KINE 601

Reliability & Validity

Reading: Huck pp 75 - 98

Reliability & Validity

- **Reliability: "consistency", test "repeatability"**

- **Factors influencing reliability scores for a given instrument**
  - the more heterogeneous the group being measured r u reliability
  - the larger the total variance in a dependent variable r u reliability
  - the more items (questions) on an instrument r u reliability

- **Example:**
  - suppose we wanted to determine the reliability of a machine used to measure back extension range
  - Subjects
    - homogenous group of healthy people: ranges between 20° and 25°
  - since the range (and variance) of measurements is small r d reliability
  - solution: include individuals with hypermobile & hypomobile spines
  - score range will u r u score variance r u reliability

Reliability & Validity

- **Notes on Reliability**
  - different reliability instruments (statistics) may give different answers
    - when reliability is critical (medical testing equipment, etc.) more than one approach or instrument should be used to assess reliability
  - instruments may give varying results depending on test subjects
    - Example: physics test given to physics students versus third graders
  - never assume an instrument is reliable on the basis of:
    - manufacturers guarantees
      - sometimes manufacturers fund their own validity and reliability studies and publish them. Some of these types of studies have been done by reputable scientists (grant incentives)
    - previously cited literature
  - reliability is better estimated with **variances** vs. **correlations**
    - ICC (variance ratio) is best but correlations are more popular in the literature
    - Chronbach's **a** – all possible "split halve" combinations
    - Kappa Coefficient – establishes rater reliability for categorical evaluations
  - two or more indices of reliability are better than one
    - Pearson **r** plus Student's t-test

Reliability & Validity

- **Validity: accuracy of measurement**
  - an instrument is valid if it measures what it is supposed to measure
  - note: a valid instrument is always reliable (accuracy requires consistency), but a reliable instrument may not always be valid
  - Which is more important: validity or reliability ????????

- **Specificity of Validity**
  - Just like reliability, validity must be evaluated within the context of it's intended purpose
    - Example: suppose we wish to measure body fat in a group of 12 year old boys using skinfolds. We enter the data and use a regression equation that was developed from 2300 subjects ages 18-65. Is our instrumentation valid?

- **True validity is difficult to establish**
  - to what do you compare acquired data to in order to establish validity
    - skinfold results are often validated by comparison with hydrostatic weighing
      - is the hydrostatic weighing valid????

Types of Measurement Validity

- **Calibration** - validation of a mechanical or electronic instrument by comparison with a known quantity or value
  - example: metabolic cart calibrated with gases of known composition

- **Face Validity** - instrument “appears” to be accurate
  - an instrument lacking face validity may be unacceptable at the onset
    - Example: the Bod-Pod's initial demonstration in the Applied Ex Sci Lab

- **Content Validity** - how well an evaluation instrument measures an intended content area.
  - like face validity, content validity is based on subjective judgements
    - does an exam measure information covered in the class & the book ?

Types of Measurement Validity

- **Criterion Related Validity** - how well performance on one instrument correlates with performance on another
  - test to be validated (target test) is correlated with criterion measure with the criterion measure being the "gold standard" with its validity already established
  - example: are GRE scores a valid estimator of GPA's in grad school ?

    - **Concurrent Validity** - target test and criterion test administered at the same time. It estimates validity of "what exists at that moment"

    - **Predictive Validity** - how well a target test will correlate with a criterion test which will be (or could be) administered in the future
      - Examples:
        - 1. how well do results from a 12-lead ECG graded exercise test predict results from an angiogram.
        - 2. how well do tumor markers predict the presence, absence, or progression of cancer.

Types of Measurement Validity

- **Construct Validity** - how well an instrument measures a hypothetical construct such as IQ, anxiety, or attitudes.
  - other important examples: quality of life, functionality, physical fitness
    - would a P.T. and O.T. define "functionality" the same way ?

- **Ways of measuring Construct Validity:**
  - **Known groups method:**
    - determine if a test can discriminate between individuals already known to have a particular trait or characteristic ? (discriminant function analysis)
  - **Factor Analysis:**
    - using a multivariate statistical technique to verify the existence of "dimensions" of a construct.
      - example: intelligence (the construct) is composed of numerous dimensions (verbal ability, quantification, reasoning….etc.) A valid test of intelligence should measure all of these dimensions. Factor analysis takes various test items and creates "factors" (scores representing a groupings of test items). If these factors are representative of these dimensions, the test is valid.
    - most often used as a data reduction technique to identify "dimensions".

Reliability & Validity

- **Final Notes on Reliability & Validity**
  - Study validity is a product of both the instrument used to collect the data and the subjects from whom the data was collected.
    - data collected using previously unvalidated instruments should be suspect
    - be wary of validity claims based on an insufficient number of subjects
    - data collected from "uncooperative" subjects negates study validity
      - description of subjects and procedures should address all possible issues
  - How "High" do the reliability and validity coefficients have to be
    - ICC's of .75 or greater indicate "good" reliability
    - ICC's of .90 or greater should be required for clinical measurements

Validity of Medical Screening Tools

- True Positive Test (TP) - test is positive and condition is present
- False Positive Test (FP) - test is positive and condition is absent
- True Negative Test (TN) - test is negative and condition is absent
- False Negative Test (FN) - test is negative and condition is present

- **Sensitivity:** % of people with the condition that test positive

$$\frac{TP}{TP + FN}$$

- **Specificity:** % of people without the condition that test negative

$$\frac{TN}{TN + FP}$$

- **Predictive Value:** % of people with a positive test that have the condition

$$\frac{TP}{TP + FP}$$

Notes on Sensitivity & Specificity and Screening Tools

- It would be desirable to have tests that were both **sensitive** and **specific**
  - usually, there is a "trade-off" between sensitivity and specificity
    - trade-off based on what constitutes a positive vs. a negative test
    - criterion for a **negative test** made **more stringent**
      - **(norm ranges made smaller)** r fewer cases missed
      - (u sensitivity and d specificity……u chance of False **+** tests)
    - criterion for **negative test** made **less stringent**
      - **(norm ranges made larger)** r more cases missed
      - (d sensitivity and u specificity…… u chance of False **-** tests)
    - sensitivity is more important when the consequences of missing a diagnosis is high
    - specificity is more important when **cost** or **risk** of further intervention is very high
      - also important from a psychological standpoint (HIV results example)
    - examples: graded exercise testing and ST-segment changes, PSA values

The Validity of Research Studies

- **Internal Validity** - the "soundness" or “quality” of the research design
  - did manipulation of the independent variable truly cause the changes seen in the dependent variable or were confounding influences present to such a degree as to undermine study results ?
  - the better the research design, the higher the internal validity.

- **External Validity** - the extent or degree of **"generalizability“**
  - Inference space
  - are the results of the study applicable to a population

  - Note: a study cannot have external validity without internal validity

Threats to Research Study Validity

- **History** - occurrence of extraneous events which might affect study results
- **Maturation** - passage of time producing changes in subjects
- **Testing** - taking a pre-test may influence scores on a post-test
  - results may only be applicable to those taking a pre-test
- **Subject Mortality** - subjects drop out of study r d statistical power
- **Instrumentation Validity and Reliability**
- **Subject Selection Bias** - experimental effect is seen because subjects were pre-selected with a contributory trait
- **Hawthorne Effect** - subject awareness of hypothesis may influence outcome

Threats to Research Study Validity

- **Selection Maturation Interaction** - subjects selected for a specific trait and that trait may disappear over the course of the study

- **Self Fulfilling Prophecy** - researcher bias in observation / data collection

- **John Henry Effect** - competitive control group tries to out-perform experimental group during post-testing

- **Placebo Effect** - experimental responses occur in the placebo group because subjects believe they are receiving the experimental treatment

- **Halo effect** - subjects respond to meet researchers expectations

- **History - Treatment Interaction** - generalization of results may be limited to a point in time when data collection occurred
  - Example: Surveying people about opinions on heart disease risk just after a national media blitz on risk reduction by the American Heart Association