Reliability and Validity

Measurement Error

- Chance fluctuations in measurement
- More precise measurement
  - Fluctuations noted more easily

Random vs. Systematic Error

- Random error
  - Chance fluctuations in measurement
  - Pushes measurements up and down
  - Average across many trials close to “exact” value
- Systematic error
  - Pushes measurements in same direction
  - Does not disappear across trials
Reliability

• Does a measure yield the same results each time you apply it?
• Reliable measures
  – Changes in measurement due to changes in what is being measured

Reliability: Thought Experiment

• Given task of developing a measure of restaurant food quality
  – How many raters sent to restaurant?
  – Which foods would raters sample?
  – How many occasions would raters visit restaurant?

Three Forms of Reliability

• “Ask many observers to make many observations on many occasions”
• Multiple observers
  – Interrater reliability
• Multiple observations
  – Internal consistency
• Multiple occasions
  – Temporal consistency/ test-retest reliability
Reliability

- Determining reliability depends on correlation
- Negative correlation
  - High scores on first test associated with low scores on second test
- Positive correlation
  - High scores on first test associated with high scores on second test

Interrater Reliability

- Degree to which multiple judges agree on observations
- Used in many real-world situations
- Judges
  - Trained
  - Independent

Interrater Reliability

- Trained judges
  - Taught coding scheme
  - Free to discuss rating with other judges
    - Only during training
- Independent judges
  - Measure/rate behavior individually
- Used when behavior cannot be assessed by simple physical actions
Internal Consistency Reliability

- Degree to which all the specific items/observations in a measure behave in the same way
- Construct of interest must be represented by more than one question or indicator

Test-retest of Reliability

- Estimates degree of fluctuation
  - Fluctuation in trait
  - Fluctuation in instrument

Increasing Reliability

- Reliability increases as numbers increase
- Results obtained are less likely to have been caused by error
- Raters
  - Systematic error / random error
- Observations
  - Systematic error / random error
- Occasions
  - Systematic error / random error
Reliability & Validity

• The ability of a test to measure what it is supposed to measure
• Related to reliability
  – The more reliable measures are usually also more valid as well.

Validity

• Internal & external validity
  – Mundane & experimental realism
• Face validity
• Content validity
• Criterion-related validity
• Construct validity

Internal Validity

• Extent to which research provides evidence of causality
• Laboratory experiments high in internal validity
  – Control for individual differences
  – Isolate independent variables from sources of contamination
Threats to Internal Validity

- Variables/factors causing changes in the DV
  - Influence ability to determine causality
  - Controlled by design and procedure

History

- Environmental (external) influences
- Occur during pre-test and post-test
  - Influence DV
- Risk increases as time between initial and follow-up assessment increases

Maturation

- Changes in the internal condition of participant
- Occurs as a function of time
- Examples
  - Aging, fatigue, learning, boredom
Instrumentation

• Changes in the measurement of the DV
  – Measurement device
    • Calibration
    • Wear
  – Experimenter
    • Change criteria
    • Loss of concentration
  – Participant
    • Practice
    • Change criteria

Selection

• Changes due to the participants included in the study
• Examples
  – Non-random assignment to condition
  – Failure of random assignment

Mortality

• Losing Ss as experiment progresses
  – Limits generalizeability
  – Can introduce bias
• Difficult to control
  – Provide evidence that lost Ss are similar to those remaining
Regression to the Mean

- Occurs when participants are selected on the basis of extreme scores
- Initial extreme scores will move to mean on subsequent tests
  - Fundamental principle of measurement
  - Random “noise” likely cause of some extreme scores

External Validity

- High external validity
  - Can generalize to other situations, people, or means of defining IV and DV
- Laboratory findings typically low in external validity
- Observational studies typically high in external validity
- Trade-off between internal and external validity

Mundane Realism

- Make experiments physically similar to real world
  - Increases external validity
- Disadvantages
  - Laborious
  - High cost
Experimental Realism

- Increases external validity
- Study is psychologically meaningful to participant
  - Participant truly experiences psychological states

Results of Asch’s Study

- Results
  - 37/50 conformed once
  - 14/40 conformed on over ½ the trials
- Study high in experimental realism

Other Types of Validity

- Face validity
- Content validity
- Criterion-related validity
- Construct validity
Face Validity

• Refers to the “face” of the test
• It doesn’t “look” like it measures what it is supposed to measure
  – Experts may inform you otherwise

Content Validity

• Is the test representative of the domain its supposed to cover?
• Evidence
  – Comes from experts in the field

Criterion-related Validity

• Tests are used to make predictions about behavior
  – SAT scores predicting academic potential
• Correlate test scores with an independent criterion
  – Concurrent validity of criterion
    • High school GPA and SAT
  – Predictive validity of criterion
    • College GPA and SAT
Construct Validity

- Direct reflection of the quality of a researcher's operational definitions
- If operational definitions capture abstract concepts
  - Construct validity is high
- Any type of research has capacity to be high in construct validity

Construct validity (cont.)

- Concerns two issues
  - Is construct being measured a valid construct?
  - Is the tool the best one for measuring that construct?
- Closely tied to the nature of theory
  - Can't be proven
  - Confidence accumulates gradually as research produces supportive results