

Analysis of Test Results

Reliability, Validity, and Item Analysis





Learning Content

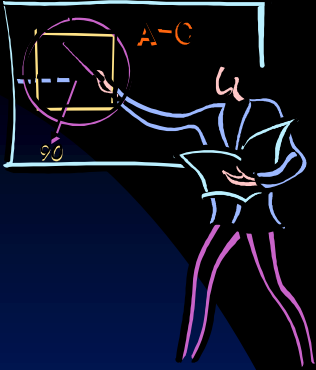
- Levels of Measurement
- Correlation Coefficient
- Reliability
- Validity
- Item Analysis

Objectives

- 1. Determine the use of the different ways of establishing an assessment tools' validity and reliability.
- 2. Familiarize on the different methods of establishing an assessment tools' validity and reliability.
- 3. Assess how good an assessment tool is by determining the index of validity, reliability, item discrimination, and item difficulty.

Levels of Measurement

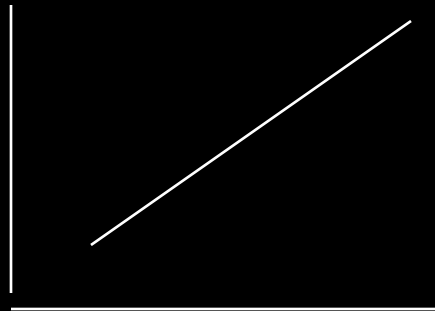
- Nominal
- Ordinal
- Interval
- Ratio



Correlation Coefficient

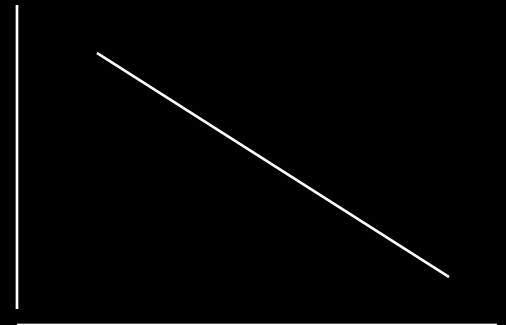
- Relationship of two variables (X & Y)
- Direction
- Positive
- Negative

X



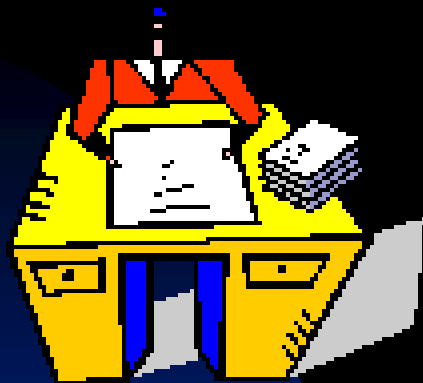
Y

Negative



Degree of Relationship

- 0.80 – 1.00 Very High relationship
- 0.6 – 0.79 High Relationship
- 0.40 – 0.59 Substantial/Marked relationship
- 0.20 – 0.39 Low relationship
- 0.00 – 0.19 Negligible relationship



Testing for Significance

- Nominal: Phi Coefficient
- Ordinal: Spearman rho
- Interval & Ratio: Pearson r
- Interval with nominal: Point biserial
- Decision rule:
 - If $p \text{ value} < \alpha = .05$: significant relationship
 - If $p \text{ value} > \alpha = .05$: no significant relationship

Variance



- R^2
- Square the correlation coefficient
- Interpretation: percentage of time that the variability in X accounts for the variability in Y.



Reliability

- Consistency of scores Obtained by the same person when retested with the identical test or with an equivalent form of the test



Test-Retest Reliability

- Repeating the identical test on a second occasion
- Temporal stability
- When variables are stable ex: motor coordination, finger dexterity, aptitude, capacity to learn
- Correlate the scores from the first test and second test.· The higher the correlation the more reliable



Alternate Form/Parallel Form

- Same person is tested with one form on the first occasion and with another equivalent form on the second
- Equivalence;
- Temporal stability and consistency of response
- Used for personality and mental ability tests
- Correlate scores on the first form and scores on the second form



Split half

- Two scores are obtained for each person by dividing the test into equivalent halves
- Internal consistency;
- Homogeneity of items
- Used for personality and mental ability tests
- The test should have many items
- Correlate scores of the odd and even numbered items
- Convert the obtained correlation coefficient into a coefficient estimate using Spearman Brown

■



Kuder Richardson (KR #20/KR #21)

- When computing for binary (e.g., true/false) items
- Consistency of responses to all items
- Used if there is a correct answer (right or wrong)
- Use KR #20 or KR #21 formula



Coefficient Alpha

- The reliability that would result if all values for each item were standardized (z transformed)
- Consistency of responses to all items
- Homogeneity of items
- Used for personality tests with multiple scored-items
- Use the cronbach's alpha formula



Inter-item reliability

- Consistency of responses to all items
- Homogeneity of items
- Used for personality tests with multiple scored-items
- Each item is correlated with every item in the test

Scorer Reliability



- Having a sample of test papers independently scored by two examiners
- To decrease examiner or scorer variance
- Clinical instruments employed in intensive individual tests ex. projective tests
- The two scores from the two raters obtained are correlated with each other



Validity

- Degree to which the test actually measures what it purports to measure



Content Validity

- Systematic examination of the test content to determine whether it covers a representative sample of the behavior domain to be measured.
- More appropriate for achievement tests & teacher made tests
- Items are based on instructional objectives, course syllabi & textbooks
- Consultation with experts
- Making test-specifications

Criterion-Prediction Validity



- Prediction from the test to any criterion situation over time interval
- Hiring job applicants, selecting students for admission to college, assigning military personnel to occupational training programs
- Test scores are correlated with other criterion measures ex: mechanical aptitude and job performance as a machinist



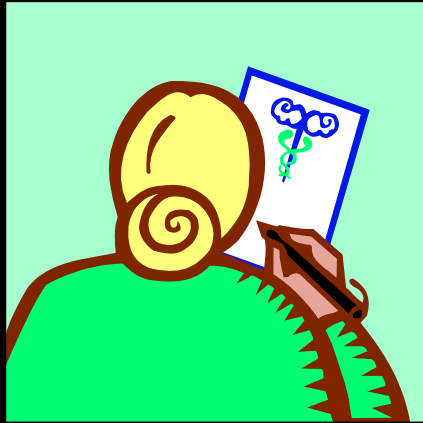
Concurrent validity

- Tests are administered to a group on whom criterion data are already available
- Diagnosing for existing status ex. entrance exam scores of students for college with their average grade for their senior year.
- Correlate the test score with the other existing measure



Construct Validity

- The extent to which the test may be said to measure a theoretical construct or trait.
- Used for personality tests. Measures that are multidimensional
 - Correlate a new test with a similar earlier test as measured approximately the same general behavior
- ● Factor analysis
- ● Comparison of the upper and lower group
- ● Point-biserial correlation (pass and fail with total test score)
- ● Correlate subtest with the entire test



Convergent Validity

- The test should correlate significantly from variables it is related to
- Commonly for personality measures
- Multitrait-multidimensional matrix



Divergent Validity

- The test should not correlate significantly from variables from which it should differ
- Commonly for personality measures
- Multitrait-multidimensional matrix

Reliability

| Type of Reliability | Nature | Measure of: | Use | Statistical Procedure |
|--|--|---|---|--|
| Test-retest | Repeating the identical test on a second occasion | Temporal stability | When variables are stable ex: motor coordination, finger dexterity, aptitude, capacity to learn | <ul style="list-style-type: none"> • Correlate the scores from the first test and second test. • The higher the correlation the more reliable |
| Alternate Form/ Parallel Form | Same person is tested with one form on the first occasion and with another equivalent form on the second | Equivalence; Temporal stability and consistency of response | Used for personality and mental ability tests | <ul style="list-style-type: none"> • Correlate scores on the first form and scores on the second form |
| Split-half | Two scores are obtained for each person by dividing the test into equivalent halves | Internal consistency; Homogeneity of items | Used for personality and mental ability tests The test should have many items | <ul style="list-style-type: none"> • Correlate scores of the odd and even numbered items • Convert the obtained correlation coefficient into a coefficient estimate using Spearman Brown |

| Type of Reliability | Nature | Measure of: | Use | Statistical Procedure |
|-------------------------------------|---|---|--|--|
| Kuder-Richardson Reliability | When computed for binary (e.g., true/false) items | Consistency of responses to all items | Used if there is a correct answer (right or wrong) | <ul style="list-style-type: none"> • Use KR #20 or KR #21 formula |
| Coefficient Alpha | the reliability that would result if all values for each item were standardized (z transformed) | Consistency of responses to all items Homogeneity of items | Used for personality tests with multiple scored-items | <ul style="list-style-type: none"> • Use the cronbach's alpha formula |
| Inter-item reliability | | Consistency of responses to all items Homogeneity of items | Used for personality tests with multiple scored-items | Each item is correlated with every item in the test |
| Scorer Reliability | Having a sample of test papers independently scored by two examiners | To decrease examiner or scorer variance | Clinical instruments employed in intensive individual tests ex. projective tests | The two scores from the two raters obtained are correlated with each other |

Validity

| Type of Validity | Nature | Use | (Statistical) Procedure |
|--------------------------------------|--|--|--|
| Content Validity | Systematic examination of the test content to determine whether it covers a representative sample of the behavior domain to be measured. | More appropriate for achievement tests & teacher made tests | <ul style="list-style-type: none"> • Items are based on instructional objectives, course syllabi & textbooks • Consultation with experts • Making test-specifications |
| Face Validity | Not a validity in a technical sense! Refers to what the test appears to superficially measure | | |
| Criterion-Prediction Validity | Prediction from the test to any criterion situation over time interval | Hiring job applicants, selecting students for admission to college, assigning military personnel to occupational training programs | Test scores are correlated with other criterion measures ex: mechanical aptitude and job performance as a machinist |
| Concurrent validity | Tests are administered to a group on whom criterion data are already available | Diagnosing for existing status ex. entrance exam scores of students for college with their average grade for their senior year. | Correlate the test score with the other existing measure |

| Type of Validity | Nature | Use | (Statistical) Procedure |
|----------------------------|--|--|--|
| Construct Validity | The extent to which the test may be said to measure a theoretical construct or trait. | Used for personality tests. Measures that are multidimensional | <ul style="list-style-type: none"> • Correlate a new test with a similar earlier test as measured approximately the same general behavior • Factor analysis • Comparison of the upper and lower group • Point-biserial correlation (pass and fail with total test score) • Correlate subtest with the entire test |
| Convergent Validity | The test should correlate significantly from variables it is related to | Commonly for personality measures | Multitrait-multidimensional matrix |
| Divergent Validity | The test should not correlate significantly from variables from which it should differ | Commonly for personality measures | Multitrait-multidimensional matrix |

Item Analysis



- Item Difficulty – The percentage of respondents who answered an item correctly
- Item Discrimination – Degree to which an item differentiates correctly among test takers in the behavior that the test is designed to measure.



Difficulty Index

| ■ Difficulty Index | Remark |
|--------------------|----------------|
| ■ .76 or higher | Easy Item |
| ■ .25 to .75 | Average Item |
| ■ .24 or lower | Difficult Item |



Index Discrimination

- .40 and above - Very good item
- .30 - .39 - Good item
- .20 - .29 - Reasonably Good item
- .10 - .19 - Marginal item
- Below .10 - Poor item