Photo credit Comstock, Inc.

Kenneth Clark

Test Reliability

Whether a classroom test is "good" is, to some degree, a matter of opinion. For me, as a teacher, the bottom line was the response of my students. I would ask a simple question: "Was that a good test?" To students, *good* means that the breadth and depth of course work are well and fairly reflected. Students' responses were, for me, one important indicator of the quality of my classroom test.

However, students take tests other than classroom tests. Given the stakes associated with standardized tests, evidence must show that such test results mean what they are intended to mean; they must be *valid,* and individual student scores must be reasonably precise and free from measurement error, that is, they must be *reliable*.

Unfortunately, meaningful informal feedback from students is difficult to obtain in large-scale testing. The determination of test validity and reliability is yielded somewhat more objectively through applying recognized—and debated—theories and statistical analytical methods. Test developers apply them to obtain evidence that the tests they produce are "good."

This article explains and demonstrates a procedure that is commonly used to determine the reliability of a test in such a way that a person who has modest arithmetical skills can carry out the same analysis on a classroom test or examination. The article also presents issues that arise from this approach to assessing test quality.

THE RELIABILITY OF A TEST

The reliability of a test is the same in nature as the reliability of a car: can it be counted on to give consistent results? Another issue, that of test validity, which is not discussed in this article, answers a question along the lines of "For what are these test results useful?" For example, a car that never starts is perfectly reliable—you can rely on the car never to start—but it is useless, or invalid, if you desire to use it for traveling.

Reliability is expressed numerically through a number of procedures, including the following. Imagine a group of four students—named A, B, C, and D—who take test 1, composed of items, that is, questions, 1, 2, and 3, each of which is worth three points. **Table 1** shows the results. To analyze a multiple-choice test, the body of the table would contain only 0's and 1's. The same procedures would apply.

TABLE 1
Results for Test 1
Student Score on Each Item

	Item			Student Score
Student	#1	#2	#3	(Average per Item)
A	1	2	3	2.00
B	0	1	2	1.00
C	1	2	3	2.00
D	0	1	2	1.00
Item average	.5	1.5	2.5	1.50 (grand average)

Results such as those in **table 1** are considered ideal for the following reasons:

- The test has distinguished among students' performance on the test items. This result is evident because of the different student average-per-item scores and is appropriate if a range of student achievement is expected.
- A student who outperforms another student on one item outperforms the other student on all items. This kind of consistency at the *item* level

A car that never starts is perfectly reliable

Kenneth Clark, kclark@edu.gov.mb.ca, is a test developer, assessment consultant, and university instructor in Winnipeg, Manitoba. He previously was a high school science and mathematics teacher and, before that, a chemical engineer.

is evidence of consistency, or reliability, at the *test* level because each item is giving the same information about relative student achievement. This consistency leads to a degree of certainty or confidence with respect to interpreting the final scores. The teacher or interpreter can feel confident in saying that student A knows more about the content area than student B because all items reveal such evidence.

One measure used to arrive at a numerical indicator of the reliability of the test is called *Hoyt's reliability*.

The term* degrees of freedom *represents the number of free* choices *within a* certain *constraint

CALCULATING HOYT'S RELIABILITY

The following method of calculating reliability is designed to detect the degree to which the previously described kind of pattern of responses exists in the data.

The principle behind the calculation is that variability exists in individual item scores, that is, twelve scores ranging from 0 through 3; in student-score averages, that is, 1.0 or 2.0; and in the average score for each item, that is, 0.5, 1.5, or 2.5. We expect students to achieve to different degrees and expect test items to have a range of difficulty, that is, differing average scores. If student achievement and item variability can account for all the variability of the individual item scores, given in the body of the table, then the test is reliable. In such a test, the ranking of students is consistent across all items, although this result is by no means intuitive.

The statistical procedure used when analyzing variability in a set of data is called an ***an***_alysis_ ***o***_f_ ***va***_riance_ (ANOVA). The purpose of this procedure is to learn the degree to which the variation of the individual item scores, also called the *total variance,* consists of differences among student average-per-item scores and differences among item averages. A third source of variation is an interaction between students and test items, something considered to be "error," which relates to the previously mentioned inconsistency in ranking. A note to statisticians: the tests of significance associated with ANOVA are not employed here. Only the partitioning of total variance is of interest.

The calculation works as follows. Variability in data can be expressed in terms of various mean squares, each represented by a symbol of the form $MS_?$. The subscript identifies the data used in the calculation, as described subsequently. Each of these mean squares is found from a related sum of squares, represented by the symbol $SS_?$. To calculate any mean square, first find the corresponding *SS* by adding the squared deviations from the grand mean of each data value of the given type as reflected in the subscript, and then divide this sum of squares by the respective degrees of freedom, $df_?$.

Degrees of freedom is a complex idea that is not fully explained here. Essentially, it represents the number of free choices within a certain constraint. For example, the constraint may be that four numbers, such as the four student scores in the last column of **table 1,** have a certain mean. Given this mean, any values could theoretically be chosen for three of the numbers, although never done in practice; but the fourth number would then be constrained by the fixed value of the mean. In this context, three degrees of freedom occur.

When the *MS*s are known, Hoyt's reliability is calculated as follows:

$$\text{Reliability} = \frac{MS_{st} - MS_e}{MS_{st}}$$

The symbols MS_{st} and MS_e will subsequently be defined precisely, but intuitively MS_{st} is a measure of the amount of variation in student average-per-item scores. MS_e is the amount of variation associated with error, a measure of the amount of interaction between students and items, with ranking of students differing from item to item. A reliability of 1 is perfect and is approached when MS_e is much smaller than MS_{st}.

The seven-step calculation

1. Calculate SS_t, the **t**otal **s**um of **s**quares based on each individual item score in the body of the table, as follows, remembering that 1.5 is the grand mean:

$$SS_t = (1 - 1.5)^2 + (2 - 1.5)^2 + (3 - 1.5)^2 + (0 - 1.5)^2 + \cdots + (1 - 1.5)^2 + (2 - 1.5)^2 = 11,$$

which is the total variation in the data.

2. Calculate SS_{st}, the sum of squares attributable to **st**udent achievement, by in essence replacing each of the individual item scores in each row with the corresponding student average-per-item score and calculating as follows:

$$SS_{st} = (2 - 1.5)^2 \times 3 \text{ scores/row} + (1 - 1.5)^2 \times 3 + (2 - 1.5)^2 \times 3 + (1 - 1.5)^2 \times 3 = 3,$$

which is the variation associated with student achievement.

3. Calculate the sum of squares attributable to differences between **i**tems, SS_i, by in essence replacing each score in a column by the corresponding item average given in the last row of the table and proceeding as follows:

$$SS_i = (0.5 - 1.5)^2 \times 4 \text{ scores/column} + (1.5 - 1.5)^2 \times 4 + (2.5 - 1.5)^2 \times 4 = 8,$$

which is the variation associated with item difficulty

4. Calculate SS_e, the amount of variation associated with **e**rror. As mentioned, the total variation is made up of three components, namely, student variation, item-score variation, and error variation; that is, $SS_t = SS_{st} + SS_i + SS_e$. This formula is used to calculate SS_e. In this example, although it never happens so nicely in practice, $SS_e = 11 - 3 - 8 = 0$. All variability in the data is explained by student variation and item variation. This result represents perfect internal consistency in how students responded, all variability being attributable to students' own levels of achievement and the variability in item difficulty, to which all students were equally exposed, as reflected in the item averages.

The respective *MS*s are found using the formula

$$MS_? = \frac{SS_?}{df_?},$$

as follows:

5. $$MS_{st} = \frac{SS_{st}}{df_{st}},$$

where $df_{st} = 3$, because the number of students is four, so the result is

$$MS_{st} = \frac{3}{3}$$
$$= 1.$$

6. $$MS_e = \frac{SS_e}{df_e},$$

where $df_e = 6$. Note that df_e is the product of df_{st} and df_i, equal to the number of free choices in the data, given the fixed column and row means. The result is

$$MS_e = \frac{0}{6}$$
$$= 0.$$

7. $$\text{Reliability} = \frac{MS_{st} - MS_e}{MS_{st}}$$
$$= \frac{1-0}{1}$$
$$= 1$$

The interpretation is that the test is internally reliable, which means that each item in the test delivers the same message about relative student achievement for the test, thus leading to some degree of confidence in the test results, that is, how students are ranked. This reliability, or a similar one, is commonly reported with published standardized tests and with many other psychological measures. Results close to 1 are ideal.

Test 2's results demonstrate how and why a reliability of less than 1 can occur. See **table 2**. The 0 and 1 entries have been switched for students B and D. The students finish with the same final scores; neither SS_{st} nor MS_{st} changes; and the total variation, SS_t, remains the same.

TABLE 2

Results for Test 2
Student Score on Each Item

	Item			Student Score
Student	#1	#2	#3	(Average per Item)
A	1	2	3	2.00
B	0	1	2	1.00
C	1	2	3	2.00
D	0	1	2	1.00
Item average	.5	1.5	2.5	1.50 (grand average)

The splendid consistency seen in the test 1 results no longer occurs. Item 1 results indicate that all students have equal ability, whereas items 2 and 3 consistently indicate that students A and C are stronger than students B and D. Test 2 is, therefore, less internally reliable. The degree of uncertainty that occurs is reflected in the appearance of a nonzero error.

To find the reliability, follow precisely the same procedure as before. This time,

$$SS_t = 11,$$

which shows no change from the previous result;

$$SS_{st} = 3; MS_{st} = 1,$$

which shows no change from the previous result;

$$SS_i = (1 - 1.5)^2 \times 4 + (1 - 1.5)^2 \times 4 + (2.5 - 1.5)^2 \times 4$$
$$= 6,$$

whereas the previous result was 8;

$$SS_e = 11 - 3 - 6 = 2,$$

whereas the previous result was 0; and

$$MS_e = \frac{2}{3 \times 2};$$

(recall that $df_e = df_{st} \times df_i = 3 \times 2 = 6$

$$= 0.33.$$

Less variation results from the items because the item averages—the last row of the table—are less dispersed. Meanwhile, the total variation, 11, and the student variation, 3, have remained the same. The missing variation, 11 – 3 – 6 = 2, is error, reflecting the inconsistency of student achievement across the items. This apparent interaction between items and students reflects the fact that how well a student achieves relative to the others in the group changes from item to item. This interaction between student achievement and item is a mark of internal inconsistency.

☛

Internally reliable means that each test item delivers the same message about relative student achievement

Hoyt's reliability for test 2 is

$$\begin{aligned}\text{Reliability} &= \frac{MS_{st} - MS_e}{MS_{st}}\\ &= \frac{1 - .33}{1}\\ &= 0.67.\end{aligned}$$

A reliability of 0.67 is too low for such high-stakes decision making as college entrance, for example. Widely administered standardized tests typically have reliabilities of approximately 0.9 and include a larger number of items that have been administered to hundreds or thousands of subjects.

By following the given algorithm, you can calculate the reliability of tests that you have created. However, do not put much stock in the result. High reliability can mean that you have devised a very good test, that instruction has been ineffective, or that your students have a wide range of abilities or interests. Low reliability can mean that you devised a test that has little to do with student knowledge, that instruction has been highly effective, or that your class is homogeneous in student ability.

INTERPRETING HOYT'S RELIABILITY

High reliability means that a teacher can be confident that Mary is better than Alan with respect to the attributes being measured by the test. If ranking students is the expressed goal of a test, then the measure of reliability explained here is one appropriate measure of test quality.

However, the measure has limitations. For example, notice that a requirement for the calculation to work is variability in student achievement, MS_{st}; the formula is reliability = $(MS_{st} - MS_e)/MS_{st}$. If student variation is 0, that is, if all students have the same student score or average per item, then reliability is undefined, since $MS_{st} = 0$. If students are very heterogeneous, the reliability approaches 1, that is, MS_{st} becomes much larger than MS_e because of the wide range of student scores. Therefore, the numerical reliability of a test is a function of the students taking the test, as well as of the test itself. Imagine giving a twelfth-grade English class a list of words like *cat, sat,* and *hat* to spell. All students would score nearly perfectly—except for random, careless errors, and these results would be obtained repeatedly if such tests were administered over and over. Despite what might appear to be a high reliability, the reliability may not be close to 1 and may even be negative if MS_e exceeds MS_{st}. Likewise, if you administer a test to a class in which every student has mastered the content area, the test may have a low reliability, although it could be a fine, content-relevant test.

Determining test quality is not appropriate where the goal is homogeneous student achievement

Meanwhile, giving a test on auto mechanics to a tenth-grade mathematics class may prove to be very reliable. A few in the class will know a good deal about cars, some will know a bit, and some will know nothing. A wide range of student scores will result, that is, MS_{st} will be high; and a student who knows nothing about cars will rarely get an item correct that a knowledgeable student cannot answer, so MS_e will be low. The test will be very reliable but of little value to the teacher. An on-topic test in a classroom where students have a wide range of achievement will lead to similar results. Students who are very strong will perform consistently better on the items than the others, a wide range of scores will occur, and test reliability will be high.

USING THE RELIABILITY COEFFICIENT AS A TEST-QUALITY INDICATOR

Using a reliability measurement such as Hoyt's reliability has significant test-design implications, as well as ethical considerations. For example, the reliability of a test is improved by restricting it to items for which student performance on the item is strongly related to student total score. In other words, test developers tend to prefer items for which a positive correlation occurs between how well students do on the item and how well they do on the total test. This correlation is a major criterion for including an item in a standardized test designed for ranking students. The higher the correlation between item score and total test score for the items on a test, the greater the reliability of the test. Test items for which test-takers' scores correlate poorly with their total test scores may be replaced. On such low-correlation items, students who are typically unsuccessful tend to do better and students who are typically successful experience difficulty. This result is often anathema to standardized-test designers and is perhaps unfair to students. Underlying this approach is a pervasive and too-often tacit assumption that general, cross-disciplinary degrees of achievement, skill, or intelligence are out there to be captured. Test items are considered worthy only if they confirm that belief. This belief is fine unless it is wrong.

Test developers and those who interpret test scores must concern themselves with other issues. Note that test reliability, as defined here, has nothing to do with the degree to which test items reflect a content area or the difficulty of the questions. Total test score, therefore, has little meaning other than allowing a description of how one student ranks relative to another on that particular test. Such test-score results are therefore generally norm-referenced and translated into percentiles, grade equivalents, or some other statistic.

Another important point is that the approach to determining test quality described in this article is not appropriate in an environment where homogeneity exists, or where it is a goal, in the level of student achievement. Recall that test items that consistently separate low achievers from high achievers are preferred. Items that fail to distinguish among students in this way are replaced by items that do. Therefore, one cannot detect improvements in the uniformity of student achievement, much as one would be unable to ever see most students become “above average.”

The move toward standards and standards testing is one response to these concerns. Standards of achievement are set, and student performance is assessed relative to those standards rather than to the performance of other students. The question of reliability is then expanded to respond to such questions as “With what consistency does this test categorize students with respect to mastery in a certain domain?” or “What probable range of scores would this student get on other tests like this one from the same content domain?” Such a view of reliability goes beyond comparing one student with another. Its use will become more frequent with the advent of standards and tests designed to measure student achievement relative to these standards.

BIBLIOGRAPHY

Crocker, Linda, and James Algina. *Introduction to Classical and Modern Test Theory.* New York: Holt, Rinehart & Winston, 1986.

Nitko, Anthony J. *Educational Tests and Measurement: An Introduction.* New York: Harcourt Brace Jovanovich, 1983.

Traub, Ross E. *Reliability for the Social Sciences: Theory and Applications.* Measurement Methods in the Social Sciences, vol. 3. Thousand Oaks, Calif.: Sage Publications, 1994.

My thanks to the Editorial Panel for helpful and thoughtful suggestions, to my colleague Harri Maharaj for his comments on an early draft, and to my colleagues Xiao Pang and Ping Yang for helpful discussions.