Reading Horizons

Volume 7, *Issue* 3 1967 *Article* 4

APRIL 1967

A Caution Concerning the Use of Standardized Tests

Ronald Crowell*

*

A Caution Concerning the Use of Standardized Tests

Ronald Crowell

Abstract

As teachers in the schools of this country, we pay lip service to the proposal that we must be concerned with the education of the individual. In our textbooks in college, we are reminded that the aim of good teaching is to bring about a change of behavior in the individual. In our classrooms, because of broad differences in individual reading ability and individual performance, we often find it necessary to group students into small, more homogeneous, groups so that we can meet the reading needs of the individual student.

A CAUTION CONCERNING THE USE OF STANDARDIZED TESTS

Ronald Crowell

As teachers in the schools of this country, we pay lip service to the proposal that we must be concerned with the education of the *individual.* In our textbooks in college, we are reminded that the aim of good teaching is to bring about a change of behavior in the individual. In our classrooms, because of broad differences in individual reading ability and individual performance, we often find it necessary to group students into small, more homogeneous, groups so that we can meet the reading needs of the individual student.

In attempting to determine the ability level or readiness level of students in order to place them in the appropriate group, we often use scores from reading tests, achievement tests, and intelligence tests. However, in most practical measurement situations the only information available to us from a test is the score of the person measured. That sounds as though it should be a useful score since we have obtained scores from all of the individuals taking that particular test. But, is it? As every person who uses tests knows, no test is perfectly accurate.

Theoretically, the obtained or observed score of any individual on any given test equals his TRUE score on that test plus a certain degree of random error, over which we have no control.(1) This random error is due to the inaccuracy of the test itself.(3) We may also recall that the accuracy of our judgments depends upon the consistency or reliability of that particular measuring instrument. We read over and over that the higher the reliability of a standardized test, the better the test is as an evaluation instrument. This certainly is true for the group as a whole. The reliability of a test is one of the most important criteria by which we choose standardized tests for use in our school systems. However, the reliability of a test does not help us directly in evaluating the scores of an individual student.

I would like to discuss two points about test reliability which we often do not consider when making decisions regarding individual students in the practical test situation. The first of these concepts is the standard error of measurement, (SEm). The SEm is one way in which we can interpret the reliability of the test. It is a statistic which has been developed for estimating the margin of error we should allow for in test scores of individual students.

When a test is given to large numbers of students the scores of the test often fall into a normal or bell shaped curve. Now, suppose we take the score of any individual who has taken this standardized test and then administer a very large number of comparable forms of this test to this student. We would find that the student does not always get the same score. In fact, as the number of forms administered gets larger we would discover that the distribution of this student's scores also begins to resemble the normal curve, although the normal curve in this case would be much smaller than the normal curve for the total group taking the test. Suppose we gave this test to this individual 100 times and kept averaging the results so that no further repetition changes his average score. The average of this series of scores we can reasonably expect to be characteristic of the student's performance so we may call this an estimate of his TRUE score. We can see that the observed score of the individual on any given administration equals his true score plus a certain amount of error that makes his observed score deviate from his true score on that particular administration of the test. If you were to mark off two points on this normal curve that would enclose the middle two-thirds of all of the scores of the previous trials, you would call these points one standard error above the true measure and one standard error below the true measure. If you were to mark off the points that enclose 95% of his scores around the estimate of the true score, you would say that this was two standard errors below his true score and two standard errors above his true score. The problem is, of course, that we really do not have 100 repetitions of any given test and so we really do not know the TRUE score of any individual. Likewise, the standard error of measurement associated with each score is unknown. However, statistical theory permits us to compute an estimate of the standard error of measurement based on an individual's single obtained score and the reliability of that test.

The point of this discussion is that we as teachers recognize that if we give any particular test to our class the scores are going to vary from very low scores for certain members of the class to very high scores for other members of the class. However, we often ignore the fact that for a variety of reasons, a test score for any specific individual is going to vary from one administration of the test to the next or from one day to the next. For example, if we give a standardized reading test to our class and Johnny Jones gets a raw score of 70 on this test, we cannot make the statement that this score for Johnny is an especially accurate indication of his reading ability

measured by this one test. The most accurate statement we can make is to use 2 SEm in defining limits around the observed score within which we would be "reasonably sure" to find Johnny's TRUE score 95 times out of 100.(3) *This number is not small.* For instance, suppose the standard deviation of a reading test taken by the group of students is 15 and suppose the SEm of any given individual taking the test is 5, that means that any person within the group is likely to shift over a large range of scores and we can be *reasonably sure* only that his TRUE score lies somewhere within plus or minus 10 points (+ or —2 SEm) of his observed score. Too often teachers make the judgment that because student A has a score of 104 on a standardized intelligence test and student B has a score of 96 on the test that student A is, in fact, "really" brighter than student B. On the basis of this erroneous judgment, student A is often placed in a more advanced group. This kind of judgment simply cannot be made.

Many of the recent revisions of standardized intelligence tests and other achievement tests such as the SCAT and the STEP tests(4) are calling attention to this concept of the SEm by reporting test results in terms of bands along a given scale instead of points on a scale. Manuals for these tests usually explain that the chances are 2 out of 3 that the TRUE score of an individual will lie somewhere in a given band and urge teachers not to regard scores as any more precise than that.

The smaller the number of items in any test, however, the smaller is the standard error that we may expect. For example, the estimated standard error of test scores for a test having anywhere from 48 to 89 test items is approximately four for a person who scores in the midrange of that test and slightly less than four for a person scoring at the extreme ranges of that test.(3)

Another concept which is equally as important as the SEm is the standard error of a difference (SEdiff). If we want to consider whether two scores on a given test are, in fact, different, we must consider the concept of the SEdiff which is larger than the SEm of an individual for one test score. For example, "think of the difference as a rope tied between two stakes, which are the two scores. Since there is a wobble in both stakes, there is bound to be more wobble in the rope than there is in either stake."(4) If we find that on any given test the SEm is three we will find that the SEdiff between any two given scores is approximately 4¼. Therefore, if we are trying to make a judgment about the significance of a gain any individual has made over a period of time and if we want to be reasonably sure that his

two scores represent a *true* difference in ability, the difference between them should be twice the SEdiff or at least 8½ points apart. Although this is an important concept, in practice it is close enough to consider two scores "really" different if the two scores are simply two SEm apart—or, as in this example, six points apart.

The second question to consider is what percent of the individuals would remain in the groups we have placed them in if their *true* scores were known? That is, what percentage of the students are we placing in the wrong group? For example, suppose we are going to divide the students in our classroom into just two groups of equal size on the basis of their scores on any one particular standardized test. Those in the half making higher scores will be placed in one group and those in the half making lower scores would be placed in another group. The answer to this question of incorrect placement also depends on the reliability of that standardized test.

If the reliability coefficient reported for that test was .96, a very high reliability, then 95% of the individuals in any one group would remain in that group on the basis of their observed or fallible scores, if the true scores on the test were actually known.(2) If the reliability coefficient is .85, a relatively common reliability coefficient, we will find that only 87% of our students would stay in the group to which we had originally assigned them and 13% of the students would move from the lower half to the upper half or from the upper half to the lower half. Although this does not seem like a very large percentage, it may be extremely important to any one person whom we have assigned to a group incorrectly. The teaching methods that we utilize might be totally inappropriate for that student of lesser ability who is assigned to the upper group and, likewise, might have a stifling or suffocating effect on the student of high ability who, by error, we have assigned to the lower group in our class. Now, this type of error is compounded in the class that we divide into three groups—which we typically do in any reading situation. If we divided a class into three groups on the basis of results from a test whose reliability was .90 we would make only 73% correct assignments. That is, we might expect more than 25% of our students to change from one group to another on any given administration of the standardized test.(2)

In this brief discussion of these two concepts, the Standard Error of Measurement and the percent of incorrect assignments, it should be apparent that if we are making judgments regarding the capabilities of individual students based on the scores from any one

standardized test, we are in danger of doing the student a great disservice by placing him in an incorrect group, making inaccurate judgments regarding his ability or by prejudicing our own view of that student's capability. This, however, is not to say that tests are not useful to us in the classroom. It does say that if we make judgments on the basis of scores of a single test the possibility exists (with a rather high probability) that we will be making incorrect decisions.

As teachers we must base our judgment not on one test score but on the basis of all data available to us. Such sources of information as informal inventories, academic histories and observations should be utilized. An awareness of the problems which can arise should make it clear that, until there are more adequate measures of the educational objectives we hope to achieve, we must be extremely careful in basing decisions regarding individual students purely on the basis of the results of one administration of one test. We must use all the tools at our disposal to assure us the most adequate picture of the student's ability.

References

1. Cronbach, Lee J., *Essentials of Psychological Testing (Second Edition)*. New York: Harper & Brothers, Publishers, 1960.
2. Ebel, Robert L., *Measuring Educational Achievement*. Englewood Cliffs, New Jersey: Prentice-Hall, Inc., 1965.
3. How Accurate is a Test Score? *Test Service Bulletin, Number 50*. New York: The Psychological Corporation, June, 1956.
4. Short-cut Statistics For Teacher-made Tests, *Evaluation and Advisory Service Series, Number 5*. Princeton, New Jersey: Educational Testing Service, 1960.