PEARSON NEW INTERNATIONAL EDITION

Measurement and Evaluation in Psychology and Education

Thorndike Thorndike-Christ

Eighth Edition

ALWAYS LEARNING™

PEARSON®

Pearson New International Edition

Measurement and Evaluation
in Psychology and Education
Thorndike Thorndike-Christ
Eighth Edition

PEARSON®

approximately by grade-point average. We give an employment test to select machine operators who are likely to be successful employees, as represented by some criterion such as high production with low spoilage and low personnel turnover. For this purpose, we care very little what a test looks like. We are interested almost entirely in the degree to which it correlates with some chosen criterion measure of job or academic success. Some other measure (often, but not necessarily, one that becomes available later) is taken as the criterion of "success," and we judge a test in terms of its relationship to that criterion measure. The higher the correlation, the better the test.

Face Validity

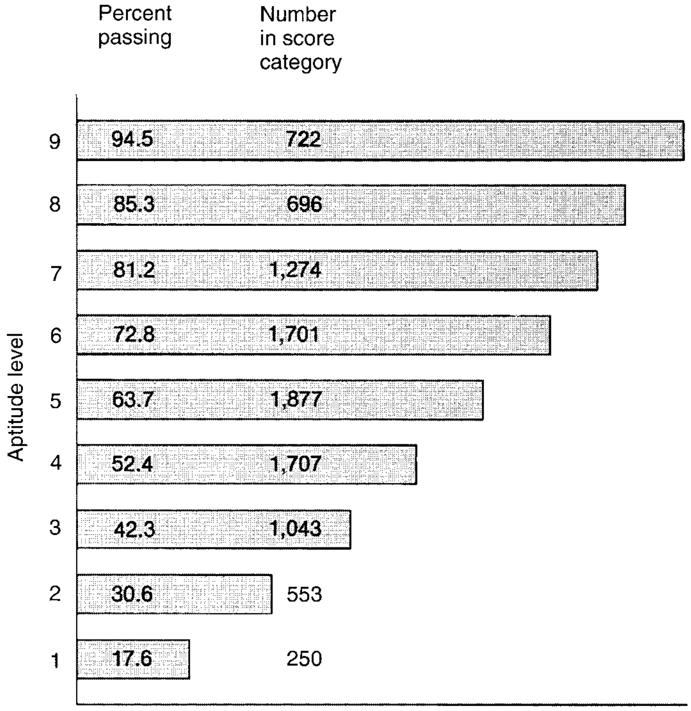
The statement that we do not care what a test looks like is not entirely true. What a test "looks like" may be of importance in determining its acceptability and reasonableness to those who will be tested. A group of would-be pilots may be more ready to accept a mathematics test dealing with wind drift and fuel consumption than they would be to accept a test with the same essential problems phrased in terms of costs of crops or recipes for baking cakes. This appearance of reasonableness is often called **face validity,** and although it is never a sufficient condition for the use of a test, it can be considered a necessary condition whenever the voluntary cooperation of the examinees is important. Unless the test looks like it could be used for the purpose the examinees have been told it will be used for, they may give less than their maximum effort or may not provide sincere responses. As Sackett, Schmitt, Ellingson, and Kabin (2001) note, "When a test looks appropriate for the performance situation in which examinees will be expected to perform, they tend to react positively. . . . Equally important, perhaps, may be the perception that one is fairly treated" (pp. 315–316).

Empirical Validity

Evaluation of a test *as a predictor* is primarily an empirical and statistical evaluation, and the collection of evidence bearing on this aspect of validity has sometimes been spoken of as **empirical** or **statistical validity.** The basic procedure is to give the test to a group that is entering some job or training program; to follow them up later; to get for each one a specified measure of success on the job or in the training program, known as the **criterion;** and then to compute the correlation between the test scores and the criterion measures of success. The higher the correlation, the more effective the test is as a predictor and the higher is its **criterion-related validity.**

There are really two broad classes of criterion-related validity, with the differentiation depending on the time relationship between collecting the test information and the criterion information. If the test information is to be used to forecast future criterion performance, as is the case in all selection decisions, then we speak of **predictive validity.** On the other hand, in some cases we would like to be able to substitute a test for another more complex or expensive procedure. This might be the case when we use a group test of intelligence in place of an individually administered one. The question then is whether the two different sources of information have a strong enough relationship that the less expensive one can be used in place of the more expensive one. If this is our concern, we would want to know that scores on one test correlate highly with scores obtained concurrently on the other. When scores on the test and the criterion are obtained at essentially the same time, we speak of **concurrent validity.** The two labels, *predictive* and *concurrent,* have more to do with the purpose of the study than with the time relationship between the assessments, but these are the terms that have evolved. Because predictive validity is the more widespread concern, we will use that term in the discussion to follow, but the procedures and issues are the same for both types of empirical validity.

Figure 5–1
Percent of cadets completing pilot training at each aptitude level. The correlation coefficient is .49.



We can picture the relationship between a predictor and a criterion in various ways. For example, the bar chart in Figure 5–1 shows the percentage of candidates passing air force pilot training at each of nine score levels on a test battery designed to assess aptitude for flight training. The chart shows a steady increase in the percentage passing as we go from low to high scores, so successful completion of training is related to aptitude level. A similar chart could be produced to show the relationship between scores on a college admissions test and subsequent grades in college or between clerical aptitude test scores and success in clerical training. The relationship pictured in the chart corresponds to a correlation coefficient of .49. A higher correlation would result in a greater difference in success rates for increasing levels of aptitude score.

The Problem of the Criterion

We have said that empirical validity can be estimated by determining the correlation between test scores and a suitable criterion measure of success on the job or in the classroom. The joker here is the phrase “suitable criterion measure.” One of the most difficult problems that the investigator of selection tests faces is that of locating or creating a satisfactory measure of success to be used as a criterion measure for test validation. This problem is most serious in employment settings but is also quite troublesome in the academic environment. It might appear that it should be a simple matter to decide on some measure of rate of production or some type of rating by supervisors to serve as a criterion measure. It might also seem that this measure, once decided on, should be obtainable in a simple and straightforward fashion. Unfortunately, identifying a satisfactory criterion measure is not so easy. Finding or developing acceptable criterion measures usually involves the tests-and-measurements research worker in a host of problems.

Each possible type of criterion measure presents its own problems. A record of actual performance has a good deal of appeal—number of widgets made, freedom from errors as a cashier, or amount of insurance sold, for example. But many jobs or positions, such as physician, teacher, secretary, or receptionist, yield no objective record of the most important elements of performance or production. And, when such records do exist, the production is often influenced by an array of factors that are outside the individual's control. The production record of a lathe operator may depend not only on personal skill in setting up and adjusting the lathe, but also on the condition of the equipment, the adequacy of the lighting in the work environment, or the quality of the material being machined. The sales of an insurance agent are a function not only of individual effectiveness as a seller, but also of the territory to be covered and the supervision and assistance the agent receives.

Because of the absence or the shortcomings of performance records, personnel psychologists have often depended on some type of rating by a supervisor as a criterion measure. Such ratings may exist as a routine part of the personnel procedures in a company, as when each employee receives a semiannual efficiency report, or the ratings may have to be gathered especially for the selection study. In either event, the ratings will typically be found to depend as much on the person giving them as on the person being rated. Ratings tend to be erratic and influenced by many factors other than performance. The problems involved with using rating procedures to describe and evaluate people are discussed in Chapters 10 and 11.

There are always many criterion measures that might be obtained and used for validating a selection test. In addition to using quantitative performance records and subjective ratings, we might also use later tests of proficiency. This type of procedure is involved when a college entrance mathematics test is validated in terms of its ability to predict later performance on a comprehensive examination on college mathematics. Here, the comprehensive examination serves as the criterion measure. Another common type of criterion measure is the average of grades in some type of educational or training program. Tests for the selection of engineers, for example, may be validated against course grades in engineering school.

All criterion measures are partial in the sense that they measure only a part of success on the job or only the preliminaries to actual job or academic performance. This last point is true of the engineering school grades just mentioned, which represent a relatively immediate—but partial—criterion of success as an engineer. The ultimate criterion would be some appraisal of lifetime success in a profession. In the very nature of things, such an ultimate criterion is inaccessible, and the investigator must be satisfied with substitutes for it. These substitutes are never completely satisfactory. The problem is always to choose the most satisfactory measure or combination of measures from among those that it appears feasible to obtain. The investigator is then faced with the problem of deciding which of the criterion measures is most satisfactory. How should we identify the most satisfactory measures?

Qualities Desired in a Criterion Measure

Four qualities are desired in a criterion measure. In order of their importance, they are (1) relevance, (2) freedom from bias, (3) reliability, and (4) availability.

We judge a criterion measure to be *relevant* to the extent that standing on the criterion measure corresponds to, or exemplifies, level or status on the trait we are trying to predict. In appraising the relevance of a criterion, we must revert to rational considerations. No empirical evidence is available to tell us how relevant freshman grade-point average is, for example, as an indicator of someone having achieved the objectives of Supercolossal University. For achievement tests, therefore, it is necessary to rely on the best available professional judgment to determine whether the content of a test accurately represents our educational objectives. In the same way, with respect to a criterion measure, it is necessary to rely on professional judgment to provide an appraisal of the degree to which some available criterion measure can serve as an indicator of what we would really like to predict. We could even say that relevance corresponds to the content validity of the criterion measure. A relevant criterion closely corresponds to the behaviors of ultimate interest.

The second most important factor in the criterion measure is *freedom from bias*. By this, we mean that the measure should be one on which each person has the same opportunity to make a good score or, more specifically, one on which each equally capable person obtains the same score (except for errors of measurement), regardless of the group to which he or she belongs. Examples of biasing factors are such things as variation in wealth from one district to another for our previous example of the insurance agent, variation in the quality of equipment and conditions of work for a factory worker, variation in "generosity" of the bosses who are rating private secretaries, or variation in the quality of teaching received by students in different classes. To the extent that the criterion score depends on factors in the conditions of work, in the evaluation of work, or in the personal characteristics of the individual, rather than on status on the trait of interest, there is no real meaning to the correlation between test results and a criterion score. A criterion measure that contains substantial bias cannot at the same time reveal relevant differences among people on the trait of interest. For a thorough discussion of issues surrounding bias and test fairness, see Camilli (2006).

The third factor is *reliability* as it applies to criterion scores. A measure of success on the job must be stable or reproducible if it is to be predicted by any type of test device. If the criterion score is one that jumps around in an unpredictable way from day to day, so that the person who shows high job performance one week may show low job performance the next, or the person who receives a high rating from one supervisor gets a low rating from another, there is no possibility of finding a test that will predict that score. A measure that is completely unstable cannot be predicted by anything else.

Finally, in choosing a criterion measure, we always encounter practical problems of *convenience and availability*. How long will we have to wait to get a criterion score for each individual? How much is it going to cost—in dollars or in disruption of normal activities? Though a personnel research program can often afford to spend a substantial part of its effort getting good criterion data, there is always a practical limit. Any choice of a criterion measure must take this practical limit into account.

The Practice of Prediction

In Chapter 2 we saw that we could use the regression line to make the best possible prediction of a person's score on one variable from their score on another. This same regression line provides the most accurate, and therefore most valid prediction of people's scores on a criterion variable. The square of the correlation between the two variables tells us the strength of the relationship; specifically, it shows how much of the variation in scores on the criterion variable is predictable from the predictor. We shall shortly see how we can express the remaining uncertainty in criterion scores.

Interpretation of Validity Coefficients

Suppose that we have gathered test and criterion scores for a group of individuals and computed the correlation between them. Perhaps the predictor is a scholastic aptitude test, and the criterion is an average of college freshman grades. How will we decide whether the test is a good predictor?

Obviously, other things being equal, the higher the correlation, the better. In one sense, our basis for evaluating any one predictor is in relation to other possible prediction procedures. Does Peter's Perfect Personnel Predictor yield a higher or lower validity coefficient than other tests that are available? Does it yield a higher or lower validity coefficient than other types of information, such as high school grades or ratings by the school principal? We will look with favor on any measure whose validity for a particular criterion is higher than that of measures previously available, even though the measure may fall far short of perfection.

A few representative validity coefficients are exhibited in Table 5–2. These give some picture of the size of correlation that has been obtained in previous work of different kinds. The investigator concerned with a particular course of study or a particular job will, of course, need to

Table 5–2
Validity of Selected Tests as Predictors of Certain Educational and Vocational Criteria

Predictor Test[a]	Criterion Variable[b]	Validity Coefficient
CogAT	TAP reading (Grade 12)	.79
Verbal	TAP social studies (Grade 12)	.78
Quantitative	TAP mathematics (Grade 12)	.79
ITED composite (Grade 9)	Cumulative high school GPA	.49
	College freshman GPA	.41
	SAT total	.84
ITBS composite (Grade 6)	Final college GPA	.44
Seashore Tonal Memory Test	Performance test on stringed instrument	.28
Short Employment Test		
Word knowledge score	Production index—bookkeeping machine operators	.10
Arithmetic score	Production index—Bookkeeping machine operators	.26

[a]CogAT = Cognitive Abilities Test; ITED = Iowa Tests of Educational Development; ITBS = Iowa Tests of Basic Skills.
[b]TAP = Tests of Achievement and Proficiency; GPA = grade-point average; SAT = Scholastic Aptitude Test.

become familiar with the validities that the tests being considered have been found to have for the criterion measure to be used.

The usefulness of a test as a predictor depends not only on how well it correlates with a criterion measure but also on how much *new* information it gives. For example, the social studies subtest of the Tests of Achievement and Proficiency (TAP) was found to correlate on the average .51 with ninth-grade social studies grades, and the reading comprehension subtest to correlate .51 with the same grades. But, the two tests have an intercorrelation of .77. They overlap and, in part at least, the information either test provides is the same as that provided by the other test. The net result is that pooling information from the two tests can give a validity coefficient of no more than .53. If the two tests had been uncorrelated, each giving evidence completely independent of the other, the combination of the two would have provided a validity coefficient of .70. Statistical procedures have been developed that enable us to determine the best weighting to give to two or more predictors and to calculate the regression equation and correlation that result from this combination. (The procedures for computing the weights for the predictor tests [regression weights] and the correlation [multiple correlation] resulting from the combination are specialized topics in statistics. A complete presentation of these methods can be found in intermediate statistics texts such as those of Cohen, Cohen, West, & Aiken, 2003; McClendon, 1994; and Tabachnick & Fidell, 2007.)

Clearly, the higher the correlation between a test or other predictor and a criterion measure, the better. But, in addition to this relative standard, we should like some absolute one. How high must the validity coefficient be for the test to be useful? What is a “satisfactory” validity? This last question is a little like asking “How high is up?” However, we can try to give some sort of answer.

To an organization using a test as a basis for deciding whether to hire a particular job applicant or to admit a particular student, the significant question is “How much more often will we make the right decision on whom to hire or to admit if we use this test than if we operate on a purely chance

Table 5–3
Two-by-Two Table of Test and Job Success

		Performance on the Job		
		Bottom Half— "Failures"	Top Half— "Successes"	Total
Score on Selection Test	Top half (accepted)			100
	Bottom half (rejected)			100
	Total	100	100	200

basis or on the basis of some already available but less valid measure?" The answer to this question depends in considerable measure on the proportion of individuals who must (or can) be accepted, called the **selection ratio,** and the prevalence of "success" in the population, called the **base rate.** A selection procedure can do much more for us if we can accept only the individual who appears to be the best 1 in every 10 applicants than if we must accept 9 out of 10. However, to provide a specific example, let us assume that we will accept half of the applicants. Let us examine Table 5–3.

Table 5–3 is set up to show 200 people in all: 100 in each half on the test and 100 in each half on the job. If there were absolutely no relationship between the test and job performance, there would be 50 people in each of the four cells of the table. Defining "success" as being in the top half on the job (a base rate of .50), the success rate would be 50 in 100 for those accepted and also for those rejected. There would be no difference between the two, and the correlation between the selection test and job performance would be zero.

Table 5–4 shows, for correlations of different sizes and a selection ratio of .50, the percentage of correct choices (i.e., "successes") among the 50% we accept. A similar percentage of correct rejections occurs in the 50% who are not accepted. The improvement in our "batting average" as the correlation goes up is shown in the table. Thus, for a correlation of .40, we will make correct decisions 63.1% of the time and be in error 36.9% of the time; with a correlation of .80 the percentage of correct decisions will be 79.5, and so forth.

Table 5–4
Percentage of Correct Assignments When 50% of Group Must Be Selected

Validity Coefficient	Percentage of Correct Choices
.00	50.0
.20	56.4
.40	63.1
.50	66.7
.60	70.5
.70	74.7
.80	79.5
.90	85.6