Reliability, Objectivity & Validity

HPS 410
Dr. Joe G. Schmalfeldt

Selecting a Criterion Score
- Criterion score – the measure used to indicate a person’s ability.
  - Can be based on the mean score of the best score.
- Mean Score – average of all trials.
  - Usually a more reliable estimate of a person’s true ability.
- Best Score – optimal score a person achieves on any one trial.
  - May be used when criterion score is to be used as an indicator of maximum possible performance.

Types of Reliability
- Objectivity
- Reliability
- Degrees of Reliability
- Standard Error of Measurement
- Spearman Brown Prophecy Formula
- Intraclass Reliability

Potential Methods to Select a Criterion Score
1. Mean of all trials.
2. Best score of all trials.
3. Mean of selected trials based on trials on which group scored best.
4. Mean of selected trials based on trials on which individual scored best (i.e., omit outliers).

Appropriate method to use depends on the situation.
Norm-referenced Test
- Designed to reflect individual differences.

In Norm-referenced Framework
- Reliability - ability to detect reliable differences between subjects.

Types of Reliability
- Stability
- Internal Consistency

Stability (Test-retest) Reliability
- Each subject is measured with same instrument on two or more different days.
- Scores are then correlated.
  - An intraclass correlation should be used.

Internal Consistency Reliability
- Consistent rate of scoring throughout a test or from trial to trial.
- All trials are administered in a single day.

Sources of Measurement Error
- Lack of agreement among raters (i.e., objectivity).
- Lack of consistent performance by person.
- Failure of instrument to measure consistently.
- Failure of tester to follow standardized procedures.
Reliability depends on:
- Decreasing measurement error
- Detecting individual differences among people
  - ability to discriminate among different ability levels

Reliability
- Ranges from 0 to 1.00
  - When $R = 0$, there is no reliability.
  - When $R = 1$, there is maximum reliability.

What is acceptable reliability?
- Depends on:
  - age
  - gender
  - experience of people tested
  - size of reliability coefficients others have obtained
  - number of days or trials
  - stability vs. internal consistency coefficient

Factors Affecting Reliability
- Type of test.
  - Maximum effort test expect $R_{xx} \geq .80$
  - Accuracy type test expect $R_{xx} \geq .70$
  - Psychological inventories expect $R_{xx} \geq .70$
- Range of ability.
  - $R_{xx}$ higher for heterogeneous groups than for homogeneous groups.
- Test length.
  - Longer test, higher $R_{xx}$
- Scoring accuracy.
  - Person administering test must be competent.
- Test difficulty.
  - Test must discriminate among ability levels.
- Test environment, organization, and instructions.
  - Favorable to good performance, motivated to do well, ready to be tested, know what to expect.
Factors Affecting Reliability

- Fatigue
  - decreases R
- Practice trials
  - increase R

Equivalence Reliability (Parallel Forms)

- Two equivalent forms of a test are administered to same subjects.
- Scores on the two forms are then correlated.

Standard Error of Measurement ($SE_M$)

- Degree you expect test score to vary due to measurement error.
- Standard deviation of a test score.
  \[ SE = S \sqrt{1 - R} \]
  - $S =$ standard deviation of group
  - $R =$ reliability coefficient
- Small SE indicates high reliability

Objectivity (Rater Reliability)

- Degree of agreement between raters.
- Depends on:
  - clarity of scoring system.
  - degree to which judge can assign scores accurately.
- If test is highly objective, objectivity is obvious and rarely calculated.
- As subjectivity increases, test developer should report estimate of objectivity.

Two Types of Objectivity:

- Intrajudge objectivity
  - consistency in scoring when test user scores same test two or more times.
- Interjudge objectivity
  - consistency between two or more independent judgments of same performance.
- Calculate objectivity like reliability, but substitute judges scores for trials.

Criterion-referenced Test

- A test used to classify a person as proficient or nonproficient (pass or fail).
In Criterion-referenced Framework:

- Reliability - defined as consistency of classification.

Reliability of Criterion-referenced Test Scores

- To estimate reliability, a double-classification or contingency table is formed.

Validity

- Understanding Validity
- Evidences of Validity
- Content Validity Evidence
- Evidence of Logical Validity
- Criterion-related Validity Evidence
- Factors that Affect Criterion-related Validity
- Construct Validity Evidence
- Criterion-referenced Test Validity

Validity

- Most important characteristic of a measurement.

Validity

- Extent to which the inferences made from specific measures are appropriate.
- It is not the measures that are validated; it is the inferences that are validated.
Validation

- Process of obtaining evidence to demonstrate that inferences made from test scores are appropriate and meaningful.

Evidence of validity:

- judged based on the intended use of the data.
- should come from various sources.
- slowly evolves and allows us to place confidence in the measurement instrument.

Construct

- The object of interest.
- Examples: aerobic fitness, upper body strength, physical activity, body composition, perceived competence.
- All validation is construct validation.

Evidences of Validity

- Content validity evidence
- Logical validity evidence
- Criterion-related validity evidence
- Construct validity evidence

Content Validity Evidence

- Logical approach
- The term content validity is used with written knowledge tests.
  - Judgments of experts are used to determine whether the items on the test represent all of the important content areas.
- A property of the test rather than a property of the use of the test scores.

Evidence of Logical Validity

- Concept of logical validity was developed for physical tests in the field of exercise science.
- Extent to which a test measures the most important components of skill necessary to perform a motor task adequately.
Criterion-related Validity Evidence

- Demonstrated by correlating the test of the construct with a criterion measure of that construct.
- Two types of criterion-related evidence designs:
  - concurrent evidence
  - predictive evidence

Concurrent Evidence of Validity

- Used when a test is proposed as a substitute for the criterion measure.
- Crucial step: selection of criterion.
- Criteria typically come from:
  - Expert judges
  - Tournament standings
  - Known valid test

Expert Ratings

- Subjective ratings by one or more experts.
  - Either point rating scale or group ranking.
  - Point rating scale is preferable.
  - Commonly used with skill tests.
- Scores from the test are correlated with scores from the expert ratings.

Tournament Standings

- Assumes tournament standing is a good indicator of overall ability in the tested skill.
- Usually used with an individual or dual activity (e.g., tennis, badminton).
- Scores from the test are correlated with tournament standings.

Predetermined Criteria

- Predetermined criteria are scores from an accepted instrument.
- Scores from the test are correlated with scores from the predetermined criteria.
- Usually used to develop a quicker and easier test than criterion.
  - e.g., skinfolds are used to evaluate body composition. What is the criterion?

Predetermined Criteria

- Criterion measure should be recognized as the gold standard.
- Multiple correlations and regression equations are often used to establish evidence of concurrent validity.
Predictive Evidence of Validity

- Used when want to predict the criterion measure some time in the future.
- For Predictive Evidence the criterion is measured some time in the future.
- For Concurrent Evidence the criterion is measured at about the same point in time.

Factors that Affect Criterion-related Validity Coefficient

- Criterion measure selected
- Characteristics of individuals tested
- Reliability
- Objectivity
- Test length

Size of Criterion-related Validity Coefficient

- Depends on:
  - criterion used
  - what others conducting similar studies have found
  - accuracy needed

Construct Validity Evidence

- Based on scientific method.
  - Hypothesize about the construct
  - Develop theory to explain construct, its relationships with other constructs, and/or the tests proposed to measure it
  - Apply procedures to confirm or disconfirm the theory
- Judge extent to which theoretical and statistical information supports the constructs.

Factor Analysis

- Used to confirm the nature of the construct.
- Can determine whether items seem to measure the same construct.

Known Difference Method

- Compare scores of two different groups that are expected to differ.
  - e.g., advanced vs. intermediate swimmers
  - e.g., comparison of group before and after training or instruction.
New 3-Stage Paradigm for Validity

- Definitional Evidence Stage
- Confirmatory Evidence Stage
- Theory Testing Stage

Definitional Evidence Stage
- Investigation of evidence to describe the nature of the construct of interest.
- Content or logical approach to collecting validity evidence is used at this stage.
- This evidence differs from other approaches to validation in that properties of the test rather than interpretations of the test are examined.

Confirmatory Evidence Stage
- Collect evidence to confirm or disconfirm the description of the construct.

Theory Testing Stage
- Test theories of how the construct of interest fits into the broader context.
- Considerations of how the construct is related to other constructs.
- Determinants of a person’s status on the construct of interest.

Criterion-referenced Test (CRT) Validity
- Criterion-referenced test -- test with a predetermined standard used to classify people as proficient or nonproficient.
- Different techniques are used to validate CRT.
Two Approaches to CRT Validity

- Logical -- Domain-referenced validity
- Statistical -- Decision validity or decision accuracy

Domain-referenced Validity

- Domain -- behavior the test is designed to measure.
- Test is logically validated by showing the test measures the behavior of interest.
- Similar to logical validity approach with norm-referenced validity.

Decision Validity or Decision Accuracy

- Accuracy of a test in classifying subjects as proficient or nonproficient.

Decision Validity

- Must be able to classify subjects as either true masters or true nonmasters.
- This is often difficult to do.
- A contingency table is then constructed.

Interpretation of Validity Coefficient “C”

- If C = .50, classification was no better than chance.
- When C = .50, validity is low.
- Values of C > .80 are desirable, but there are no set rules for evaluating size of C.

For Criterion-referenced Test Validity:

- Should have evidence of both
  - domain-referenced validity and
  - decision validity.
Phi Coefficient

- Another estimate of CRT validity.
- Correlation between two dichotomous variables.
- Range from -1.0 to +1.0