Research Psychometric Properties

There are two important psychometric properties – reliability and validity. Reliability deals with the extent to which a measure is repeatable or stable. Essentially reliability refers to the consistency of a measure. A reliable measure would return the same results from time to time, assuming the underlying phenomenon which we were measuring has not changed. Reliability is of particular importance because it establishes a ceiling for validity. A measure can be no more valid than it is reliable. Methods for establishing reliability include:

- **Test re-test** – As depicted in the graphic below, test re-test reliability consists of administering the measure at Time 1, then after some period of time administering it again at Time 2. A reliable measure would produce similar results at both points in time. On a job satisfaction measure those who scored high at Time 1 would also score high at Time 2 while those who scored low at Time 1 would also score low at Time 2. The problem with this method of examining reliability is that if sufficient time has not elapsed between Time 1 and Time 2 individuals may simply remember their Time 1 responses and instead of reliability being examined it’s simply individual memory that’s being examined. On the other hand, should too much time elapse between the two measures, it might appear that the measure in not reliable when actually the phenomenon being measured has changed between Time 1 and Time 2.

- **Parallel forms** – Similar to test re-test reliability, parallel forms consists of two versions of the measure and one version is administered at Time 1 while the second version is administered at Time 2. The memory problems associated with the test re-test method of examining reliability are corrected with this method in that the individuals have not seen the version 2 items previously. One difficulty of parallel forms reliability is that it requires developing a considerable number of items that measure the same thing. Parallel forms reliability is depicted below.
Internal consistency – This method of establishing reliability essentially relies upon how well the several items within the measure “hang together.” Suppose that a measure of customer satisfaction consists of 10 statements regarding how satisfied the customer is with service, prices, etc. Basically the reliability of the measure is examined by estimating how well the items produce similar results. The graphic below depicts the relationship between Item 1 and Item 2, but to establish Internal Consistency reliability we’d hope to see similar relationships between all possible pairs.

Inter-rater – This form of reliability is examined when there are multiple judges or raters evaluating the same phenomenon. Inter-rate reliability is essentially the extent of consensus among the raters. When there is a high degree of consensus it would be evidence of inter-rater reliability.

Validity refers to the extent to which a study actually captures or measures what it purports to examine. Two broad categories of validity are external and internal validity. External validity refers to the generalizability of the research findings. For example, would our findings from a particular study also apply to other groups, other geographical
areas, other types of jobs, etc.? Internal validity essentially refers to how confident the researcher is in the study results. This type of validity encompasses a range of issues from how rigorous the study was in terms of research design, sample, measures, etc. to whether the measures are measuring what they are supposed to measure. In this description of validity we will focus on this second area, the extent to which measures capture what they are purported to capture. Although the following will be focused on examining internal validity in an employment testing setting, the discussion is appropriate in a variety of settings. Techniques for examining internal validity include:

- **Face** – The concern of face validity is to not alienate the test taker. Basically face validity is to what extent, to the naïve, casual test-taker does the test measure what it purports to measure. If an employee is completing a measure of job satisfaction, but realizes that there are items that appear to be geared toward honesty or integrity, the validity of the job satisfaction measure may be diminished because of the lack of face validity.

- **Content** – Content validity focuses on the extent to which an employment test samples the work domain. For example, a typing test would be content valid when testing applicants for the position of typist.

- **Construct** – A construct is a mental definition or abstraction such as intelligence, job satisfaction, customer satisfaction, or organizational commitment. Construct validity refers to how well a measure of a construct actually measures the underlying construct such as how well does an IQ test actually measure intelligence.

- **Criterion-related** – There are two types of criterion-related validity in employment testing; concurrent and predictive. In concurrent criterion-related validity the test or measure that is thought to be predictive of performance is administered to current employees and the results on the test are compared with some criterion of performance such as quantity produced, sales, or supervisory evaluations. Ideally we’d like to see a very positive relationship so that higher performers score higher while lower performers score lower. If that were the case, we could conclude that our measure predicts performance and that it is a valid test and we could use it for employee selection. However, frequently the results in our concurrent test are as depicted below
In the above example, there is very little relationship between how employees scored on the test and their job performance. A problem that sometimes occurs with concurrent criterion-related validity studies is known as restriction of range. In our current employee group we don’t have any poor or even below average employees; they’ve all been terminated or quit because they couldn’t meet performance requirements or they weren’t even hired in the first place. If they could have been included in our assessment, our results might have looked like:

As is obvious from the above graph, there is clearly a relationship between our test and performance and we would have concluded that the test was valid had we had the full range of employees in our study.

A second type of criterion-related validity assessment is known as predictive. In a predictive study applicants, rather than current employees, are administered our measure that we think will be predictive of performance. The test results are not used as a basis upon which to hire, the actual selection is based upon some other factors or all applicants are hired. After a suitable time (e.g. six months or a year), performance of those individuals is determined and is compared with
their test scores from the test we’re seeking to validate. If there is a strong relationship between the two, we’d conclude the test is valid.