Introduction

At its heart, the process of validating pre-employment tests is all about amassing evidence demonstrating that a test score is related to a target job in some meaningful fashion that permits distinguishing between those who are likely to perform well from those who are not. As a result, in the truest sense of the word, it is not a test cannot be said to be valid or not valid but rather the conclusions draws from the individual test scores. Validity, therefore, refers to the determination—based on the available evidence—as to whether the test scores serve as an effective a proxy that is indicative of an examinee’s likely success on the job.

In other words, determining a test’s validity is a lot like presenting a legal case to a jury. The question you should keep in mind is, “Does the evidence suggest, beyond a reasonable doubt, that a person’s test score is an acceptable estimate of his or her likely future job performance?” Continuing the analogy, it is obvious that some kinds of evidence carry more weight than others. DNA or eye witness evidence, for example, is usually considered strong. Circumstantial evidence is generally considered weak—unless there is so much of it that it tips the balances toward being highly convincing.

There are a number of strategies that have been proposed for demonstrating a test’s validity (many of them newer and more reliant on some of the more advanced statistical or theoretical approaches). To learn about these strategies, the reader is encouraged to review the Society for Industrial and Organizational Psychology’s Principles for the Validation and Use of Personnel Selection Procedures (available from www.siop.org) or the American Psychological Association’s Standards for Educational and Psychological Testing (available from www.apa.org).

Perhaps the most widely recognized guide for the validation and use of pre-employment tests is the Federal Uniform Guidelines on Employee Selection Procedures (1978). While allowing for the development of other approaches to establishing test validity as the state-of-the-art develops, specifically endorses three general test validation strategies. These three approaches are a) content validation, b) criterion-related validation, and c) criterion validation. Because the courts tend to defer to the Guidelines as the standard of practice, most employers utilize one or more of these strategies to help reduce their exposure to legal liability resulting from charges of employment discrimination.
Criterion Validation: An Overview

This brief article discusses criterion validation; an approach used to validate approximately 15% of US pre-employment tests (Biddle, 2007). Criterion validation is a strategy that uses statistical techniques to demonstrate that an examinee’s performance on a test (his/her test score) is predictive of their actual job performance as measured by some criterion (e.g., job performance, number of products assembled, number of calls handled, etc.)

It should be obvious that this is a very different approach from the most common validation strategy—content validation—which relies on the assertion that the pre-employment test is measuring the actual knowledge, skills and abilities required on the job as identified through a job analysis.

Since criterion validation focuses on the relationship between test score and job performance, the key to validation using this approach is the use of a statistical test known as the “correlation coefficient” where each examinee’s test score is correlated with his or her job performance. If the correlation coefficient equals or exceeds $r=0.20$, it means the test is sufficiently related to job performance to make judgments about a candidates likelihood of job success based on his or her test score. In other words, you can use test score to determine who is appropriate to hire.

Note that a correlation of $r=0.20$ is the minimum that should be considered acceptable. Ordinarily, you would like to see a correlation that is larger—which indicates the relationship between test score and job performance is stronger. The larger the correlation, the better! To illustrate the importance of the size of the correlation between test score and job performance consider the following example.

A correlation coefficient can be converted to a statistic called to coefficient of determination by squaring it (i.e., multiplying it by itself). The coefficient of determination (also called r-squared) tells you the percent of variation in one score that is related to the variation in the other score. In other words, it can tell you the percent of variation in job performance that tends to be related to—or explained by—the variation in his or her test score.

If there is a correlation of 0.20 between a group of test scores and job performance, it means that about 4% of job performance is related to test performance (0.20 multiplied by 0.20). Stated differently, about 96% of the variation in job performance seems to be related to factors other than the test score! While it could be argued that the test, while helping a little to identify effective employees, it is not measuring a very large amount of what makes a person a good employee. In other words, it is a weak predictor of performance and many might question its use for determining who is hired and who is not.

On the other hand, if the correlation between a pre-employment test and job performance is 0.60, the coefficient of determination becomes 36% (0.60 multiplied by 0.60). In this case the test is explaining 36% of the variation in employee performance. Thus, making hiring decisions based on scores from the second test will be much more defensible because it is demonstrably more related to—and therefore predictive of—job performance.
How To Conduct A Criterion Validation Study

On the face of it, conducting a criterion validation study appears to be quite simple. All that is needed is a pre-employment test and a measure of job performance. Step one is to administer the test to a large sample of current or potential employees. Step 2 is to collect performance ratings on the criteria of interest (e.g., job performance, customer service ratings, number of phone calls fielded in a given time). Finally, step 3 requires that the correlation between test score and job performance rating be computed. If the correlation is .20 or greater, the test can be said to be at least marginally valid.

As simple as this sounds, one might ask, “Why doesn’t everyone use this validation approach?” There are several very good reasons. These are briefly presented below:

1. Knowledge of statistics: Because criterion validation is a purely statistical process, it is important that anyone who plans to use this approach become well informed with respect to the issues, problems, requirements and practices of establishing relationships between variables using a correlation coefficient.

2. Sample size requirements. Statistics such as correlation coefficients require minimum sample sizes to be considered both valid and stable (i.e., repeatable). Considering the need to demonstrate that a test is valid across shifts, geographic work locations, gender and ethnic groups, etc., BCG typically recommends a sample size of at least 300.

3. Requirement that scores from both the test and criterion be available. In order to calculate correlation coefficients, two values are required. In the case of criterion validation, each test score must also have an associated job performance/criterion measure. This means that the sample size requirement of 300 listed above requires 300 hires—not just 300 test takers. Therefore, unless an organization is large, the sample size requirement alone tends to result in criterion validation not being an acceptable option for validating pre-employment tests.

4. Criterion measures are frequently of low quality. Criterion validation requires a measure of performance that is to be correlated with test scores. For many jobs, performance ratings are of questionable validity. For example, supervisor ratings tend to be influenced by factors other than job performance. These influences can range from a tendency of some supervisors to rate employees more harshly while others tend to rate them more highly. Occasionally, supervisors unconsciously rate employees who are similar to themselves highly while rating others lower. The best type of performance measure to use is one that is highly objective such as number of products manufactured, number of customer service calls handled, etc.

5. Restriction of range. It is not uncommon for supervisors to rate the majority of their staff as simply “average” or “meets expectations” with the result being little variability in ratings. This tendency may even be built into the performance appraisal system in that any deviation from “meets expectations” requires documentation which tends to discourage using the entire range of performance scores. In this case, a test which
typically has significant amounts of variation (people score from very low to very high) will appear to have little or no validity in terms of predicting performance not because it is invalid but because there is a problem with the way job performance was measured.

6. From a theoretical perspective it is difficult to develop tests. This pitfall of criterion validation warrants an article all on its own. In its simplest form, the issue is as follows. A valid pre-employment test must be reliable, meaning that it provides consistent results. A reliable test tends to be one that is “homogeneous” meaning that it consists of items that tend to be related to each other and therefore measure the same general knowledge domain. However, a test constructed of homogeneous items tends to have a low correlation between its total test score and a job-related criterion because all of the items are measuring a very similar construct. Therefore, a test that has maximal predictive validity must consist of items with low correlations among each other (resulting in a test that is not homogeneous) and yet each item has a high correlation with the criterion. As a result, a very powerful predictive test will look strange because it will consist of items that have little or nothing to do with each other. This will make the test less acceptable to test-takers. Furthermore, there is no theory to drive, in a consistent manner, an employment test containing such items.

From this article, it should be clear that criterion validation as a strategy for demonstrating the validity of using test scores to make hiring decisions. The approach can be quite effective. It is easy for people (i.e., members of the jury) to understand that the relationship between test score and job performance is such that as test score increases job performance also tends to increase. As a result, hiring those with the highest test score is likely to result in hiring those who will perform the best.

It should also be clear, however, that there are a number of methodological and practical issues that make criterion validation less than ideal for the vast majority of employers. These include the requirements for large sample sizes, a lack of high-quality performance measures, the need to have a strong background in the application of statistics to ensure that all relevant statistical issues are identified and addressed, and the theoretical difficulties in constructing heterogeneous tests that are highly predictive which still appearing to be job related.