for the next stages of instruction. If the test scores are not very reliable, if they misrepresent students' true level of knowledge, inappropriate decisions might be made that could have negative effects—temporary or lasting—on the students. Having inadequate information may be worse than having no information at all.

The emphasis in this unit will be on the reliability of scores from teacher-made achievement tests that are intended to be used for making norm-referenced score interpretations. (The purpose of norm-referenced interpretations is to describe the test performance of a student by comparing his or her score with the scores of other students. That is, we want to obtain a rank ordering of students' scores that represents the actual differences in achievement among students. We make "reference" to this rank ordering of scores to judge how high or how low a particular student's score is.) The ideas presented in this unit also can be applied to some extent to the use of scores from standardized achievement and aptitude tests and to scores obtained to make content-referenced (criterion-referenced) interpretations. But the focus of our discussion will be norm-referenced tests prepared by teachers (or publishers of textbooks and other instructional materials) to measure the achievement of students, whether at the elementary school or the college level. Because of this focus, our discussion will not include much theoretical background about reliability, and it will be limited to four of the several methods that can be used to estimate score reliability. (Some of these other topics, classical reliability procedures and reliability of scores from criterion-referenced tests, will be dealt with in other units in this series.)

David A. Frisbie is Associate Professor of Measurement and Statistics and Assistant Director of the Iowa Basic Skills Testing Program, 316 Lindquist Center, Iowa City, IA 52242. His specialization is achievement testing.

Series Information

ITEMS is a series of units designed to facilitate instruction in educational measurement. These units are published by the National Council on Measurement in Education. This module may be photocopied without permission if reproduced in its entirety and used for instructional purposes. Barbara S. Plake, University of Nebraska-Lincoln, has served as the editor for this module.

Objectives

The objectives of the instruction provided in this module are to help the learner to do the following:

1. Explain the meaning of test score reliability.
2. Identify and describe a variety of factors that can influence the magnitude of a student's test score.

3. Describe methods that can be used by teachers to estimate the reliability of the scores obtained by students on classroom tests.
4. Explain how the level of score reliability relates to the appropriate use of test scores.
5. Identify factors associated with test takers, the examination, and the testing conditions that can affect the size of the reliability coefficient for a set of test scores.
6. Explain the relative merits of the alternative actions that might be taken when a set of scores is too unreliable.

Test Score Reliability

Consistency of Measurements

The scores from tests are measurements in much the same way as the numbers obtained from using a scale at the meat market, a carpenter's ruler, or a cook's measuring spoons. Any of these measurements may be more or less consistent, depending on the quality of the measuring tool and the care taken by its user. Measurement errors can explain why a student's score on the same test will vary from one day to the next or why a student might score lower on one test than on another comparable test.

We should not expect test scores to be perfect measurements, but there is only so much error that we should be willing to tolerate. A geography test should yield scores that put students in the same relative order, whether given on Tuesday or Wednesday. If a slightly different set of 40 test questions had been used, essentially the same relative ordering of student scores should have resulted. That is, anytime a classroom test is given, we would like the resulting scores to be generalizable over testing occasions, over sets of similar test questions, and over slightly varying testing conditions. We need to be able to depend on the scores to be consistent enough to be useful. If we cannot rely on the scores as accurate measures of achievement, we cannot use them to make instructional decisions or to communicate progress to students or their parents.

It is probably apparent by now that reliability is a property of a set of test scores, *not* a property of the test itself. An English test could yield fairly accurate scores on a certain day when given to a particular class, but could yield fairly inconsistent scores when given to a different class or to the same class on another occasion. Why this is so will become more apparent when, in the next section, we examine the factors that can influence reliability. Even though we may hear some people refer to a test as "very reliable," what they really mean is that the scores we obtained from a particular group on a certain day and under certain testing conditions could be reproduced by giving an equivalent test to that group under these same conditions. That is, the scores it yielded are highly consistent.

One way of describing reliability, therefore, is in terms of reproducibility. To the extent that we are able to obtain the same results on subsequent measurements of the same students, our measurements are consistent. Obtaining the "same results" could mean that (a) everyone obtains the same score on both measurement occasions or (b) the relative *order* of students' scores on the two occasions is the same (but individuals' scores may be different from one time to the next). It is this second meaning that is used most frequently to define a reliability coefficient. It is an index of the level of consistency of a set of scores: "The reliability coefficient for a set of scores from a group of examinees is the coefficient of correlation between that set of scores and another set of scores on an equivalent test obtained independently from the members of the same group" (Ebel & Frisbie, 1986, p. 71).

Factors That Contribute to Inconsistency

Errors occur in the process of measuring students' achievements even though we may not be able to observe them and students may not be able to notice that they have occurred. What causes test scores to be different from what they really ought to be? For what reasons do students who take the same test at the same time and in the same place obtain scores that differ from one another? We must be able to answer these fundamental questions to understand the conditions under which errors may occur and to prevent them or to minimize their effect on the test scores.

As an example, Scott and Marcy might obtain different scores on the same literature test because Scott knows more about the content covered by the test. Of course, this is the kind of difference we want test scores to reflect, and no inaccuracies are involved if this is the sole explanation for the score difference. All other explanations for the difference are potential sources of inaccuracy, factors that contribute to measurement error. Here are some examples:

1. Marcy did not read the instructions carefully and forgot to answer the five questions on the back side of the last page. These items were marked wrong.
2. Because of his more extensive test-taking experience, Scott was able to detect certain idiosyncrasies in this teacher's item-writing habits and to use these clues to choose correct answers or to eliminate some wrong choices in several multiple-choice items. Marcy was unaware of these unintended clues.
3. Marcy was unable to concentrate fully on the test because she was continually blowing her nose and sneezing (or because she was so tired from having stayed up so late the previous night, or because the adrenaline was still flowing from the argument she had just before class with her friend Stacy).
4. Scott was fortunate in that the two essay questions related closely to what he had most recently studied, but Marcy had concentrated her study in several other areas instead. She might have been much more successful had a different pair of essay questions been asked.
5. Though Marcy is an above-average reader, for some unexplainable reason she had to reread nearly everything because she seemed unable to concentrate well enough to comprehend on the first reading. Scott's attention did not seem to fluctuate in any unusual way during the test.
6. Marcy sat near the air conditioning vent and became so cold as the test progressed that she began to shiver and feel as if she needed to use the restroom. The conditions around Scott were less extreme and did not seem to affect his attentiveness as he worked through the test.
7. The teacher recognized both Scott's and Marcy's handwriting when scoring the essay responses. He seemed particularly lenient with one of Scott's incomplete responses and probably should have awarded Marcy a few more points than

he actually did for one of her responses.

8. Scott guessed correctly on four of five multiple-choice items, but Marcy was correct on only two of the six guesses she made.

In each of the eight illustrations given above, the errors that occurred affected Scott and Marcy differently and probably affected all other students in other various ways. We call these errors random because if we were to give these students a different, equivalent test or give them the same test again, we would expect these errors to have a somewhat different effect the second time. For example, Marcy might not forget to answer the last few questions, but Scott might; there might be fewer items with which Scott can employ his test-taking skills to improve his score; Marcy might feel perfectly healthy, but Scott might have a cold or feel depressed or be very tired; the two essay questions might have content that is equally familiar to both; Marcy's concentration might be high enough that she needs to do little, if any, rereading; both might be somewhat affected by the heat and humidity because the air conditioner is not working; the teacher might show no leniency in scoring Scott's essay responses, there might be no incomplete responses from Scott, or Marcy might be awarded "too many" points for a unique or creative expression in one of her responses; or, finally, each student might have guessed correctly on one of four multiple-choice items. Each type of error might be present or absent in a specific testing situation for a given test taker. Sometimes the effect of an error will be fairly large, sometimes it will be fairly small, and sometimes it will be absent altogether. When they do occur, some types of errors are positive and some are negative. If we were to give the same test to a person repeatedly (and assume each administration to be independent of the others), the effect of each type of random error should vary over each of these occasions. But if Scott's testwiseness helped him get exactly two items correct every time, this error would not be considered random for him.

As long as these kinds of errors are not predictable from student to student on a particular testing occasion, we call the errors random. Random errors cause students tested at the same time to obtain scores that differ from one another for the wrong reasons. If we could compute an error score for each student, some would have positive scores (their observed scores were too high), some would have negative scores, and some would have a score of zero. Because these are *random* errors that we are talking about, the average of the error scores of all students on this occasion should be zero.

Errors may be systematic rather than random. Systematic errors affect all examinees to nearly the same extent and cause *all* scores to be higher or lower than they really ought to be. These kinds of errors, then, affect the absolute size of the students' scores, but they do not cause students to have scores that differ from one another by an appreciable amount. These incidents might cause systematic errors to influence the scores of Scott and Marcy (and all their classmates) in the same way on the literature test:

1. A fire drill shortened the testing period by 10 minutes.
2. Three of the multiple-choice questions were so easy (all the wrong responses were very implausible) that nobody missed any one of them.
3. There was one 7-point essay for which nobody really knew the answer, but everyone wrote something, and everyone got at least three points, even if the response was wrong or irrelevant.

Notice that such systematic errors would interfere with our ability to make domain-referenced (absolute) score interpretations, but they would not affect negatively any norm-referenced (relative) interpretations we make. Systematic errors tend to inflate or deflate all examinees' scores by a constant amount, but random errors affect nearly every individual's score in a somewhat different way.

An error may be systematic for one student but not for the whole class. For example, a student who reads very poorly may obtain a score that is 25% lower than it ought to be on every literature test. That is, the student knows a good deal more than his or her test score shows: Poor reading skills systematically mask this student's true level of achievement as represented by the test score. For this student, error due to reading skills is systematic, but for the class, reading ability may not be a detrimental factor, or it may have a differential effect across individuals. The example of Scott's using his testwiseness to get two items correct every time illustrates systematic error in Scott's score. Note that systematic errors do not cause a student's score to be inconsistent from test to test.

Fortunately, not all of these potential errors—random or systematic—are likely to happen every time we administer a test, and not all of the errors are likely to have a marked effect on students' test scores. The cumulative effect of a host of small errors can distort a test score enough, however, that the level of achievement it represents is quite misleading. Consequently, we need to estimate how much error is associated with a set of scores so that we can judge whether the scores are useful for our original testing purpose.

Learning Exercise A:

Determine whether each situation described is likely to contribute (a) random error, (b) systematic error, or (c) no error to the test scores:

1. Some students were able to determine that choice "a" never seemed to be the correct answer for any of the multiple-choice items.
2. Everyone knew the definition for "parallel lines" and, consequently, everyone got that item right.
3. Everyone missed the science item that contained the word "proselyte" because they did not know the meaning of the term.
4. Ben guessed correctly on two true-false items, but Jessica guessed correctly on five true-false items.
5. Scores on the 5-point essay item ranged from 3 to 5. Everyone who wrote something was awarded at least 3 points, regardless of the quality of the response. Tom even got 3 points for writing, "I don't know."

Answers:

1. The condition describes variable testwiseness because only *some* students benefited from the discovery about choice "a"—random.
2. There is no apparent error implied by the statement;

everyone got the item right because they possessed the requisite knowledge—no error.

3. The statement implies that the only reason the item was missed was because of a vocabulary word unrelated to science content. The score of each student was lowered by 1 point—systematic.

4. Good luck due to guessing is distributed to examinees differently; some have much good luck, some have a little, and some have bad luck—random.

5. It can be deduced that everyone wrote something and some responded well enough to earn a perfect score. We cannot tell if everyone's score was systematically higher than it should have been by 3 points, but Tom's was—random.

Methods of Estimating Reliability

A reliability coefficient is a number that provides an index of the amount of error associated with a particular set of test scores. It can range from 1.00, indicating perfect reliability or no measurement error, down to 0.00, indicating that the presence (abundance) of random error is the only reason why students obtained scores that differed from one another.

The reliability coefficient is not the only index available for describing how consistent the scores from a test are. The standard error of measurement (SEM), another such index, is particularly useful as an aid to score interpretation because it can be computed and expressed in terms of raw score points or standard score units, depending on the type of score being used. The SEM instructional unit describes the various approaches for representing and computing the SEM and the relative merits of each approach.

There are many methods available for computing reliability coefficients, but each method does not yield the same result. The main reason for this seeming inconsistency is that each method is designed to detect the presence of only certain kinds of errors. For a given testing situation, some methods will detect certain errors and other methods will not; some methods will treat certain errors as though they were random but other methods will treat those same errors as if they were systematic. (This idea will be given more lengthy attention in the module that covers classical reliability.) For our purposes, we will focus our attention on the methods that are most commonly used and most appropriate to use to detect the errors that plague teacher-made test scores most frequently.

Most teachers probably do not compute reliability coefficients, partly because of the computational difficulties of some methods and partly because of an incomplete understanding of reliability. The computational burden has been eased considerably by the widespread availability of microcomputer software that will compute reliability coefficients, test score means and standard deviations, and other useful test statistics. These developments, coupled with the increased availability of small test-scoring machines (optical scanners) that can be used directly with a microcomputer, have made the test-scoring and test-score-evaluation processes efficient for teachers to accomplish. Thus it is reasonable to encourage and expect teachers to be more attentive than they have been in the past to the quality of their tests and, in particular, to the reliability of the scores derived from them.

Methods of internal analysis are the most appropriate to use with scores from classroom tests because these methods can detect errors due to content sampling and to differences among students in testwiseness, ability to follow instructions, scoring bias, and luck in guessing answers correctly—the types of errors most likely to affect teacher-made test scores. These methods include coefficient alpha, Kuder-Richardson Formula 20 (K-R20), Kuder-Richardson Formula 21 (K-R21), and adjusted K-R21 (K-R21′). We will give some attention to the computation of these coefficients, primarily as a means of fostering an understanding of each coefficient and the various factors that affect its magnitude.

Coefficient alpha (α). Alpha can provide an estimate of the reliability of scores from tests composed of any assortment of item types—essays, multiple-choice, numerical problems, true-false, or completion. The formula to use is

$$\alpha = \frac{k}{k-1}\left[1-\frac{\Sigma S_i^2}{S_t^2}\right] \quad (1)$$

where k is the number of items, ΣS_i^2 is the sum of the variances for the separate test items, and S_t^2 is the variance for the set of student total test scores. The scores shown in Table 1 for 15 students can be used to illustrate the computation of alpha. The number of items, k, is 12 (10 multiple-choice and two essays). The variance of each item can be calculated using the raw score formula

$$S^2 = \frac{n(\Sigma X^2)-(\Sigma X)^2}{n^2} \quad (2)$$

where X represents an individual's score and n is the number of individuals for whom scores are available. For Item 1, $\Sigma X = 10$, $\Sigma X^2 = 10$, $n = 15$, and $S^2 = .222$. For Item 11, $\Sigma X = 85$, $\Sigma X^2 = 549$, and $S^2 = 4.49$. Can you verify these values for Item 12: $\Sigma X = 50$, $\Sigma X^2 = 188$, and $S^2 = 1.42$? Can

TABLE 1

Test Item Scores for 15 Students on 10 Multiple-Choice Items and 2 Essay Items

Item number	Student A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	0	1	1	1	1	1	0	1	1	1	0	1	1	0	0
2	1	1	1	1	1	0	0	1	1	0	0	0	0	0	0
3	1	1	0	0	0	1	0	1	0	1	1	0	1	0	0
4	0	1	1	1	1	1	1	0	0	0	0	0	0	0	0
5	1	1	1	1	0	1	1	1	0	0	1	0	0	0	0
6	0	0	0	1	1	1	1	1	1	0	1	1	1	1	1
7	1	0	1	0	0	1	1	0	0	0	0	0	0	0	1
8	1	1	1	1	1	1	1	0	1	1	1	0	0	0	1
9	0	1	1	0	1	1	1	1	0	1	0	1	1	0	1
10	1	1	0	1	1	0	0	1	0	1	0	0	1	0	0
11	7	10	6	5	6	8	3	5	6	3	3	5	9	3	6
12	2	4	3	3	4	5	4	3	5	2	1	4	5	2	3

Note. Items 11 and 12 were essays having maximum scores of 10 and 5 points, respectively. Items 1–10 were multiple-choice items scored 1 for a correct answer and 0 for an incorrect answer.

you show that the variance for Item 2 is .249? The sum of the variances for all 12 items is 8.203. Finally, the variance of the test scores, S_t^2, can be found with this same formula (2) once the total scores for Students A through O have been summed. Verify that the score for Student A is 15, for B is 22, and for H is 15. The variance of the test scores, then, can be found to equal 18.25 ($\Sigma X = 217$, $\Sigma X^2 = 3413$, and $n = 15$). Now alpha can be computed for this 12-item test as follows:

$$\alpha = \frac{12}{11}\left[1 - \frac{8.203}{18.26}\right]$$
$$\alpha = 1.09\,[1 - .4492]$$
$$\alpha = 0.572$$

K-R20. This formula, a simplified version of coefficient alpha, is to be used only when scoring is dichotomous (1 for correct and 0 for wrong), as is usually done with objective tests. The similarity between K-R20 and alpha can be seen by comparing the two computational formulas (Equations 1 and 3):

$$\text{K-R20} = \frac{k}{k-1}\left[1 - \frac{\Sigma pq}{S_t^2}\right] \quad (3)$$

Here p is the proportion of the group that answered an item correctly and q is the proportion that answered it incorrectly. With this kind of scoring (i.e., one-zero) $\Sigma pq = \Sigma S_i^2$. In Table 1, the values of p and q for Item 1 are .67 and .33. The value of pq is .221, the same as the value computed earlier for the variance of Item 1. Can you show that the variances for Items 4 and 7 are .240 and .222, respectively? Can you verify that $\Sigma pq = 2.293$ for Items 1–10? Before we can find the value of K-R20 for Items 1–10, we must compute the variance of the scores on the 10-item test. After finding the total score for each person (A = 6, B = 8, etc.), the variance of these scores can be found ($\Sigma X = 82$, $\Sigma X^2 = 504$, and $S_t^2 = 3.716$). Finally, K-R20 can be calculated as follows:

$$\text{K-R20} = \frac{10}{9}\left[1 - \frac{2.293}{3.716}\right]$$
$$\text{K-R20} = 1.11\,[1 - .6171]$$
$$\text{K-R20} = 0.425$$

K-R21. *Both* α and K-R20 require computation of item variances, a rather time-consuming task when the number of items is large and electronic computing is not available. K-R21 is simpler to compute than either of the others but, like K-R20, it is appropriate to use only when items are scored dichotomously. The formula is

$$\text{K-R21} = \frac{k}{k-1}\left[1 - \frac{\bar{X}(k-\bar{X})}{kS_t^2}\right] \quad (4)$$

where k is the number of items, $\bar{X}$ is the mean score, and S_t^2 is the score variance. For the first 10 items in Table 1, the mean for the 15 students is 5.47 ($\Sigma X = 82$), and the variance is 3.716.

$$\text{K-R21} = \frac{10}{9}\left[1 - \frac{(5.47)(10-5.47)}{10(3.716)}\right]$$
$$\text{K-R21} = 1.11\left[1 - \frac{24.78}{37.16}\right]$$
$$\text{K-R21} = 0.370$$

The value of K-R21 is nearly always less than the value of K-R20, using the same set of scores, because certain assumptions required for the use of K-R21 are rarely met. The value of K-R21 should be thought of as an estimate of K-R20. In fact, it could be regarded as the lowest value K-R20 would have if *it* had been computed.

K-R21′. Because K-R21 is so convenient to calculate and is considered the lowest possible value that K-R20 would have, some researchers have developed an adjusted K-R21 formula to obtain a closer approximation to the K-R20 value (Wilson, Downing, & Ebel, 1977):

$$\text{K-R21}' = 1 - \left[\frac{(.8)(\bar{X})(k-\bar{X})}{kS_t^2}\right] \quad (5)$$

The value of K-R21′ for the first 10 items in Table 1 is 0.466. Can you verify this? In this case, the K-R21′ did only slightly better than K-R21 at estimating the value of K-R20: K-R21′ overestimated by .041, and K-R21 underestimated by .055. Ordinarily, K-R21′ and K-R21 both will underestimate K-R20, but K-R21′ will be appreciably more accurate.

Of the four methods described above, only coefficient alpha can be used universally without regard to the nature of the scoring method. For essay tests, objective tests, numerical or problem tests, or any combination of them, the reliability of the scores can be estimated with coefficient alpha. The Kuder-Richardson methods are appropriate only when dichotomous scoring is used.

Interpreting Reliability Coefficients

There are no absolute standards that can be used to judge whether a particular reliability coefficient is high enough. Some relative standards have evolved based on what has been observed about the reliability of scores under certain circumstances. For example, most published standardized tests yield scores that have reliabilities in the range .85–.95, values regarded by most as highly acceptable. Teacher-made tests, on the other hand, tend to yield score reliabilities that average about .50.

The standards for minimally acceptable values for test score reliability need to be established in the context of score use. That is, how reliable the scores must be depends mostly on how the scores will be used—what kinds of decisions will be made and how much weight the test score will have in the decision. Experts in educational measurement have agreed informally that the reliability coefficient should be at least .85 if the scores will be used to make decisions about individuals *and* if the scores are the only available useful information. (This ought to be a very rare circumstance.) However, if the decision is about a *group* of individuals, the generally accepted minimum standard is .65.

Usually, we can tolerate reliabilities around .50 for scores from teacher-made tests if each score will be combined with other information—test scores, quiz scores, observations—to assign a grade for quarter or semester work. It is the reliability of the score that results from combining the collection of measurements that should concern us the most; it is this score, not the score from any one test, on which grading decisions will be made. When an important decision is made using a single score, we need to be concerned about the reliability of that single set of scores. For example, if a teacher uses a placement test to determine the most appropriate starting point in instruction for each student, those decisions will be important and little additional corroborating information will be at hand. Consequently, our standard for

acceptable reliability for such placement test scores should be noticeably higher than that for achievement test scores that will be used only for grading.

Factors That Affect Reliability Estimates

If we understand the factors that contribute to test score inconsistencies and if we compute reliability estimates for the scores from our tests, we should be able to use and interpret the test scores prudently. But that is not enough! We must be able to build tests that will help us achieve score reliability estimates that are at least minimally acceptable, and we must be able to revise our tests so that "improved" versions will yield more reliable scores in the future. To this end, we will consider some of the factors associated with the test that can be manipulated or controlled to enhance score reliability. Other factors associated with either the examinees or the testing conditions will be considered, too.

Test length. Scores from a longer test are apt to be more reliable than the scores from a shorter one. This is true because the longer test is likely to yield a greater spread of scores. Intuition suggests that a more dependable, more reproducible rank ordering of students can be achieved with a 10-item test, for example, than with a 5-item test. When there are more categories into which individuals can be sorted (0–10 rather than 0–5), the scores we assign to students can reflect the differences in their actual achievement more consistently. As the number of separate pieces of information we obtain about the achievement of each examinee increases, we can become increasingly accurate as we rank order the individuals in terms of their achievement.

Test content. Tests that measure the achievement of a somewhat homogeneous set of topics are likely to yield more reliable scores than tests that measure a potpourri of somewhat unrelated ideas. Each of the methods of internal analysis described above for estimating reliability is an index of item homogeneity, an indication of the extent to which all the items in the test measure a single domain of content. (The theoretical explanation for this phenomenon relates to item intercorrelations, a topic beyond the scope of this instructional module.) A test that has items that measure reading comprehension, computational skills, and knowledge of the principles of test construction probably will yield less reliable scores than a test of comparable length that measures only one of these traits.

Item difficulty. All the items in a test need to be in the moderate range of difficulty, neither too hard nor too easy for the group, to help identify differences in achievement among students. An item that everyone in a class answers correctly (What color is the White House?) does not help to show who has achieved more or less; neither does an item that everyone misses (In what country is Lake Fromme located?). Consequently, in the small amount of time available for testing, the very easy or very difficult test items do little to further our purpose for testing. In fact, they take up valuable testing time and return very little information that helps us rank order individuals precisely.

Item discrimination. Items that discriminate properly are answered correctly by most of the students who earn high scores on the test and are missed by most of those who earn low test scores. Items that discriminate properly help to accumulate high scores for those who have learned and keep low achievers from obtaining high scores on the test. If a certain item is answered correctly only by low-achieving students, it would be discriminating improperly because it would elevate the scores of the wrong students. And an item that fails to discriminate (equal numbers of high- and low-achieving students answer it correctly) elevates test scores randomly or "indiscriminately." Highly discriminating items help to distinguish between examinees of different achievement levels and, consequently, they contribute substantially to test score reliability. In fact, the single most useful action to take in an attempt to improve the reliability of scores from a certain test is to improve each item's ability to discriminate. The test with the highest average item discrimination index is likely to yield scores of highest reliability.

Group heterogeneity. The reliability estimate will be higher for a group that is heterogeneous with respect to achievement of the test content than it will be if the group is homogeneous. When a group is very homogeneous, it is more difficult to achieve a spread of scores and to detect the small differences that actually exist. The scores we obtain in such situations usually are so similar to one another that we are not sure if the differences are real or due strictly to random error. When interindividual differences are greater, as in a more heterogeneous group, the rank ordering of individuals is likely to be replicated more easily on a retest.

Student motivation. If students are not motivated to do their best on a test, their scores are not apt to represent their actual achievement levels very well. But when the consequences of scoring high or low are important to examinees, the scores are likely to be more accurate. Indifference, lack of motivation, or underenthusiasm, for whatever reasons, can depress test scores just as much as anxiety or overenthusiasm may.

Student testwiseness. When the amount of test-taking experience and levels of testwiseness vary considerably within a group, such backgrounds and skills may cause scores to be less reliable than they otherwise would be. When all examinees in the group are experienced and sophisticated test takers or when all are relatively naive about test taking, such homogeneity probably will not lead to much random measurement error. The rank order of scores is likely to be influenced only when there is obvious variability in testwiseness within the group. Students who answer an item correctly because of their testwiseness rather than their achievement of content, cause the item to discriminate improperly. As we have seen earlier, poor item discrimination contributes to lowered reliability estimates.

Time limits. It is customary for classroom achievement tests to be administered with generous time limits so that nearly all, if not all, students can finish. However, when time becomes a factor, when the test can be regarded as speeded, the result is a reliability coefficient that somewhat misrepresents score accuracy. The reliability estimate obtained under speeded conditions by the methods of internal analysis is artificially high, an artifact of the method itself.

Security precautions. Occurrences of cheating by students during a test contribute random errors to the test scores. Some students are able to provide correct answers for questions to which they actually do not know the answers. Copying of answers, use of cribs or cheat sheets, and the passing of information give unfair advantage to some and cause their

scores to be higher than they would be on retesting. The passing of information from class to class when the same test will be given to different classes at different times also reduces overall score reliability. The effect is to reduce artificially the differences between students and, consequently, to make real differences more difficult to detect accurately. Of course, similar outcomes will be observed if some students gain access to copies of the test prior to its use.

Learning Exercise B:

What, if any, effect is each of these actions likely to have on the reliability of the test scores?

1. Students who scored in the upper 10% on the last test need not take the test to be given tomorrow.
2. Students signed an "honor agreement" the day before the test, indicating that they would not cheat and that they would report any student whom they observed cheating.
3. The test was made extra long so that, at most, only about 10% of the students would be able to finish it in the hour available for testing.
4. The teacher reused a large number of items from previous classes and selected those items on the basis of how well they had discriminated the last time they were used.
5. The first 3 of the 25 items were written to ensure that no student would get any one of them wrong.

Answers:

1. Excusing the "best" students from the past test probably will cause the "new" group to be more homogeneous. Thus, reliability is likely to be lower than it would be had the entire class been tested.
2. The honor pact may help to reduce the amount of cheating that might have taken place. Reliability probably will be higher than it might have been had no pact been offered to students.
3. A test of any length that is as speeded as this one appears to be probably will yield an inflated value for the reliability estimate. In this case, the accuracy of the scores is *not* improved, but our estimate of the accuracy is.
4. The fact that the items are likely to be reasonably high in discrimination should cause the reliability of the scores to be fairly high. For this generalization to hold, we should expect the current class to be similar in achievement to past classes, and we should not expect the nature of instruction to have changed much.
5. The first three items will not contribute to high accuracy because such easy items do not discriminate levels of achievement. Reliability will be lower in this instance than it would have been if the three items were replaced with items that showed positive discrimination.

What If the Reliability Is Too Low?

Low reliability is symptomatic of an unhealthy testing situation just as high fever indicates unhealthy body tissue. We cannot tell in either case what the problem is, but the symptom suggests where to look. Was it the test, the nature of the examinees, or the testing conditions? Perhaps it was a combination. We need to determine a plausible explanation so that we can decide whether the scores can be used for their intended purpose. When reliability is questionable, we need to consider the various options available for using the scores appropriately.

Suppose the reliability of the scores (perhaps K-R21') from a unit achievement test turns out to be 0.33 and the teacher decides that this value is unsatisfactory. The scores have *some* value, but less than had been hoped originally. If it is practical to do so, the first priority should be to correct or improve the factors that contributed to the low reliability and then retest. A variety of reasons, notably test development and/or test administration time constraints, may preclude the retesting alternative. So if a decision is made to retain the scores, a subsequent decision might be to discount the scores, that is, to allocate less weight to them in decisionmaking than had been planned originally. If the scores were to count as 25% of the final grade, for example, their weight might be dropped to 20% or a bit less. Another way to discount or reduce the weight of a set of scores is to give a new, revised version of the test and then combine the scores from the two administrations. If the combined score is given a weight of 25%, then the first set of scores with the low reliability necessarily will have less weight than intended originally.

When a set of scores is discounted, as described above, decisionmaking can be affected in significant ways. If the discounted scores related to significant content for which it is particularly difficult to write test items, the validity of the decisions will be questionable. If the discounted scores related to instructional objectives that represented mostly higher order thinking skills, problem-solving, or application of important principles, subsequent decisions might be grossly misleading. In sum, the user must be cognizant of the effect of the discounting and weigh it in the trade-off with using relatively unreliable scores.

If the reliability coefficient for a set of achievement test scores is low, this means that the consistency of the scores is highly suspect. We should conclude that the scores are contaminated by errors of measurement if we have ruled out factors that might have an artificial impact on the magnitude of the number. If the reliability estimate indicates that we have obtained measurements that are error laden, it also signifies that we have not measured very well the important achievements we set out to measure. Thus, the scores should not be used with confidence to make the judgments about grades, placement, or need for remediation as we had hoped to make. Unless we have succeeded in measuring accurately, our measurements will be relatively useless for *any* purpose.

Summation

Test scores are used by teachers to gain information about students for making instructional decisions and for evaluating student progress. The quality of those decisions and the accuracy of those judgments relate closely to the dependability of the information on which they are based. Many of the decisions are important enough that we cannot afford to assume that the test scores are dependable or consistent enough. The extent to which measurement errors have infiltrated the scores needs to be estimated so that the teacher can decide whether full confidence can be placed in the scores.

Reliability is the term reserved by testing specialists to

refer to the consistency of a set of scores. And it is the scores that have error in them, not the test. Even expertly prepared tests produce imperfect scores because behaviors of test takers and conditions in the testing environment can cause test scores to be either higher or lower than they ought to be. So it is not the reliability of the test that we question, but it is the scores that contain excess baggage or deficiencies that distort their meanings.

The kinds of errors that interfere with our measurements can be classified as random or systematic. Random errors occur for some test takers and not others, or they occur in different degrees for all examinees in a group. Memory fluctuations, lucky or unlucky guessing, testwiseness, content sampling, fatigue or emotional strain, and subjective scoring are examples of factors that might contribute random errors to scores on teacher-made tests. Systematic errors could be caused by faulty test items, scoring abnormalities, or interruptions in the test administration. The random errors affect the relative ranking of students' scores, but the systematic errors simply increase or decrease all students' scores by the same amount without affecting the ordering of scores.

A number of methods have been devised to compute a reliability coefficient, a numerical index of the amount of error present in a set of test scores. A reliability coefficient can provide the kind of "hard" evidence that a teacher needs to decide if a set of scores is accurate enough to be useful for its original purpose. Of the many procedures available, the methods of internal analysis are the most practical ones for teachers to use in estimating the reliability of the scores from their tests. These methods include coefficient alpha and three Kuder-Richardson formulas. The numerical index that results from any one of these computational procedures varies from 0.00, indicating complete inconsistency, to +1.00, indicating perfect or error-free measurement.

Aspects of the test that tend to contribute to high reliability estimates are length, homogeneous item content, items of moderate difficulty, and items high in discrimination. High reliability can be expected if the group being tested is heterogeneous with respect to the content measured, is motivated to perform at its best, and is experienced in test taking. If testing conditions are established so that time limits are fairly generous, opportunities for cheating are eliminated, and distractions due to noise, temperature, and lighting are removed, score reliability will be enhanced.

Occasionally, a teacher may obtain scores that are not useful as originally intended; they should be replaced by retesting or should be given less weight in decisionmaking than originally intended. How high the reliability coefficient must be depends on situational factors: What other relevant information is available? How important is the decision? What factors probably cause the reliability estimate to be so low or so high? Because the test, certain characteristics of the examinees, and the testing conditions all can contribute concurrently to measurement error, how to interpret the reliability coefficient is no straightforward matter. For the same reasons, however, teachers cannot afford to use test scores blindly, as though the scores were infallible.

References

Ebel, R. L., & Frisbie, D. A. (1986). *Essentials of educational measurement* (4th ed.). Englewood Cliffs, NJ: Prentice-Hall.

Wilson, R. A., Downing, S. M., & Ebel, R. L. (1977). *An empirical adjustment of the Kuder-Richardson 21 reliability coefficient to better estimate the Kuder-Richardson 20 coefficient.* Unpublished manuscript.

Additional Readings

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* New York: Holt, Rinehart and Winston.

For those who desire greater theoretical depth about the basis for reliability, chapter 6 provides a thorough discussion. Methods of estimation are illustrated in chapter 7, and the appropriate formulas are derived in chapter 6. Both chapters conclude with sets of useful exercises.

Ebel, R. L., & Frisbie, D. A. (1986). *Essentials of educational measurement* (4th ed.). Englewood Cliffs, NJ: Prentice-Hall.

A section in the second chapter, "Kinds of Achievement Tests," provides background on the distinction between norm-referenced and content-referenced score interpretations and its implications. The relationship of this distinction to reliability and methods of estimating it is dealt with in chapter 5, "The Reliability of Educational Tests." This chapter also includes some theoretical foundation and some practice exercises for computing reliability coefficients.

Gronlund, N. E. (1985). *Measurement and evaluation in teaching* (5th ed.). New York: Macmillan.

Chapter 4 presents sections on the meaning of reliability and factors that influence reliability estimates, both of which complement this module well. A section in the chapter "How High Should Reliability Be?" describes an analytical approach to answering that question.

Nitko, A. J. (1983). *Educational tests and measurement: An introduction.* New York: Harcourt Brace Jovanovich.

Some basic theoretical notions are described in chapter 15, and a method for estimating scorer reliability for essays is illustrated.

Stanley, J. C. (1971). Reliability. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed.). Washington, DC: American Council on Education.

This chapter presents practical understandings and theoretical foundations of reliability. As one of the most comprehensive treatments in the literature, it is a classic reference for any issue related to reliability. Its date does not diminish its value as a source of information about reliability with respect to teacher-made norm-referenced tests.

Wilson, R. A., Downing, S. M., & Ebel, R. L. (1977). *An empirical adjustment of the Kuder-Richardson 21 reliability coefficient to better estimate the Kuder-Richardson 20 coefficient.* Unpublished manuscript.

The theoretical justification for the K-R21′ approach used in this module is provided, and the research study that supports its use is described.

Self-Test

Indicate whether each statement is true or false.

1. Any measurement errors that contribute to unreliability in test scores can be eliminated somehow.
2. Error-free test scores are perfectly reliable.
3. Most errors of measurement are the result of errors in scoring.
4. The operational definition of test reliability includes "equivalent forms" as an essential element.
5. Reliability is a name given to the characteristic of a test, and accuracy is a name given to a characteristic of a set of test scores.
6. The amount of error associated with a certain set of scores could be influenced by both the nature of the test items and the conditions under which the scores were obtained.
7. If Sarah takes the same test on two consecutive days

and receives the same score both times, it is appropriate to conclude that the test was reliable.

8. By our definition, a set of test scores would be considered perfectly reliable only if each examinee in a group obtained exactly the same score on two equivalent forms of the test.

9. In a group of seventh graders, the levels of test sophistication are more likely to be a source of random than systematic error.

10. If all the students in a class were bothered by a cramp in their little toes during an exam, the reliability of the test scores likely would be lowered as a result.

11. Loud noises from construction equipment outside a testing room are more likely to contribute systematic errors than random errors to the scores.

12. Because all students are free to make guesses for objective test items, guessing seldom influences the reliability of objective test scores.

13. The procedure used by a teacher for scoring essay items can contribute both systematic and random errors to the test scores.

14. A test item that was so difficult that no student answered it correctly would cause the test scores for that group to be less reliable than they ought to be.

15. Only tests composed of items susceptible to guessing will yield scores of less than perfect reliability.

16. It is easier to obtain high reliability in scores from a test if the items are homogeneous rather than heterogeneous in content.

17. To yield reliable scores, a test must be composed of items that vary widely in difficulty, from very difficult to very easy.

18. A 43-item test is likely to yield more reliable scores than a 55-item test when both are composed of comparable items and given to the same group.

19. If some of the poorly discriminating items in a test are replaced by similar items that are higher in discrimination, the reliability of the scores from the new test should be higher than that obtained from the original test.

20. The scores on a junior-high-level science test are likely to be more reliable for a group of seventh graders than for a combined group of seventh and eighth graders.

21. Scores obtained under conditions of examinee anonymity are likely to be less reliable than those obtained when examinees are identified by name.

22. A group of examinees that has had little test-taking experience should be expected to obtain less reliable scores than a group that is highly sophisticated in test taking.

23. For a classroom achievement test, a higher reliability estimate is likely to be obtained if the time limits are too strict rather than too generous.

24. If several students improve their scores by cheating during a test, the reliability of the scores of the whole class is likely to suffer.

25. The use of coefficient alpha to estimate reliability requires that dichotomous scoring (only 0 and 1) be used.

26. When computing coefficient alpha for a test composed of four 3-point essay items, the appropriate value of k to use is 12.

27. The larger the sum of the item variances in relation to the total score variance, the more reliable the test scores will be.

28. The K-R20 procedure is always appropriate to use whenever the K-R21 procedure is appropriate.

29. When computed on the same set of scores, K-R20 never will be smaller than K-R21.

30. The purpose of K-R21′ is to provide a direct estimate of K-R21.

31. To obtain reliability estimates using the methods of internal analysis, a correlation coefficient must be calculated.

32. The reliability of scores from a test composed of both essay and true-false items could be estimated by at least one of the methods of internal analysis.

33. For a 40-item multiple-choice test given to 30 students, reliability estimates could be obtained using either the K-R21′ method or coefficient alpha.

34. There is more error associated with scores yielding a reliability estimate of 0.72 than there is with scores yielding an estimate of 0.46.

35. Scores from classroom achievement tests tend to be about as reliable as those from standardized achievement tests when both are from the same group.

36. It is more likely for a reliability of 0.38 to be associated with scores from a short vocabulary quiz than with scores from a literature unit test.

37. If the reliability of the scores from a test is estimated to be 0.77, the difference between Wendy's score of 32 and Lisa's score of 37 is more likely due to *real* achievement differences than to measurement errors.

38. It would be reasonable for a teacher to discard the scores from a test if the K-R21′ obtained from the scores was only about 0.54.

39. A reliability estimate of 0.46 for a set of scores is a certain indication that the test is of low quality.

40. All sets of test scores have *some* value for decisionmaking, regardless of their level of reliability.

Self-Test Answer Key and Explanations

Letters following each explanation refer to the following subsections of the text: A = "Consistency of Measurements"; B = "Factors That Contribute to Inaccuracy"; C = "Methods of Estimating Reliability"; D = "Interpreting Reliability Coefficients"; and E = "Factors That Affect Reliability Estimates" (all may be found in the section titled "Test Score Reliability"). I = introduction.

1. **F** To eliminate measurement errors we must be able to control the factors that cause them. Though we can curb some of these errors successfully, we cannot eliminate errors due to accidental incidents in the testing situation, guessing

correctly by some examinees, or sampling of items from the universe of all relevant test items. (A)

2. **T** Though error-free scores are not practically feasible, statements such as these describe the meaning of the term "reliability." (I)

3. **F** Many factors associated with the examinee and with the testing conditions contribute to measurement error. Scoring errors represent only one such factor. (B)

4. **T** The operational definition includes the correlation between scores on equivalent tests as the essence of obtaining a reliability coefficient. (A)

5. **F** Reliability and accuracy are interchangeable terms that describe a set of test scores. Reliability is not a constant characteristic of a test, but one that varies with the nature of the group to which it is given and the conditions under which it is given. The errors associated with reliability are in the scores, not in the test. (A)

6. **T** The eight examples given in this section show that test item content, testing conditions, and conditions within individual examinees concurrently affect the amount and type of error that might affect the scores. (B)

7. **F** First, having information about the consistency of only Sarah's scores is insufficient for making decisions about reliability. Second, we can draw conclusions about the reliability of the scores, but not about the test itself. Finally, it should be recognized that a retest need not yield exactly the same score for each person in order for us to make high claims about reliability. As long as examinees are in the same order (or nearly so), the correlation between scores will be high. (A)

8. **F** The definition indicates that a correlation coefficient will be used to assess reliability. A perfect correlation, 1.00, could be obtained if each person's scores were the same both times *or* if the relative order of each individual's scores were the same both times. (A)

9. **T** It is reasonable to expect considerable variability in test-taking skills within a group of seventh graders. This variation will permit some students to get higher scores than others simply because of a higher level of testwiseness. These would be random rather than systematic errors in that all examinees' scores would not be affected in the same way. (B)

10. **T** Even though this ailment affected everyone, it is reasonable to presume that students varied in their ability to cope with such a distraction. Therefore, it is reasonable to label the potential effect as random rather than systematic error. (B)

11. **F** See the explanation for Item 10. (B)

12. **F** It is not the opportunity to guess but rather the act of guessing that influences score reliability. As long as students vary in their chance success when they guess, random rather than systematic errors are likely to occur. (B)

13. **T** Example 7 in the text describes how random errors from essay scoring might occur, and the third illustration describing systematic errors relates to essay scoring. (B)

14. **T** Such a test item does not provide information that helps to differentiate levels of achievement: It does not discriminate. Consequently, had a more discriminating item been used, the reliability would have been higher. (E)

15. **F** Several illustrations in this section show that guessing on objective items is only one of many potential sources of measurement error. (B)

16. **T** The methods of internal analysis recommended for computing reliability coefficients are based on this notion. If the content measured by a test is too diverse, there is a good chance that some of what is being measured is extraneous or unrelated ability. (E)

17. **F** Test items that are moderate in difficulty are more likely to be good discriminators than items that are extreme in difficulty. And tests composed of items that discriminate will yield the most reliable scores. (E)

18. **F** Generally, the longer a test is, the more reliable its scores will be. (E)

19. **T** When the average discrimination level of the items in a test is improved, more reliable scores should be obtained when the revised test is administered to a comparable group. (E)

20. **F** Because the group of seventh graders is likely to be more homogeneous in their science achievement than the combined group, the reliability of their scores probably will be lower. When the real differences in achievement among individuals are small, those differences will be difficult for test scores to reflect with high accuracy. (E)

21. **T** Under conditions of anonymity, students are less apt to be motivated to perform at their best. When the consequences for examinees are the same, regardless of their scores, students probably will vary in both their enthusiasm for completing the test and their desire to maximize their scores. (E)

22. **F** Only when individuals in the group vary with respect to testwiseness should we expect score reliability to be affected. When examinees are at the same low level of sophistication, any errors that affect scores are likely to be systematic rather than random. (E)

23. **T** When the test is speeded, the reliability estimate could be artificially higher than it really ought to be. Generous time limits provide conditions for obtaining reliability estimates that reflect the actual level of score accuracy better. (E)

24. **T** The score improvement due to cheating will introduce random errors that would be absent if the test were given under more secure conditions. All scores are used to estimate reliability, and the estimate is for the entire set of scores, not just for some that may be affected by certain kinds of errors. (E)

25. **F** The restriction of dichotomous scoring applies only to the Kuder-Richardson methods. Alpha is the only choice (among methods described in this module) when essays or problems with varying point values are included in the test. (C)

26. **F** There are four test items and, therefore, the value of k should be four. The maximum number of points does not enter into the computation of alpha directly. (C)

27. **F** Equation 1, the one used to compute alpha, should be examined to help answer this question. The sum of the item variances (ΣS_i^2) is associated with errors of measurement. As it increases in size, the fraction in which it is the numerator also increases. When that fraction is subtracted from one, the reliability coefficient results. The larger the fraction, then, the smaller the coefficient will be. (C)

28. **T** The use of K-R21 assumes that all items in the test are equal in difficulty, but the use of K-R20 does not require this condition. Otherwise, the two methods are equally appropriate. (C)

29. **T** K-R20 equals K-R21 only when all items are equal

in difficulty. When the item difficulties vary, K-R21 will always be smaller than K-R20. (C)

30. **F** Both K-R21 and K-R21′ are intended to be estimates of K-R20. K-R21′ is expected to be a closer estimate most of the time. (C)

31. **F** None of the equations shown in this section include a correlation coefficient as a variable to be used in computing the reliability coefficient. The reliability coefficient that results, however, is interpreted like a correlation coefficient. (C)

32. **T** Coefficient alpha, one of the methods of internal analysis, is appropriate to use when both objective and essay items are included in the test. (C)

33. **T** If we assume that dichotomous scoring is required for this multiple-choice test, then either procedure is appropriate. Since K-R21′ is an *estimate* of K-R20 (or alpha), it would be preferable to estimate reliability with alpha. (C)

34. **F** The amount of error associated with the scores decreases as the size of the reliability coefficient increases. When reliability is at its maximum, 1.00, there is no error included in the scores. (D)

35. **F** Published tests tend to have very high reliability estimates associated with them because extensive efforts have been made to develop high-quality items and because large heterogeneous groups usually are used to obtain the estimates. Scores from teacher-made tests tend to be much less reliable for essentially opposite reasons. (D)

36. **T** Because the literature unit test is probably longer and because its items may be more carefully developed, it is reasonable to expect its scores to be more reliable than those from a quiz. (E)

37. **T** The reliability coefficient is high enough that, even though we know that measurement errors exist, errors do not dominate these scores. The score difference seems too large to be explained only by measurement error when the reliability is 0.77. (D)

38. **F** A K-R21′ estimate of 0.54 is about what we usually observe for scores on teacher-made tests. Without knowing more about how the scores are to be used, a definitive response cannot be made. Whatever the purpose, however, scores with that high a reliability ought to have *some* value. (D)

39. **F** Factors beyond the test itself can cause reliability estimates to be as low as 0.46. The nature of the examinees (very homogeneous) or unusual testing conditions (too hot, too noisy) can cause the scores from a "good test" to be low in reliability. (E)

40. **F** Some scores are so dominated by errors that they are useless for their original purpose. Scores with low reliability misrepresent examinees' ability levels as well as the differences that appear to exist among examinees. (D)