RELIABILITY AND VALIDITY

Definition of Reliability:

Reliability refers to the consistency of a measure. A test is considered reliable if we get the same result repeatedly. For example, if a test is designed to measure a trait (such as introversion), then each time the test is administered to a subject, the results should be approximately the same.

Reliability can be split into two main branches: internal and external reliability.

Internal reliability

This describes the internal consistency of a measure (i.e. consistency within itself), such as whether the different questions (known as 'items') in a questionnaire are all measuring the same thing.

One way to assess this is by using the split-half method, where data collected is split randomly in half and compared, to see if results taken from each part of the measure are similar. It therefore follows that reliability can be improved if items that produce similar results are used.

External reliability

This assesses consistency when different measures of the same thing are compared, i.e. does one measure match up against other measures?

Discrepancies will consequently lower inter-observer reliability, e.g. results could change if one researcher conducts an interview differently to another.

Such reliability issues can be improved by standardising procedures (i.e. making sure that procedures are carried out the same way each time), for instance by implementing interviewer training, and/or practice through pilot studies.

Test-Retest Reliability

Test-retest reliability is a measure of the consistency of a psychological test or assessment. This kind of reliability is used to determine the consistency of a test across time. Test-retest reliability is best used for things that are stable over time, such as intelligence.

Test-retest reliability is measured by administering a test twice at two different points in time. This type of reliability assumes that there will be no change in the quality or construct being measured.[2] In most cases, reliability will be higher when little time has passed between tests.

The test-retest method is just one of the ways that can be used to determine the reliability of a measurement. Other techniques that can be used include inter-rater reliability, internal consistency, and parallel-forms reliability.

Inter-Rater Reliability

This type of reliability is assessed by having two or more independent judges score the test.[3] The scores are then compared to determine the consistency of the raters estimates.

One way to test inter-rater reliability is to have each rater assign each test item a score. For example, each rater might score items on a scale from 1 to 10. Next, you would calculate the correlation between the two ratings to determine the level of inter-rater reliability.

Another means of testing inter-rater reliability is to have raters determine which category each observation falls into and then calculate the percentage of agreement between the raters. So, if the raters agree 8 out of 10 times, the test has an 80% inter-rater reliability rate.

It is important to note that test-retest reliability only refers to the consistency of a test, not necessarily the validity of the results.

Parallel-Forms Reliability

Parallel-forms reliability is gauged by comparing two different tests that were created using the same content.[4] This is accomplished by creating a large pool of test items that measure the same quality and then randomly dividing the items into two separate tests. The two tests should then be administered to the same subjects at the same time.

Internal Consistency Reliability

This form of reliability is used to judge the consistency of results across items on the same test.[1] Essentially, you are comparing test items that measure the same construct to determine the tests internal consistency.

When you see a question that seems very similar to another test question, it may indicate that the two questions are being used to gauge reliability.

Because the two questions are similar and designed to measure the same thing, the test taker should answer both questions the same, which would indicate that the test has internal consistency.

Definition of Validity:

Validity is a measure of how well a test measures what it claims to measure.[1]

Psychological assessment is an important part of both experimental research and clinical treatment. One of the greatest concerns when creating a psychological test is whether or not it actually measures what we think it is measuring.

For example, a test might be designed to measure a stable personality trait but instead, measure transitory emotions generated by situational or environmental conditions. A valid test ensures that the results are an accurate reflection of the dimension undergoing assessment.[2]

Validity is the extent to which a test measures what it claims to measure. It is vital for a test to be valid in order for the results to be accurately applied and interpreted.

Types of Validity

Validity isn't determined by a single statistic, but by a body of research that demonstrates the relationship between the test and the behavior it is intended to measure. There are three types of validity.

Content Validity

When a test has content validity, the items on the test represent the entire range of possible items the test should cover.[3] Individual test questions may be drawn from a large pool of items that cover a broad range of topics.

In some instances where a test measures a trait that is difficult to define, an expert judge may rate each item's relevance. Because each judge is basing their rating on opinion, two independent judges rate the test separately. Items that are rated as strongly relevant by both judges will be included in the final test.

Criterion-Related Validity

A test is said to have criterion-related validity when the test has demonstrated its effectiveness in predicting criterion or indicators of a construct, such as when an employer hires new employees based on normal hiring procedures like interviews, education, and experience.[4]

This method demonstrates that people who do well on a test will do well on a job, and people with a low score on a test will do poorly on a job. There are two different types of criterion validity:

- **Concurrent validity**: This occurs when criterion measures are obtained at the same time as test scores,[5] indicating the ability of test scores in estimating an individual's current state. For example, on a test that measures levels of depression, the test would be said to have concurrent validity if it measured the current levels of depression experienced by the test taker.
- **Predictive validity**: This is when the criterion measures are obtained at a time after the test.[6] Examples of tests with predictive validity are career or aptitude tests, which are helpful in determining who is likely to succeed or fail in certain subjects or occupations.

Construct Validity

A test has construct validity if it demonstrates an association between the test scores and the prediction of a theoretical trait.[7] Intelligence tests are one example of measurement instruments that should have construct validity. A valid intelligence test should be able to accurately measure the construct of intelligence rather than other characteristics such as memory or educational level.

Essentially, content validity looks at whether a test covers the full range of behaviors that make up the construct being measured. The procedure here is to identify necessary tasks to perform a job like typing, design, or physical ability.

In order to demonstrate the content validity of a selection procedure, the behaviors demonstrated in the selection should be a representative sample of the behaviors of the job.

Face Validity in Psychological Testing

Another method that is used rarely because it is not very sophisticated is face validity. It is based only on the appearance of the measure and what it is supposed to measure, but not what the test actually measures.

Face validity is one of the most basic measures of validity. Essentially, researchers are simply taking the validity of the test at face value by looking at whether a test *appears* to measure the target variable.[8] On a measure of happiness, for example, the test would be said to have face validity if it appeared to actually measure levels of happiness.

Obviously, face validity only means that the test *looks* like it works. It does not mean that the test has been proven to work. However, if the measure seems to be valid at this point, researchers may investigate further in order to determine whether the test is valid and should be used in the future.

Essentially, face validity is whether a test seems to measure what it is supposed to measure. It involves taking the test at face value.

A survey asking people which political candidate they plan to vote for would be said to have high face validity. The purpose of the test is very clear, even to people who are unfamiliar with psychometrics.

A complex test used as part of a psychological experiment that looks at a variety of values, characteristics, and behaviors might be said to have low face validity. The exact purpose of the test is not immediately clear, particularly to the participants.

Obviously, while face validity might be a good tool for determining whether a test seems to measure what it purports to measure, having face validity alone does not mean that a test is actually valid. Sometimes a test looks like it is measuring one thing, while it is actually measuring something else entirely.

Reference: *Test and Measurement in Research Methodology BY A.K. Singh*

Kelley, T. L. (1927).*Interpretation of educational measurements. New York*: Macmillan.