Reliability

Reliability of a test pertains to reliable measurement which means that the measurement is accurate and free from any sort of error. Reliability is one of the most essential characteristic of a test. If a test gives same result on different occasions, it is said to be reliable. So Reliability means consistency of the test result, internal consistency and consistency of results over a period of time.

According to Anastasi and Ubrina (1982)

"Reliability refers to the **consistency of scores** obtained by the same persons when they are re-examined with the same test on different occasions, or with different sets of equivalent items, or under other variable examining conditions."

TYPES OF RELIABILITY

1. Test Retest Reliability

- It involves administration of the same test at 2 different times.
- It estimates the error related to administering a test at 2 different times

It involves three steps

- Administering a test to a group of individual
- Re administering the same test on the same sample
- Correlating the first set of scores to the 2nd

- Value of correlation tells us about reliability.
- Reliability of any test can be expressed in terms correlation which is reliability co efficient.
- Higher the correlation, higher the reliability.
- This method is used for assessing the permanent traits e.g., intelligence , aptitude etc.
- Such traits do not change over period of time.
- Errors occur due to re administering of a same test.

- Error variance caused due to uncontrolled testing conditions such as sudden noise, changes in weather test taker conditions e.g., fatigue, illness, emotional states of test taker etc.
- Higher test retest reliability illustrates that the scores are free from random daily changes.
- So the scores can be generalized over difference occasions
- Another error is carry over effect (when performance in the second session is influenced by participation in the first session)
- Participants often remember answers of the items in the previous session.
- Carry over effect is not problematic if the changes over time are systematic e.g., if the scores of all participants increase by 5 points.
- When something affects only few participants then it's called random carry over effect.
- These errors can be controlled by choosing appropriate interval between 2 sessions.
- Short interval leads to practice effect.
- Long interval leads to other developmental changes in individuals.
- Appropriate interval is 2 weeks.
- Gap between 2 administrations should be mentioned in test manual.

Advantages

- Gives equivalent test content on both administrations.
- Less difficult to develop 1 form instead of 2.

2. Alternate form Reliability

- It involves administering 2 different tests at 2 different times.
- These both tests are equivalent in terms of content, statistical characteristics.
- Errors in alternate form are item sampling and content sampling.
- Item sampling is related to number of items (it can be controlled by developing equal number of items in both forms). E.g., 30,30 items in form 1 & 2.
- Content sampling is related to content of items (item content should be same in both forms). E.g., if Q#1 is related to test retest reliability then Q#1 of form 2 should be related to same content.

Advantages

- Useful for follow up studies
- Used for Educational tests

- Controlling carry over effect

Disadvantages

- Expensive method
- Time consuming

3. Split Half Reliability

- Test is divided into two equal halves.
- Scores are compared for 1^{st} & 2^{nd} half.
- The results of both halves are compared
- One way of dividing a test in equal halves by dividing it in odd and even numbers.
- E.g., items 1, 3, 5, will be compiled in first half while items 2, 4, 6, will be compiled in second half.
- This method assumes that items are pre-arranged in approx. same difficulty level.
- If items are in form of clusters then each cluster should be placed completely in 1 half.
- E.g., items 1-5 are related to comprehension of paragraph. These items will be placed in first half while items 6-10 are related to another paragraph will be placed in second half.
- The longer the test the more sample behavior it will assess.
- E.g., 100 item test can be divided into 50, 50, items in both halves.

Advantages

- Requires single administration
- Simpler method
- It can manage problems/errors occur in test retest and parallel form.
- 3. Split Half Reliability

Disadvantages

- Many ways to split test and each way gives different reliability estimate.
- E.g., odd, even splitting and simple dividing test in equal halves. Both will lead to different results.

4. Internal Consistency method

- Also called Kuder Richardson formula 20/KR-20.
- Focuses on homogeneity of items
- It measures inter item consistency

- It calculates reliability based on # of items.
- Proportion of responses that are correct and proportion of items that are incorrect.
- E.g., items of DEPRESSION TEST should assess symptoms of depression
- 4. Internal Consistency method

Advantages

- Used for items that are dichotomous (yes/no)
- Formula is simple and less precise

Disadvantages

- Only applicable for tests that are measuring single skill area.

5. Scorer Reliability

- It determines the extent to which the results are objective.
- This method deals with the error related to scorer variance.
    - E.g., some test require judgment on the part of scorer.
- It can be calculated by taking sample of papers scored by 2 examiners. The resultant correlation of both examiners rating will be a measure of scorer reliability.
- High correlation indicates that both examiners are scoring test papers similarly.

Standard Error of measurement

- The **standard error of measurement** of a test provides an estimate of the amount of **error** in a test.
- An individual's *true score* would equal the average of his or her scores *(observed scores)* on every possible version of a particular test in order to account for measurement error associated with a test design. Because the latter is impossible, standardized tests usually have an associated standard error of measurement (SEM), an index of the expected variation in observed scores due to measurement error. The SEM is in standard deviation units and can be related to the normal curve.
- Standard Error of measurement
- If a student takes the same test repeatedly, with no change in his level of knowledge and preparation, it is possible that some of the resulting scores would be slightly higher or slightly lower than the score that precisely reflects the student's actual level of knowledge

and ability. The difference between a student's actual score and his highest or lowest hypothetical score is known as the standard error of measurement.

- In the example below, a student who correctly answered 30 of the 60 questions on a grade-8 science test had a scale score of 403. The standard error of measurement at this achievement level was 15 scale score points.
- So his score range in between 385-418
- False Negative
- Because of the standard error of measurement, the potential exists that a small percentage of students may score lower than anticipated on a test, given their level of knowledge and preparation. Testing experts refer to this phenomenon as a "false negative."
- False Positive
- Conversely, the possibility exists that a small percentage of students may score higher than otherwise would have been expected. Testing experts refer to this phenomenon as a "false positive."