KCP learning

TUA

factsheet 26: test validity

Test Validity

We have looked at the importance of standardisation and test reliability. However, a test may be both standardised and reliable, but still useless for a particular assessment purpose.

For example, an assessment criterion for a senior management position might be "Ability to Understand and Use Complex Numerical Data to Make Well Reasoned Decisions".

If a test is constructed to measure advanced numerical reasoning, but the test consists of questions calling only for arithmetic computation, we may find that the test has an acceptable reliability but does not measure what it is supposed to measure. In other words, it measures something (arithmetic computation) but it is not a valid predictor of numerical reasoning.

What is Test Validity?

The validity of a test is therefore concerned with the extent to which it actually measures what it is supposed to measure. In assessment, a test's validity is a measure of it's relevance to the job (or training) content. Our aim in evaluating a test in terms of its validity is to determine the extent to which the test scores will predict something about the individual's performance or satisfaction in a particular job.

Forms of Validity

Face Validity is concerned with the extent to which a test appears to have relevance to a particular job. Face validity is important since poor perceived relevance can cause resistance on the part of the candidates. A number of older personality questionnaires have been criticised for containing questions with low perceived relevance: for example, the 16PF (R B Cattell's personality questionnaire) includes the item "If I had a gun in my hand that I knew was loaded, I would feel nervous until I unloaded it", to which candidates must respond "Yes", "In between", or "No". An individual applying for an administrative position may ask "What has this question to do with my job application?".

Faith Validity is sometimes used synonymously with Face Validity, although strictly speaking the two are slightly different. Faith validity is often taken to mean that we decide to use a test because we want to believe that it work.

Content Validity is similar to face validity in that it concerns the extent to which test items are representative of the attribute to be measured. Content validity can be assessed in a structured way by performing a detailed job analysis, thereby deriving descriptions of the job content, which can then be matched with the test content. In a simple case, it may be discovered that a clerical job involves quick and accurate checking of detailed written information: if a test samples this ability by presenting candidates with items requiring fast and accurate checking, its content validity would be high.

KCP learning TUA

factsheet 26: test validity (cont'd)

An ability test may be said to have content validity, provided that the cognitive processes involved in the test are similar to cognitive processes required for successful performance on the job. This means that the superficial detail of the test may be different from the job content so long as the underlying processes required are similar. For example, a checking test may have content validity for many clerical positions which involve quick and accurate checking, even if these positions have nothing to do with subject matter of the test.

The distinction, then, between face and content validity is that the latter requires that aspects of the job/training content are actually represented in the test content, whereas the former (face validity) merely requires that the test content superficially **appears** to be related to job/training content.

Content validity, in particular, is a crucial attribute of a good test. Many instruments are selected for use solely on the basis of their reliability and content validity, the argument being that if a test's content represents elements of the job or training, and the test is a reliable measuring instrument, it is justifiable to use the test as an assessment instrument.

However, neither face nor content validity provides a statistical (or "empirical") evaluation of the validity of a test. In order to obtain a **validity coefficient** we must use a criterion-related method of evaluating validity.

Criterion-Related Validity Like reliability, the validity of a test, or questionnaire scale, can be expressed as a correlation coefficient, and is referred to as a **validity coefficient.**

In order to obtain such a coefficient, we must correlate the test scores of a given sample with a **criterion measure** for the same sample. This criterion measure would typically be a measure of job performance (e.g. ratings by managers or supervisors), but could also be a different kind of criterion such as attendance, length of tenure, rapidity of promotion, training success/failure and so on.

There are two main types of criterion-related validity, **concurrent validity** and **predictive validity**.

Concurrent validity is based on the correlation between test scores and the criterion-measure when both sets of data are collected at the same time (i.e. concurrently). This means that the sample is made up of job incumbents who are usually taking the test solely for the purpose of test validation.

Predictive validity involves correlating test scores with a criterion-measure collected subsequently to the testing session. Typically test scores are obtained from applicants at the time of their selection. Criterion data is then collected for the successful applicants at an appropriate time (e.g. 1 year after recruitment).

factsheet 26: test validity(cont'd)

Pros and cons of concurrent and predictive validation

Concurrent validity has the following advantages:

- It provides an expedient means by which a validation coefficient can be obtained. There is no delay in obtaining data.
- It is a highly appropriate means of validating instruments used for assessment of current attributes (e.g. instruments used in management development to diagnose current strengths and limitations).

Disadvantages of the concurrent method are:

- If the test is intended for use in selection concurrent validity is theoretically less appropriate than predictive validity because a concurrent validation result cannot show that the test is related to future performance. Concurrent validity is therefore less suited to instruments used to assess potential, rather than current attributes.
- Concurrent validity is based on a sample of job incumbents. In completing the tests incumbents may differ from job applicants in terms of their motivations. In particular, incumbents completing self-report questionnaires will tend to be less motivated to give socially desirable" responses than job applicants. Validity based on job incumbents cannot, therefore, be directly generalised to the selection situation.

Predictive validation has the advantage that:

- It demonstrates relationships between test scores and a future criterion (such as job performance). Predictive validity is therefore well suited to tests used in selection.

However, the disadvantage of predictive validation is that:

- There is a delay in obtaining results. For this reason concurrent validation studies are much more popular with test users in organisations than predictive studies.
- Criterion-related validities (concurrent and predictive) are highly susceptible to the problems of: restriction in range and poor quality of criterion reliability

Construct validity The construct validity (sometimes called "construct-related" validity) of a test concerns the extent to which the test measures a theoretical construct or trait, such as intelligence, neuroticism, mechanical comprehension. Such constructs are developed to theoretically account for observed patterns of responses or behaviours. Construct validity involves the accumulation of any data which clarifies the nature of the construct measured by the instrument.

For example, if a general intelligence test could be shown to correlate strongly with a wide range of tests of other abilities, this would provide a form of construct validity supporting the notion that the test measures general intelligence. Equally if factor analysis of items in a personality questionnaire revealed a factor consisting of questions relating to extraversion, and a scale made up of these questions correlated with other self-report measures of extraversion, these findings would suggest that these questions together measure a construct of extraversion.

Construct validity is of central importance in psychometrics.

KCP learning

TUA

factsheet 26: test validity (cont'd)

Evaluating tests in terms of validity

Evaluation of face or content validity are qualitative processes as described in Factsheet 36.

Validity coefficients are, in practice, often quite low. It is not uncommon for a concurrent validation coefficient to be in the range **0.2 to 0.3, and it is quite unusual to find validity coefficients higher than 0.5.**

This is partly because validity is limited by the reliability of the test itself, and the reliability of the criterion. It may also be partly due to restriction of range caused by selection effects. However, it is also because a single attribute (such as verbal reasoning ability) is normally only one of several or many attributes required for success. Managers with good verbal reasoning ability may not be successful because of other limitations (e.g. in interpersonal skills). The correlations of test scores with criterion will thus be weakened by "exceptions" from the general tendency of high scorers to be better performers.

A test may well be useful in assessment even if its validity is as low as 0.2.

This notion is illustrated in Factsheet 27 which shows that a trainability test correlates with training success at 0.30 (i.e. the validity coefficient is 0.30), but the expectancy table shows that **80% of candidates who scored above average on the test also performed above average in training**.

Validity generalisation

When tests are correlated with performance measures on apparently similar jobs, it has often been noted that validity coefficients vary considerably from one job to another.

This led to a view that validity is situation-specific. However, it has since been shown that much of this variability may be attributable to statistical artefacts arising from sampling error (small sample sizes), poor criterion reliability, and restriction of range. (Note: these same problems have also resulted in valid tests being underestimated or mistakenly rejected).

New statistical techniques were developed to combine data from many different studies. This work demonstrated that the validity of tests of verbal, numerical and reasoning aptitudes can in fact be generalised across jobs for more than was originally realised. In other words, performance in many occupations appears to depend, at least to some extent, on a relatively small number of core abilities.

Most psychologists now accept the concept of validity generalisation; that is, if a test is shown to be valid for one job in one organisation, it is likely to also be valid for similar jobs in other organisations.

KCP learning TUA

factsheet 26: test validity (cont'd)

We have described how a validity coefficient can be an underestimate of the true validity of a test if it is undermined by problems such as criterion-unreliability or restriction of range due to selection effects.

However, validity coefficients can be corrected for both of these problems. Correction formulas can be applied and **result in an increased, corrected validity coefficient** which represents a more realistic estimate of the true validity of the test than the obtained (uncorrected) coefficient.

This training programme does not require delegates to be competent in the application of these formulae. It is, however, important to know of their existence and they are included below for reference.

Correction for criterion unreliability

This is sometimes known as "correction for attenuation" and requires an estimate of the reliability of the criterion data (e.g. job performance ratings). One way of arriving at such an estimate is to collect performance ratings for each test-taker from two separate managers or supervisors. The correlation between the two sets of ratings provides a measure of the criterion-reliability.

The correction formula is:

$$rca = \frac{rxy}{\sqrt{ryy}}$$

where rca = validity coefficient corrected for attenuation
rxy = uncorrected validity coefficient
ryy = estimated reliability of criterion

Correction for restriction in range

Like the formula for correcting reliability for restriction of range (see Section 15.8), this formula depends on knowledge of the test standard deviation of both the restricted and non-restricted groups. In a validation study all applicants tested would typically comprise the non-restricted group. The restricted group consists only of those selected candidates who are included in the validation sample.

The formula for estimating the unrestricted validity from the restricted validity coefficient is:

$$rcr = \frac{v.rxy}{\sqrt{v_2.rxy_2 - rxy_2 + 1}}$$

where rcr = validity coefficient for restriction of range
rxy = applicant validity coefficient
rxy_2 = validation group validity coefficient
v = SD of test in unrestricted (applicant group)
v_2 = SD of test in restricted (validation group)