The Buck Stops Here

- Scientific Rigor
  - Function of Methods Used
    - To define the problems and develop an evidence based aims and hypothesis
    - To measure the variables with attention to potential sources of measurement error and bias.
    - To use and interpret statistical and other analysis precisely.
    - To limit generalizations.
**Objectives**

- Describe measurement error and its impact on the research process.
- Identify how item analysis can assist in maintaining rigor of measurement instruments.
- Describe reliability & validity strategies.
- Explain the concept of responsiveness and its associated techniques.
- Clarify sensitivity and specificity.

**The Reality**

- No amount of enthusiasm makes up for lack of scientific rigor
- Written research proposals are chains of reasoning
- Assessing validity & reliability involves critical thinking in regards to the preponderance of the evidence.

**Item Analysis**

- Not usually reported in the literature unless the study is one that seeking to establishing the psychometric properties of an instrument.
- An instrument that has to many items will artificially increase the scores used in reliability testing as well as respondent fatigue.
- Goal
  - To determine that each item is measuring the concept that it intents to measure.
  - To delete items that are redundant or are measuring another concepts.
- Technique: Item to Total Correlation
  - Items that score a correlation score >0.70 are redundant of other items on the scale
  - Items that score a correlation score <0.30 are measuring a different concept from the main body of items.
Measurement Error

- An measurement’s results varies as a result of accuracy in the measurement (true scores) and other factors (commonly called error)
- The degree of deviation between true scores and obtained scores when measuring a characteristic
- Observed Data = True Score + Error ($X_O = X_T + X_E$)
- Too much error can lead to misleading conclusions about study findings

Errors of Measurement

- Error score does not necessarily mean “wrong” score
- Error component is a composite of other factors that are also being measured by the researcher
- Error score includes other factors
- Example: Pain score (how much of the score is from “anxiety” around pain)?

Types of Error

- Random Error
  - Inconsistent, random variation
  - Error without qualification
  - Cannot find a pattern in error
- Systematic Error
  - Consistent error, not random
  - Bias
  - Has a consistent pattern
Sources of Measurement Error

- Situational Contaminants
  - Was there disruptions during the measurement?
- Response-Set Bias
  - Where a respondent always chooses the same answer or one that they believe the investigators want
- Transitory Personal Factors
  - Such as a headache, or increased stress
- Administration Variations
  - Are score different at different times of the year?
- Instrument Clarity
  - Was the instrument written at the correct literacy level?
- Response Sampling
  - Was the convenience sample biased in some way such as all female?

Reliability: Data Collection

- Degree of consistency with repeated measurements

Categories of Reliability

- Stability
  - Used with same people (patients) on separate occasions (over time) and get same answers
- Internal Consistency
  - All subparts/items are measuring same general thing
- Equivalence
  - Equitable results from two or more instruments or observers
Reliability: Stability

- Sometimes called test, re-test reliability
- Is the agreement of measuring instruments over time.
- To determine stability, a measure or test is repeated on the same subjects at a future date.
- Results are compared and correlated with the initial test to give a measure of stability.
- Typically the Spearman-Brown coefficient.
- Scores equal to or greater than 0.70 are usually considered sufficient.

Reliability: Internal Consistency

- This form of reliability is used to judge the consistency of results across items on the same test.
- Essentially, you are comparing test items that measure the same construct to determine the tests internal consistency.
- Statistical Techniques
  - Split-Half
    - Items are divided into 2 sections, then a correlation between the two sections is determined.
  - Cronbach's Alpha
    - The average of all possible split half reliabilities for a set of items
    - By convention, a lenient cut-off of .60 is common in exploratory research; alpha should be at least .70 or higher to retain an item in an "adequate" scale; and many researchers require a cut-off of .80 for a "good scale."

Aspects of Reliability: Equivalence

- Equivalency reliability is the extent to which two observers or items measure identical concepts at an identical level of difficulty.
- Inter-Rater Reliability
  - Are different observers using the same instrument measuring the same phenomena equivalent?
  - A statistical measure of inter-rater reliability is Cohen's Kappa
    - Ranges from -1.0 to 1.0 where
    - Large numbers mean better reliability,
    - Values near zero suggest that agreement is attributable to chance, and
    - Values less than zero signify that agreement is even less than which could be attributed to chance.
- Instrumental Equivalency
  - Are two (presumably parallel) instruments administered at about the same time equivalent?
Validity: Data Collection

- Degree to which a data collection instrument measures what it is supposed to be measuring.
- 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.

Reliability & Validity

- Not totally independent of each other
- An instrument that is not reliable cannot possibly be valid
  - erratic, inconsistent, inaccurate
- However, an instrument can be reliable and not valid

Levels of Validity

- There are many types of validity:
  - Face validity
  - Content validity
  - Construct validity
  - Criterion-related validity
Validity: Face Validity

- Face validity is concerned with how a measure or procedure appears.
- Does it seem like a reasonable way to gain the information the researchers are attempting to obtain?
- Does it seem well designed?
- Does it seem as though it will work reliably?
- Unlike content validity, face validity does not depend on established theories for support.

Validity: Content Validity

- Content validity evidence involves the degree to which the content of the test matches a content domain associated with the construct.
- Ask experts, based on judgment
  - Adequacy of the “content” area
  - Do my questions adequately get to the area of interest?
    - Yes
    - No
    - Maybe
- Content Validity Index
  - Statistical measure of agreement among the experts.

Validity: Construct Validity

- Validity of a test or a measurement tool that is established by demonstrating its ability to identify or measure the variables or constructs that it proposes to identify or measure.
- The judgment is based on the accumulation of statistical findings, usually correlations, from numerous studies using the instrument being evaluated.
- Most common strategy: Factor Analysis
  - Types
    - Exploratory
    - Confirmatory
  - How many factors did the analysis reveal?
  - What variance was explained by all of the factors?
    - Usually want to explain more than 70% of the variance.
Validity: Criterion-related Validity

- A measure of how well one variable or set of variables predicts an outcome based on information from other variables.
- Criterion of comparison must be valid itself!
- Types
  - Concurrent (Known Groups)
    - A measurement/instrument is given to two divergent groups. If the measurement is valid, the scores should diverge.
  - Predictive
    - A measurement's ability to predict scores on another measurement that is related or purports to measure the same or similar construct.

Responsiveness

- Controversial as to whether or not this is separate from or just one type of validity.
- Looking for the effect size change of the instrument across time.
- For some investigators, it is also an assessment of an instrument:
  - Ceiling Effect
    - A ceiling effect occurs when test items aren't challenging enough for a group of individuals. Thus, the test score will not increase for a subsample of people who may have clinically improved because they have already reached the highest score that can be achieved on that test.
  - Floor Effect
    - The floor effect is when data cannot take on a value lower than some particular number. Thus, it represents a subsample for whom clinical decline may not register as a change in score, even if there is worsening of function/behavior etc.

Sensitivity & Specificity

- Assess the properties of a diagnostic instrument.
- Sensitivity and specificity describe how well the test discriminates between patients with and without disease.
  - Sensitivity is the proportion of patients with disease who test positive.
  - Specificity is the proportion of patients without disease who test negative.
Reliability and Validity in Qualitative Studies

Traditional Criteria for Judging Quantitative Research
- internal validity
- external validity
- reliability

Alternative Criteria for Judging Qualitative Research
- credibility
- transferability
- dependability
- conformability

References


References to Consider


For more information please contact: Nursing.Research@kp.org http://nursingpathways.kp.org/scal/research/index.html