An NCME Instructional Module on

Understanding Reliability

Ross E. Traub, *Ontario Institute for Studies in Education*
Glenn L. Rowley, *Monash University*

The topic of test reliability is about the relative consistency of test scores and other educational and psychological measurements. In this module, the idea of consistency is illustrated with reference to two sets of test scores. A mathematical model is developed to explain both relative consistency and relative inconsistency of measurements. A means of indexing reliability is derived using the model. Practical methods of estimating reliability indices are considered, together with factors that influence the reliability index of a set of measurements and the interpretation that can be made of that index.

No measurement is perfect. For some measurements, a source of imperfection is obvious, as when we use a bathroom scale that has been calibrated in two-pound intervals. Even if the scales worked perfectly, the best we could hope for is a reading within a pound or two of the correct weight. But if the scales were poorly calibrated or the inner mechanism were faulty, the error could be considerably greater.

Measuring procedures with finely graded or finely calibrated scales foster the impression that the measurements obtained using the procedures will be very precise. But repeated application of such a measuring procedure to the same person or object may reveal quite startling fluctuations—for example, bodily measurements such as heart rate or blood pressure. Inconsistent measurements are a bane to persons engaged in research. Scientists have learned to repeat their measures several times when it is important to obtain results in which they can be confident. The average of a set of repeated measurements provides a more precise estimate of what is being measured than does a single measurement, and, as a bonus, the amount of variation in the repeated measurements shows how inconsistent the numbers produced by the measuring procedure can be. Unfortunately, the measuring procedures we use in education usually cannot be repeated as easily as can some of the measuring procedures used in the physical sciences.

The foregoing discussion is not to deny that test scores often appear to possess great precision. With a test that is scored objectively, we may easily convince ourselves that a score of 49 means that the individual answered precisely 49 questions correctly, not 48 and not 50. It is true that as a count of correct answers the test score contains no error. But as a measure of some ability possessed by the examinee, we cannot be so confident. Can we be sure that another person of the same ability would have obtained exactly the same score? Can we even be sure that the same person would have obtained the identical score had the test been administered on a different occasion or under different circumstances? We cannot, and it is for this reason that we must turn our attention to the question of error in test scores—error in the sense of uncertainty, not error in the sense of a mistake. How can we interpret a test score that provides uncertain information about a person, and how can we obtain information about the degree of uncertainty to attach to test scores? These are the questions that will be addressed in this module.

The Concept of Reliability

In everyday life, the concept of reliability is closely associated with the idea of consistency. Let us consider several familiar examples. An automobile is a reliable starter if its motor invariably starts at the turn of the ignition key. An employee is reliable if she does all the things expected of her on the job. A vacation guide is reliable if it contains information that is consistent with the experiences of the traveller. Notice that the attribution of reliability in these examples can be made only after several—perhaps a great many—repetitions of some chain of events or behaviors. Thus, a car will not be judged a reliable starter until it has been started successfully many times over a period of months. To be thought of as reliable, an employee must perform her duties faithfully each day for several months, perhaps years.

These examples may suggest that the term *reliable* applies only when an event or behavior occurs with perfect consistency or predictability. But this is not so. A baseball player who makes a hit on average only once for every three times at bat will nevertheless be called a good (that is, a reliable) hitter. Clearly, reliability is not necessarily an all-or-nothing concept. There can exist degrees of reliability such that one person or object is judged more reliable than another. For example, a typist who makes five errors on average in each 100 words typed will be judged more reliable than the typist who makes ten errors on average in the same number of words typed, even though neither person is perfectly reliable in the sense of producing error-free typing. Similarly, some cars may be more reliable starters than other cars. If there is a difference in the starting reliability of two cars, it can be discovered through

Ross E. Traub is a professor in the Department of Measurement, Evaluation, and Computer Applications at the Ontario Institute for Studies in Education, 252 Bloor St. W, Toronto, Ontario, Canada, M5S 1V6.

Glenn L. Rowley is a professor of education at Monash University, Clayton, Victoria, Australia, 3168.

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.

Table 1
Examinee Scores on Two Forms of a Thirty-Item Multiple-Choice Test of Mathematics

Examinee I.D.	Form A	Form B
1058	27	16
51	26	26
57	26	26
1084	26	26
559	25	26
2	25	25
1132	25	23
266	24	28
1051	24	26
1050	24	21
71	24	15
1079	23	22
.	.	.
.	.	.
.	.	.
438	7	8
413	7	7
45	7	6
286	7	6
1007	6	13
455	6	12
598	6	12
1147	6	12
139	6	8
524	6	8
180	6	4
564	5	16

repeated attempts, made under similar circumstances, to start both vehicles.

The concept of reliability is also used to describe and assess the scores that persons achieve on educational tests. In this application, reliability carries some of the connotations of its use in everyday speech. To make the connection clear, it will be helpful to illustrate the meaning of consistency of educational measurement, and then suggest a way of indexing it.

Table 1 contains a selection of the scores obtained when two 30-item multiple-choice mathematics tests were administered to 199 eighth grade students. Math Test A and Math Test B were designed to measure the same achievements and to be equally difficult. (Such tests are usually referred to as alternate forms.) We note that many students, perhaps most, gained quite similar scores on the two tests, but that a small number increased their scores dramatically (e.g., student #564, who scored 5 on Math Test A and 16 on Math Test B). Others did less well on the second testing (e.g., student #1058, who scored 27 on Test A but dropped to 16 on Test B). If we are to regard the two tests as measuring the same achievement, we must concede that there are, at least occasionally, substantial errors of measurement. But giving all our attention to cases such as these two may create the impression that there is no consistency at all in the test scores. This would be wrong, too. Because the data in Table 1 have been ordered from high to low on Math Test A, we are able to recognize that those who scored high on Test A tended also to score high on Test B, and likewise for those who scored low. So, instead of asking "Are the scores from the two tests consistent?" or "Are there errors of measurement?" we need to ask "How consistent are the scores from the two tests?" or, conversely, "How much error or inconsistency do the scores contain?" These questions may be addressed graphically or by computing the appropriate summary statistics.

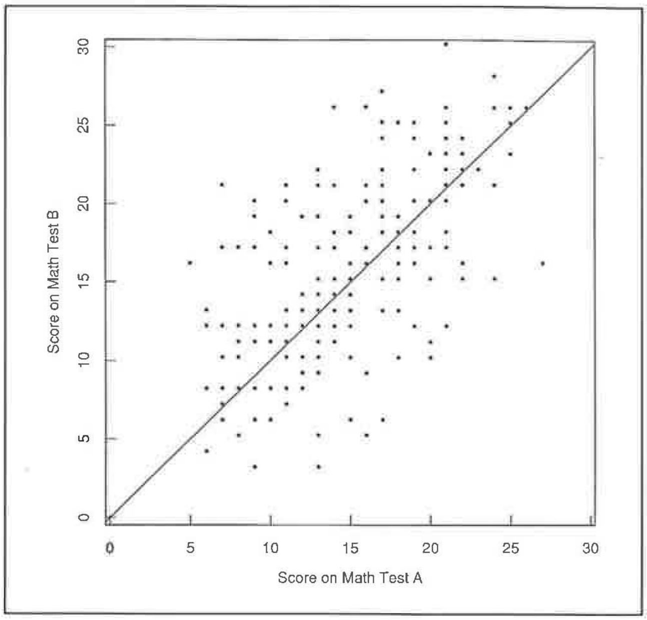

FIGURE 1. *A plot of scores on Math Test A against scores on Math Test B for 199 eighth-grade students*

Graphical Treatment

Figures 1 and 2 illustrate the notion of consistency of educational measurements. In Figure 1 the scores that each of the 199 students achieved on Math Test A are plotted against the scores achieved by the same students on Math Test B. If each

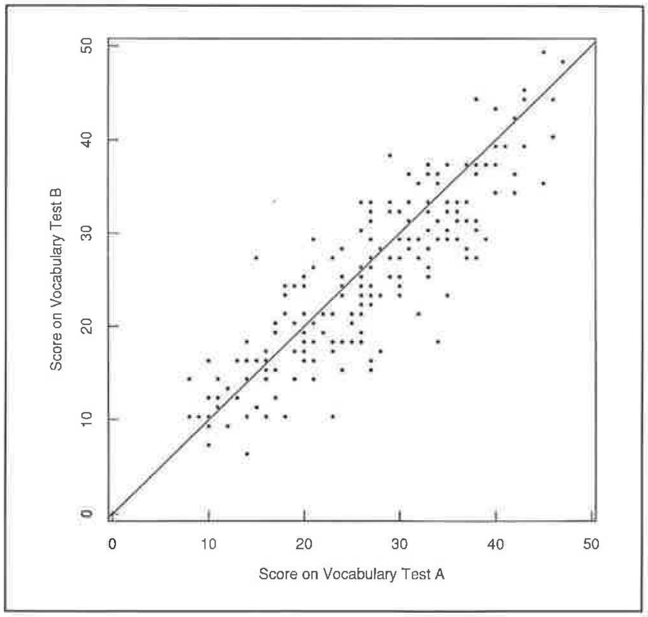

FIGURE 2. *A plot of scores on Vocabulary Test A against scores on Vocabulary Test B for 199 eighth-grade students*

student had achieved identical scores on the two tests, the plotted points would all lie on the diagonal line through the origin with a 45-degree slope. This line defines perfect consistency of measurement—every student receiving the same score on each measurement occasion. The fact that the plotted points in Figure 1 scatter about this straight line is evidence of unreliability or inconsistency in measuring the mathematics achievement of these students. The degree to which the points scatter is an indication of the amount of unreliability or inconsistency in the scores. We note that one effect of measurement error is that the two tests will rank the students differently. The two students referred to previously (#1058 and #564) can be identified on the scatter plot, and the alert reader will note that in the complete data set there are students whose scores changed by even greater amounts.

Figure 2 is a plot of the scores that the same 199 students achieved on two multiple-choice tests of vocabulary. Here again the forty-five degree line through the origin has been drawn on the figure. Notice that this line is not ideal for representing the trend in this plot. Scores on Vocabulary Test A were, on average, about two points larger than scores on Vocabulary Test B, so more of the plotted points in Figure 2 lie below the 45 degree line than above it. Although it is possible that these two tests measure the same characteristic, one (Test B) is more difficult that the other. This suggests a less stringent notion of perfect consistency in the measurements provided by two tests. We have perfect *relative* consistency when there exists a general rule for obtaining any examinee's score on one test from his or her score on the other test. This includes all situations in which the plot of scores on two tests lies entirely on a line. It is not necessary that the line be straight, but for the argument to be made that the tests measure the same characteristic, albeit in different ways, it must be true that as the scores on one test increase, the scores on the other test also increase. Usually we limit our attention to situations in which the line of relationship defined by the scores on two tests is straight.

The data plotted in Figures 1 and 2 give evidence of a particular kind, namely, evidence of degree of consistency in the scores achieved on two tests that were intended to measure the same characteristic and that were administered to the same group of examinees on two different occasions. Several other kinds of consistency may be of interest. Examples include consistency in the scores on just one form of a test, which has been administered on two different occasions to the same group of examinees, and consistency in the scores on two different forms of a test, which have been administered on the same occasion. Interest also extends, for example, beyond test scores per se to the consistency with which different markers grade the same set of essays, or the consistency with which different judges rate the performances of different figure skaters in Olympic competition.

By visual inspection alone, we can readily conclude that the scores shown in Figure 2 are more consistent (hence more reliable, containing less error) than those in Figure 1. In sections to follow, we will look at ways of indexing the degree of reliability and the amount of measurement error in sets of data like these.

Formalization

The notion that is called classical reliability can be developed in a formal sense by concentrating first on the measurement of one characteristic of just one person. To be concrete, let us suppose that the person is an eighth grade student and the characteristic of interest is achievement in mathematics. Further, let us imagine that this student can be retested many times with the same instrument, and that the student's mathematics achievement is *not* affected by the process. (There are, of course, many reasons why frequent retesting cannot be done in practice. People learn from writing tests, they remember responses previously given, and they become tired and even unwilling. None of this matters. What we are doing is setting up a model; any one testing occasion will be regarded as one of the many that could possibly have taken place. Our aim is to investigate the implications of the model, and then apply them to the data from only one or perhaps two testing occasions.) We would not expect all the scores resulting from this repeated testing experiment to be the same, just as we would not expect the repeated measurements that could be made of the length of a table, using a tape-measure that has a suitably fine scale, to be all the same. In both cases, any variation in the numbers—scores, measurements—is presumed to be due to errors of measurement. (There are many sources for errors in tests scores, for example, inconsistent behavior on the part of examinees, mistakes in marking, different interpretations of examinees' responses, changes in an examinee's mood, motivation and mental alertness, guessing, and even sheer luck in the choice of questions for the test.)

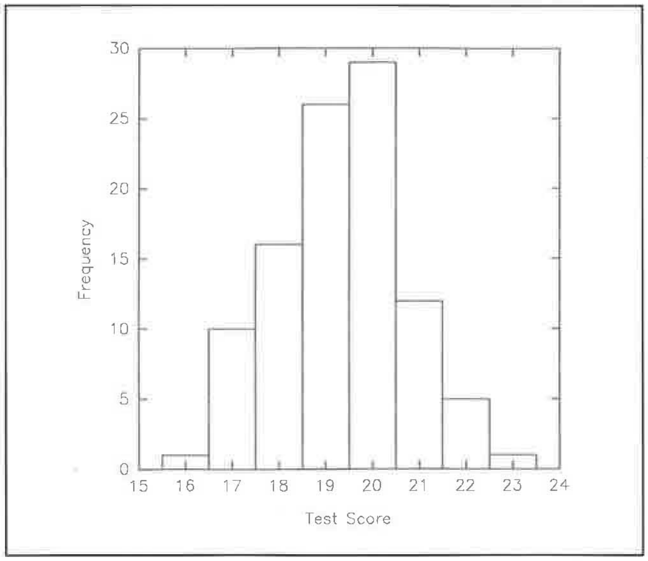

FIGURE 3. *Histogram of 100 test scores for one (hypothetical) examinee*

Further study of the repeated measurements of a single characteristic of one person can be facilitated by a bar graph or histogram, such as that in Figure 3. We see that the depicted frequency distribution of repeated measurements has a central point where the frequency (or density) of scores is greatest. We also see that the scores spread themselves over a portion of the scale of measurement. If Figure 3 represents the results of an experiment in which the mathematics achievement of a student is tested repeatedly, we can take the central point defined by the arithmetic mean of the distribution to be the score that best represents the student's mathematics achievement. By convention, we refer to this central point as the student's true score. (With reference to Figure 3, the true score of the examinee is estimated to be 19.3, the mean of this distribution. The standard error of measurement for these scores is estimated to be 1.4, the standard deviation of this distribution.)

The term *true score* can easily be misinterpreted. It is not intended to convey any notion of absolute truth, of what a person is "really worth" on a test. It means no more and no less than what is stated in the definition: the person's average score over repeated (necessarily hypothetical) testings. A person with well-developed test-taking skills may score so consistently high on a test that we believe the scores overestimate his or her achievement. Nevertheless, such a person will have a high true score. The overestimation is not, in this formulation, seen as measurement error. Instead, it is seen as bias due to test-taking skills, which enhance the person's test scores but which, we believe, are unrelated to the ability the test is intended to measure. It is only scatter about the true score, not bias in the true score itself, that is presumed due to errors of measurement.

The smaller the variance or standard deviation, the smaller the effect that errors of measurement have, in general, on test scores.

A useful index of the size or extent of these errors is the variance or its square root, the standard deviation, of the scores in the distribution of repeated measurements. The smaller the variance or standard deviation, the smaller the effect that errors of measurement have, in general, on the test scores. Given two tests of the same characteristic, each yielding scores on the same scale, the test with a standard deviation of errors that is smaller provides greater consistency of measurement than the test with a standard deviation of errors that is larger. (The standard deviation of repeated measurements is called the standard error of measurement.)

The nature of educational testing (and most psychological testing) is such that repeated measurements of the same person are usually impossible to obtain. The concept of reliability has been developed in the more realistic context in which many individuals are measured on the same characteristic, rather than the situation where a characteristic of one person is measured repeatedly. Instead of learning about the distribution of measurement errors for one person, what we must do is study the test behavior of many persons, thereby trying to learn about the average size of the measurement errors across the group of persons. In principle, we are able to do this, provided we have a minimum of two measurements for each person, these having been obtained using more-or-less equivalent tests or procedures. The difference between the two measurements for a person gives us an indication of the size of the measurement error for that individual. By itself, one such difference for one examinee is not good information, for by chance it may be much larger or much smaller than the average of many separate determinations of the difference. The differences between two more-or-less equivalent measurements of each member of a relatively large group of examinees provide us with good information about the distribution of measurement errors across the group (even though this information does not necessarily reflect how differences in a large number of repeated testings of a single individual would be distributed). The strategy of testing a group of examinees two or more times is one that we will follow in order to learn about measurement error.

Suppose now that everyone in a well-defined population of students is measured just once with the same test instrument. All the scores so obtained can be assembled into a frequency distribution. The question is how to interpret the scores in this distribution using the concepts of true score and error of measurement. Two extreme situations can be imagined. In the first, all the scores in the distribution are considered to be true scores, with no errors of measurement. In such a case as this, all differences in observed scores could be expected to recur in another administration of the test. And all the variance in the scores is true-score variance. The second extreme situation is that in which all examinees have identically the same true score; hence, all observed-score variation is due to variation in errors of measurement. In this situation, we could not expect the differences in observed scores to be maintained in the scores from another administration of the test.

Neither of these extreme situations is realistic for scores on educational and psychological measures. We expect such scores to include a true-score component—how else could we explain the relatively high degree of consistency apparent in Figure 1 in the way examinees are ordered by their scores on the two math tests? But, we also expect the observed scores to include an error component—how else could we explain why the examinees in Figure 1 are not ordered in exactly the same way by the scores they obtained on the two math tests? Thus, we expect the variation in any set of observed test scores to be due partly to variation in true scores and partly to variation in error scores. But we cannot tell, just by studying the frequency distribution of observed scores from one administration of a test, how much variation is attributable to each. Classical reliability theory provides us with ways of estimating these amounts.

The theory of classical reliability begins with the proposition that any time a person takes a test, the observed score X may be broken up into two components: the true score T (as defined previously) and an error score E, which is the difference between the observed and true scores. Formally, we write

$$X = T + E \tag{1}$$

Over the many possible testing occasions we imagine are possible in the repeated testing experiment described earlier—only one of these occasions usually occurs in reality—T is the same for an individual over all occasions, but we accept that X is likely to fluctuate, depending on which testing occasion actually is chosen for the individual. We attribute this variation in X to fluctuations in the error score, E. If we assume that in the long run observed scores will scatter symmetrically on either side of T, then the error scores will average out to zero, whether the true score be high or low. In statistical terms, a consequence of the way in which true and error scores are defined is that error scores will be uncorrelated with true scores. Because this is so, it follows that the variance of the observed scores for a set of examinees will consist simply of the sum of two separate and uncorrelated variances, one due to true scores and the other due to errors of measurement. Symbolically,

$$\sigma_X^2 = \sigma_T^2 + \mathrm{Ave}\,(\sigma_E^2) \tag{2}$$

where

σ_X^2 is the variance of the observed scores,
σ_T^2 is the variance of the true scores, and
$\mathrm{Ave}(\sigma_E^2)$ is the variance of the error score distribution for each examinee, averaged over all the examinees who were tested. Hereafter, $\mathrm{Ave}(\sigma_E^2)$ is simply denoted by σ_E^2.

Thus an observed-score variance of 100 might result from a true-score variance of 90 and an average error variance of 10. But it could also result from a true-score variance of 40 and an average error variance of 60. We need classical reliability theory to help us distinguish between situations as fundamentally different as these.

The above expression for the relationship among the observed, true, and error variances for an educational measure makes plain the fact that for fixed, observed-score variance, any increase or decrease in true variance is accompanied by a corresponding decrease or increase in error variance. It is also clear from this relationship that if the variances of the errors associated with the measurements of different examinees are approximately equal, in which case σ_E^2 is approximately the same for any group of examinees, then any difference in observed score variance for different groups of examinees must be due to differences between the groups in true score variance. This means that the "range of talent" in the group of examinees being tested affects the relative magnitude of the true-score variance to the error-score variance. Group heterogeneity on the characteristic being measured is an important factor to be considered in assessing the reliability of a test, a point to which we will return later in this module.

Indexing Reliability

A natural way to think of indexing reliability, given the relationship among variances expressed in Equation (2), is as a ratio, specifically the ratio of true-score variance to observed-score variance. This ratio is near one when most of the observed-score variance is attributable to true scores; it is near zero when the true variance is small relative to the observed variance. The ratio of true-score variance to observed-score variance is one way of defining the concept of reliability for educational and psychological tests. Here and subsequently, the symbol $\rho_{xx'}$ is used to denote the reliability coefficient. Symbolically, $\rho_{xx'}$ is defined as follows:

$$\rho_{xx'} = \sigma^2_T/\sigma^2_X \qquad (3)$$

Because of the additive relationship between true- and error-score variances (they sum, as noted previously, to the observed-score variance), it can easily be shown that this ratio is identically equal to one less the proportion of observed-score variance attributable to error variance:

$$\rho_{xx'} = 1 - (\sigma^2_E/\sigma^2_X) \qquad (4)$$

The concept of reliability defined and indexed in this way has several useful properties:

- It is a dimensionless number (i.e., it has no units).
- The maximum value of the coefficient of reliability is one, when all the variance of observed scores is attributable to true scores.
- The minimum value of the coefficient is zero, when there is no true-score variance and all the variance of observed scores is attributable to errors of measurement.
- In practice, any test that we may use will yield scores for which the reliability coefficient is between zero and one; the greater the reliability of the scores, the closer to one the associated reliability coefficient will be.

It is common for test users and developers to see reliability as an important property of the scores examinees attain on a test, and to see the reliability coefficient as a vital indicator of test-score quality. It would be rare for publishers of tests not to provide data on reliability in their test manuals, especially if they aspire to any degree of respectability for their tests. But, while accepting that reliability is an important property, we should not imagine that a high reliability coefficient alone is sufficient to demonstrate the high quality of a set of test scores. A test that yields highly reliable scores may measure abilities that are not considered important, and the test scores may be interpreted incorrectly or used for inappropriate purposes. These issues are addressed at length in measurement texts (e.g., Allen & Yen, 1979; Crocker & Algina, 1986). Moreover, it must be stressed that reliability is not simply a function of the test. It is an indicator of the quality of a set of test scores; hence, reliability is dependent on characteristics of the group of examinees who take the test, in addition to being dependent on characteristics of the test and the test administration. These matters will be discussed in more detail presently.

Estimating Reliability

As we have defined reliability to this point, it consists of a ratio of two quantities, one of which—the observed-score variance—can be computed easily whenever a test is administered, and the other of which—the true-score variance—cannot be computed directly because true scores are unobservable. (Throughout, we assume that a sufficiently large sample of examinees is tested so that the effect of statistical sampling error on an estimate of variance can be safely ignored.) To obtain estimates of the unobservable quantities, the true-score variance (σ^2_T) and average error variance (σ^2_E), it is necessary once again to appeal to the notion of repeated measurement.

Suppose that two forms of a test are administered on different occasions to a large sample of examinees. (Later we will consider other ways in which repeated measurements can be obtained.) We need to make certain assumptions about the two tests. Earlier it was indicated that they should measure the same ability. Formally, we will assume that any examinee that we choose will have identical true scores on the two tests, or that, if true scores on the two tests differ, the relationship between them will be of a particular type known as a linear relationship. This means that the graph formed by plotting examinees' true scores on one test against their true scores on the other test will be a straight line, never a curve. True scores on the two tests, therefore, should rank the examinees in exactly the same way. If this is the case, and if errors of measurements are truly random and therefore independent of each other and of true scores, it can be shown by some reasonably simple algebra, that the unobservable quantity

$$\rho_{xx'} = \sigma^2_T/\sigma^2_X$$

is exactly equal to the coefficient of correlation between scores from the two tests. The proof of this result is not given here, but interested readers may turn to a measurement text such as that by Allen and Yen (1979) or Crocker and Algina (1986), where the result is demonstrated.

How realistic are the assumptions of the previous paragraph? In practice, we can never know that true scores on two tests are either exactly equal or perfectly linearly related, since we can never know true scores. But we can construct tests so as to make it likely that examinees' true scores on the tests are equal or nearly so. For example, we can write items in matched pairs, where each item in a pair resembles the other greatly, both in the skill or knowledge tested and in the level of difficulty of the tasks presented. Tests constructed in this way are referred to as alternate forms. The aim in constructing alternate forms is that "it shouldn't matter which one we use," and that they should measure the same ability at the same level of difficulty. If this aim were achieved exactly, we could refer to the two forms as parallel. Parallel test forms satisfy the assumptions outlined earlier—in particular, that any person will have the same true score and the same variance of measurement errors on either test form. In practice, the best we can do is construct alternate forms with sufficient care that we are willing to believe they are parallel, or very nearly so. If they were exactly parallel, we would be able to calculate the value of the reliability coefficient; because the best we can say is that we believe the forms to be approximately parallel, we should regard the correlation between scores from alternate test forms as an approximation to the reliability coefficient. The closer the two forms are to being truly parallel, the better the approximation will be. Although we cannot assert categorically that the two mathematics tests used to obtain the data plotted in Figure 1 are parallel, if we treat them as parallel, the scores of the 199 students on these tests yield an estimate of reliability of 0.63. The corresponding estimate of reliability for the vocabulary test scores plotted in Figure 2 is 0.86, corroborating our earlier suggestion that the vocabulary test scores appear more reliable than the math test scores.

What can be done to estimate reliability if there is only one form of a test? One possibility is to administer the test twice, with an interval of time between the two administrations. This approach has obvious drawbacks—examinees may remember questions and learn the answers in the interval between administrations, in which case their true scores will be different from one administration to the other; they may remember the answers they gave to questions on the first administration and repeat them, right or wrong, on the second administration, in which case the errors of measurement on the two administrations will not be independent in the way they are assumed to be.

Despite the foregoing drawbacks, a second administration is the only way to obtain the repeated measurements needed to estimate reliability when a test consists of only one task or exercise, such as a test that gave examinees the task of writing a single essay. But if the test is composed of a number of parts or items, as is the typical multiple-choice test and the essay test that contains several separate questions, then the index known as Coefficient Alpha can be computed. In effect, the separate parts of the test serve as repeated measures, and the interrelationships among scores on these parts provide information about reliability. (If the items of the test are scored correct/incorrect, as in the typical multiple-choice test, a special form of Coefficient Alpha known as the Kuder-Richardson Formula 20 or KR-20 index can be computed.) For further information about Coefficient Alpha or the KR-20 index, the reader is directed to a measurement text, such as the aforementioned book by Allen and Yen (1979) or that by Crocker and Algina (1986).

Reliability is dependent on characteristics of the test, the conditions of administration, and the group of examinees.

Interpreting Reliability Coefficients

The formulas developed in classical reliability theory lend themselves to several different but complementary interpretations of the reliability coefficient. First, note that the reliability coefficient is, by definition, equal to the proportion of observed-score variance that is attributable to true scores. Therefore reliability may be thought of as true, or systematic, variance. If we estimate the reliability of a test to be 0.91, for example, we may interpret this as telling us that an estimated 91 percent of the observed variance in scores is due to systematic differences in examinee performance, the remainder to chance differences. But also, we note from Equation (4) that the reliability coefficient is one less the proportion of error variance. A relatively high reliability coefficient, therefore, indicates a relative lack of error variance. For the test just described, we may conclude that a proportion of (1 − 0.91), or 9 percent, of the observed variance is due to measurement error. Finally, since the reliability coefficient has also been shown to equal the correlation between parallel measures, we may think of reliability in terms of correlation. A reliability coefficient of 0.91 tells us that the scores from the test could be expected to correlate 0.91 with the scores from a parallel test.

The size of the estimated reliability coefficient for a set of test scores will depend on the sources of errors that potentially can affect the test scores. Which sources of errors these are depends on the way the test scores are obtained. We would expect, for example, the estimate of reliability obtained by correlating scores for the same test administered on two occasions separated by a week to be larger than the estimate obtained by correlating scores from alternate forms administered on occasions separated by two months. The possibility of score inconsistencies due to changes in examinee ability and motivation, changes that vary randomly in size from one examinee to another, is more likely over a two-month interval than over a week. Also, the alternate forms situation will include score inconsistencies, if they arise, due to the different items comprising the forms. The latter source of inconsistency cannot affect scores when the same test is administered on both occasions. In studying the information in test manuals, it is important to note not only the size of the reliability coefficients reported, but also the type of estimate reported, the kinds of error that it acknowledges, and the population of examinees that was sampled.

More recent developments in the theory of reliability, known as generalizability theory, have shown us how to think about multiple sources of error and how to design the gathering of data so as to permit us to estimate the separate contributions of several sources of error to test scores. Generalizability theory may be seen as an extension of classical reliability theory that recognizes and distinguishes among the many potential sources of error in test scores or other data about the behavior of individuals. Alternatively, classical reliability theory may be seen as a special case of the more general theory, which is applicable when you do not wish to distinguish among the various sources of measurement error. More information about generalizability theory can be obtained from a reading of Brennan (1983), Brennan and Kane (1979), Crocker and Algina (1986), and Feldt and Brennan (1989).

What Makes a Test Reliable?

This is actually the wrong question, since a test by itself is neither reliable nor unreliable. When a test is used to assign scores to individuals, the scores that are obtained may be reliable or they may be unreliable; it is the scores that have the property of reliability, and not the test itself. Nevertheless, we often need to ask ourselves the following question: Under what circumstances do tests produce reliable scores? We may consider this question by turning our attention, first, to the test itself, second, to the conditions under which the test is administered, and, third, to the group of examinees being tested. It is the interaction among all three of these factors that determines the reliability of a test.

The Test

Test length. Generally speaking, longer tests yield more reliable scores than shorter tests. Any test may be thought of as a sample of tasks, and if the sample is too small, chance in the selection of tasks will play too great a part in determining the scores that students obtain.

The relationship between reliability and test length has been shown to follow a simple mathematical relationship, described by the Spearman-Brown formula:

$$\rho_{nn'} = n\rho_{11'}/[1 + (n - 1)\rho_{11'}] \tag{5}$$

where

$\rho_{11'}$ is the reliability of the original test,
$\rho_{nn'}$ is the reliability of the lengthened test, and
n is the factor by which the test is lengthened.

To see how this formula works, consider the situation in which a classroom teacher has administered a 30-item test to a class and estimates the reliability of the test to be 0.72. She wonders whether it would be worth lengthening the test to obtain greater reliability. Suppose that the school scheduling is such that doubling the number of items is the maximum that she could consider. Putting n equal to 2 in the Spearman-Brown formula, she arrives at a version of the formula that predicts the effect of doubling test length on reliability:

$$\rho_{22'} = 2\rho_{11'}/(1 + \rho_{11'}) \quad (6)$$

In the particular case we are considering, the predicted reliability for the 60-item test would be

$$\rho_{22'} = 2(0.72)/(1 + 0.72)$$

or, approximately, 0.84. She may then decide whether the increased reliability is sufficient to warrant the extra work in developing a longer test, and the extra testing time required to use it. Note also that, if the length were doubled again, to 120 items, the Spearman-Brown formula would predict that the reliability would increase further to 2(0.84)/(1 + 0.84) or, approximately, 0.91. There are two things to note about this prediction. First, beyond a certain point, the reward, in terms of increased reliability, for increases in test length begins to diminish; the last 60 items have increased the predicted reliability rather less than did the previous 30. Second, the prediction is becoming exceedingly risky at this point. For the Spearman-Brown formula to apply, it is necessary that the additional items function in the same way as those already present. They should be of the same type (multiple-choice, short-answer, etc.) and should test similar knowledge and skills. But also it is necessary that the students should approach them similarly; if the length of the test is such that fatigue, boredom, or resentment begin to affect the students' behavior, we could not expect the formula to give us sensible predictions.

Item type. Generally speaking, more reliable scores come from tests in which the items can be scored objectively than from tests in which the scoring involves an element of subjectivity. Objective tests are usually more reliable than essay tests of the same length (measured in terms of total testing time) for two reasons: first, they eliminate scorer inconsistency as a source of measurement error, and, secondly, they are able to cover more content, thus reducing the unreliability that can result from luck in the selection of questions. In general, then, for a given testing time, the greater reliability will be obtained by using a larger number of shorter, more objectively scorable items than from using a smaller number of longer, less objectively scorable tasks.

Item quality. Poor quality items will detract from the reliability of almost any test. When items are unclear or even ambiguous, error will be introduced by the varying interpretations that students may place on the items. When an item is much too difficult for the students being tested, they will either not answer it or guess. Guessing adds an element of randomness to scores: some gain a mark through chance; others of equal ability are not so rewarded. When an item is so easy that all students can answer it correctly, it does not detract from test reliability, but it does nothing to enhance it. The effect of adding such an item to a test is the same as that of adding one mark to each student's score—while the scores are all that much higher, the capacity of the test to make reliable distinctions among students is unaffected. Items that contribute most to test reliability are those that discriminate—in the technical sense, this refers to items on which students who possess the knowledge and skill needed to answer the question correctly have a better chance of success than students not in possession of this knowledge and skill. Items that are either very easy or very difficult for all the students being tested cannot be good discriminators. Therefore, it can be said that in order to maximize reliability a test should be pitched at a level of difficulty that matches the abilities of the students, neither too easy for them nor (the worse of the two) too difficult.

The Conditions of Administration

Here we refer to the physical conditions under which the test is administered (e.g., the light, noise and temperature levels of the testing room), the instructions used to set the task for the examinees, the time limits imposed (if any), and the person administering the test (e.g., the individual's vigilance in detecting copying and other forms of cheating, and ability to cope with the contingencies that inevitably arise during the administration of a test). To the extent that these factors vary unpredictably from one administration of the test to another, and to the extent that the conditions in effect during an administration affect some examinees differently from the way they affect other examinees, test scores will vary for reasons other than differences among the examinees in the knowledge and skill being tested. Effects of this kind work to reduce reliability by introducing unwanted sources of random variation or measurement error into the test scores. The above ideas are elaborated with reference to two aspects of the administration of a test.

Instructions. Consider the matter of guessing the answers to multiple-choice test items. It has been common practice in the past for test developers *not* to consider guessing in the instructions that examinees receive at the beginning of a test session. Consequently, examinees have differed in their test-taking behavior for reasons deriving from differences in willingness to guess, not differences in knowledge and skill in the subject being tested. Greater control of guessing behavior, albeit imperfect control, can be achieved by giving all examinees the same instruction for how to respond to items that cannot be answered correctly from knowledge. For example, instructing examinees to answer every question and to guess if necessary and, in addition, informing examinees that wrong answers will receive the same (zero) credit as omitted answers, should have the effect of reducing, possibly even eliminating, the effect on test scores of differences among examinees in willingness to guess.

Time limits. An important condition of the test administration is whether or not the time allowed is sufficiently long for all or almost all examinees to finish the test. If the test is speeded, then one of the abilities required to do well is the ability to work quickly. Tests that must be worked quickly may appear to be more reliable than tests that do not emphasize speed for the reason that speed of work may be an attribute on which examinees differ consistently. Reliability will be enhanced, however, at the cost of the test's validity, for the speeded instrument will be measuring something different from what it was intended to measure. More important, from the perspective of assessing test reliability, is the fact that no method of estimating reliability from the results of a single administration of a test will be satisfactory because, however the test is divided so as to provide the repeated measurements needed for estimating reliability, the scores for some parts will be more affected by the time limitation than the scores for other parts. (We expect performance of items near the end of a test to be most affected by the imposition of a strict time limit.) The only way to gauge the reliability of highly speeded tests is to employ at least two alternate forms, administering them in separately timed sessions.

The Group of Examinees

The size of the reliability coefficient depends on the range of true differences in ability in the group of examinees tested. A group in which the range of ability is narrow will yield a lower index of reliability than a group in which the range of ability is broad, even though the measuring instrument is the same for both groups. This property has interesting consequences, which have caused some to question the value of the reliability coefficient in certain applications of testing. Consider the situation, for example, in which the same test is used in successive years for the same course. The instructor becomes more effective each year, and the students achieve the course objectives more fully. Soon we will reach the situation in which many students will be scoring so highly there will be little further room for improvement. As more students reach this level, the scores of the group will become compressed into the top of the score range, and the score variance will become smaller. Consequently, the reliability coefficient will become successively smaller, and, in an apparent paradox, the more effective the instruction, the less reliable the measurement will appear to be. There is no paradox, however, provided we keep in mind just what property of a test is indexed by the reliability coefficient. We are looking at the capability of the test to make reliable distinctions among the group of examinees with respect to the ability measured by the test. If there is a great range of ability in a group, a good test should be able to do this very well. But if the examinees differ very little from one another, as they will if the test covers a limited range of tasks in which all examinees are highly skilled, reliable distinctions will be difficult to make, even with a test of high quality. A low reliability coefficient does not necessarily mean that the test is of poor quality, and in some circumstances a poor test might measure with high reliability. (These considerations come into play in assessing the quality of so-called criterion-referenced tests.) We should not, therefore, think of reliability as telling us all that we need to know about test quality. A reliability coefficient tells us about *a* quality of a test (and one that we usually value), but not about *the* quality of the test.

Self-Test

Indicate whether each of the following 10 statements is true or false. If a statement is false, revise it to be true.

1. A reliability coefficient describes the consistency with which a test measures some characteristic of one person.
2. If the students in a group are tested twice, using parallel forms of a test, if the pair of scores for each student defines a point in a scatter plot of Form I scores against Form II scores, and if all the plotted points lie on the 45 degree line through the origin (0,0 coordinate) of the graph, then the reliability of the test, as estimated by the coefficient of correlation between scores on the two test forms, is 1.
3. Through an error of computer programming, all the university applicants who took an admission test were credited with 10 more correct answers than they really earned on the test. This mistake added error of measurement to the test scores.
4. Suppose a class of students, none of whom has studied the branch of mathematics known as calculus, is given a multiple-choice test of the common derivatives of differential calculus. Each student guesses the answer to every question. The variance of the students' scores on this test will be composed of some true-score variance and some error-score variance.
5. A sample of 100 students from a well-defined population is administered two parallel forms of a test, the administrations being separated by a week. If the coefficient of correlation between the scores on the two test forms is 0.9, then these test scores provide an estimate of reliability for the test equal to 0.9^2 or 0.81.
6. Two test publishers, A and B, each develop two parallel forms of a test of punctuation skills. The reliability of Publisher A's test is estimated by administering both test forms to a sample of fifth grade students. Publisher B obtains an estimate of reliability for its test by administering both forms to a sample of students drawn from the fifth, sixth and seventh grades. If the tests for both publishers are equal in length and if the administrations of the parallel forms of each publisher are separated by one week, the estimates of reliability for both tests will most likely be about the same.
7. The scores on a multiple-choice test will be more reliable than the scores on a free-response test of the same knowledge, provided both tests are of the same length and the two groups of examinees involved in the reliability-estimation experiments, one group for the multiple-choice test and one for the free-response test, are randomly equivalent in ability and knowledge.
8. A coefficient of correlation between the scores for a group of examinees on parallel forms of a speeded test yields an acceptable estimate of reliability for the test.
9. If a test is doubled in length, the reliability of scores on the lengthened test will very likely be twice the reliability of scores on the test at its original (undoubled) length.
10. If the correlation between scores on parallel forms of a test is used to estimate reliability, then the range of possible values for the reliability coefficient must be -1 to $+1$.

Self-Test Key and Explanations

1. False. Reliability coefficients describe the consistency with which test scores are assigned the members of a population of persons. A reliability coefficient involves the notion of true-score variance. If we have several test scores for one person, and these scores measure the same characteristic of the person in the same way, then any inconsistency in the scores is an indicator of error of measurement. The true score of the person, within reasonable limits, is assumed not to differ from one of these measurements to another. (See section of *Understanding Reliability* entitled "Formalization.")
2. True. The coefficient of correlation between the scores of a sample of examinees on parallel forms of a test provides an estimate of the reliability of the scores examinees earn on either test form. The fact that the plotted scores lie on a straight line with a positive slope (i.e., scores on the test defining the ordinate or vertical axis of the plot increase as scores on the test defining the abscissa or horizontal axis increase) means that the correlation coefficient will be $+1$. (See section of *Understanding Reliability* entitled "Graphical Treatment.")
3. False. Measurement error is random from person to person, not systematic and constant for all persons, as in this question. The computer programming error results in each person having an apparent true score that is 10 points larger than it should be. (See section of *Understanding Reliability* entitled "Graphical Treatment.")
4. False. Students who are totally ignorant of calculus, as these students are alleged to be, will have to answer the multiple-choice questions by guessing. All differences among their test scores will then be due only to chance, with the students who receive higher scores being luckier (not more knowledgeable) than those who receive lower scores. In this case all variance in test scores must be due to error of measurement. (See section of *Understanding Reliability* entitled "Formalization.")
5. False. The reliability of a test, defined as the ratio of true-score variance to the variance of the observed test scores, is equal to the coefficient of correlation between scores on parallel forms of the test, not the square of

the coefficient of correlation. (See section of *Understanding Reliability* entitled "Estimating Reliability.")

6. False. In addition to such factors of the test as length and to such conditions of the test administrations as the length of time between them (here, one week for the test of each publisher), estimates of reliability depend on the range of ability in the group tested. It is very likely that the range of differences in punctuation skills is wider in the group of students tested by Publisher B than it is in the group tested by Publisher A. All other things being equal, then, we expect the estimate of reliability for the test of Publisher B to be larger than the estimate of reliability for the test of Publisher A. (See section of *Understanding Reliability* entitled "What Makes a Test Reliable.")
7. Uncertain. The reliability of a multiple-choice test is attenuated or reduced by the guessing that can occur when examinees who don't know the answer attempt the question anyway. This source of unreliability either doesn't exist for the free-response test or is greatly reduced by the fact that the examinee who doesn't know the correct answer cannot simply choose one of a small set of multiple-choices for his or her answer. The examinee who guesses the answer to an item on a free-response test must produce a response, which in the face of total ignorance is unlikely to be correct. On the other hand, free-response answers must be scored by judges, and judges rarely achieve unanimous agreement on the marks to be assigned a free-response answer, especially one of any length. This source of unreliability, disagreements among judges as to the worth of answers, does not affect the scoring of multiple-choice tests. Which of the multiple-choice and the free-response tests will be the more reliable depends on which source of unreliability, guessing or scorer disagreements, affects test scores the most. An empirical study is required to answer this question. (See section of *Understanding Reliability* entitled "What Makes a Test Reliable.")
8. True. Parallel forms of a speeded test, if separately and independently administered to a sample of examinees, will provide independent estimates of each examinee's ability to perform the test. These scores may be correlated to produce an estimate of the reliability of the test. Two scores derived from examinee performance of only one form of a speeded test, e.g., the performance of odd-numbered items versus the performance of even-numbered items, are not independent when the test is speeded and hence do not provide a satisfactory basis for estimating reliability. (See section of *Understanding Reliability* entitled "Estimating Reliability.")
9. False. The relation between length and reliability is not one of simple proportionality. The Spearman-Brown formula provides an estimate of the reliability of a lengthened test. If a test of reliability 0.6 is doubled in length, the reliability of the lengthened test is estimated to be .75 [$=(2 \times 0.6)/(1 + 0.6)$]. (See section of *Understanding Reliability* entitled "What Makes a Test Reliable.")
10. False, at least in theory. The reliability coefficient is, by definition, the ratio of two variances, and a variance is always greater than or equal to zero. Assuming the denominator of the ratio, the observed-score variance, is greater than zero, it follows that the reliability coefficient must in theory always be greater than or equal to zero. Practice can, of course, differ from theory. In practice, the estimate of a reliability coefficient might be negative, as it would be if two supposedly parallel forms of a test produced scores that gave rise to a negative coefficient of correlation. But if a negative parallel-forms estimate of reliability were obtained, we would be led to question whether the forms really were parallel measures of the same characteristic. Alternatively, we would question the procedure followed in administering the two tests, or some other feature of the experiment that was conducted to obtain the scores that were correlated. (See sections of *Understanding Reliability* entitled "Formalization" and "Estimating Reliability.")

Annotated References

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory.* Monterey, California: Brooks/Cole.

An introductory treatment of measurement concepts, including a review of basic statistics and a development of classical reliability theory. Validity and test construction are also considered, together with such other matters as the scaling of test scores and the equating of scores on one test to those on another test.

Brennan, R. L. (1983). *Elements of generalizability theory.* Iowa City, Iowa: ACT Publications.

A thorough and relatively advanced treatment of the subject. An understanding of analysis of variance and the estimation of components of variance is required.

Brennan, R. L., & Kane, M. T. (1979). Generalizability theory: A Review. In R. E. Traub (Ed.), *New Directions for Testing and Measurement,* No. 4. (pp. 33–51). San Francisco: Jossey-Bass.

This chapter provides a relatively brief introduction to generalizability theory. Less demanding mathematically than Brennan's book, it is nevertheless necessary to possess an understanding of analysis of variance.

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

Textbook coverage of test theory, including classical reliability theory and generalizability theory. The required statistical concepts are reviewed in Chapter 2. Among other topics considered are test construction, item analysis, and validity.

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed.) (pp. 105–146). New York: Macmillan.

An up-to-date review of classical reliability theory and generalizability theory. A knowledge of statistics, including analysis of variance, is required.

Frisbie, D. A. (1988). Reliability of scores from teacher-made tests. *Educational Measurement: Issues and Practice, 7,* 25–35.

Another instructional module in the ITEMS series on the topic of reliability, one that overlaps substantially with the present module, and otherwise complements it. Easy to read, with low demand for background in mathematics and statistics.

New Editor Appointed
for
Educational Measurement: Issues and Practice

Beginning with issue 1, 1992, Dr. John Fremer will serve as editor of *Educational Measurement: Issues and Practice (EM:IP).* Dr. Fremer, who will serve a 3-year editorship, is now receiving manuscripts for possible publication in *EM:IP.* Please send manuscripts to the following address:

Dr. John Fremer
Mail Stop 07-E
Educational Testing Service (ETS)
Princeton, NJ 08541