26:010:557 / 26:620:557
Social Science Research Methods

Dr. Peter R. Gillett

Associate Professor

Department of Accounting & Information Systems

Rutgers Business School – Newark & New Brunswick

Overview

- Foundations of Measurement
- Reliability
- Validity
- Measurement Theory
- Research Design

Foundations of Measurement

- Four general levels of measurement
  - Nominal
    - Labels without ‘number’ meanings
    - All members of a subset are assigned the same numeral
    - No two subsets are assigned the same numeral
  - Ordinal
    - Objects in a set are rank ordered
    - This ordering is transitive
    - Equal spacing of numerals does not imply to equal spacing of the underlying properties

Foundations of Measurement

- Four general levels of measurement
  - Interval
    - At least ordinal
    - Equal distances on the scale represent equal distances in the property being measured
    - E.g., Celsius scale
  - Ratio
    - Highest level
    - At least interval
    - Has a natural absolute zero
    - Ratios of values are meaningful

Foundations of Measurement

- Many numerical measurements are, strictly speaking, at best ordinal
- However, they may be approximately 'interval' scales
  - We can often assume this if we have multiple measures and they are all substantially and linearly related
- Ratio scales are rarely *required*

Foundations of Measurement

- Likert scales

  strongly disagree – disagree – neutral – agree – strongly agree

  -2 -1 0 1 2

- Units of analysis
  - Individual
  - Group
  - Company / firm

Foundations of Measurement

- Measurement Scales
  - Equality of units
  - Comparability of scales
  - Transformations of scores (we will return to this later)
    - Percentiles
    - Standard scores
    - Standardized scores
    - Normalized scores
  - Building composite scores out of standardized components

Foundations of Measurement

- Description
  - Classification (qualitative)
  - Measurement (quantitative)
- Variables
  - Qualitative
  - Quantitative
    - Ranked
    - Scalar
      - Discontinuous
      - Continuous
- Scales

Reliability

- Reliability
  - Dependability
  - Stability
  - Consistency
  - Reproducibility
  - Predictability
  - Lack of distortion

Reliability

- Will we get the same results if we measure the same objects with the same instrument
  - Stability, dependability, predictability
- Are the measures obtained from the instrument ‘true’ measures
  - Lack of distortion
- How much measurement error is there?

Reliability

- Reliability is the *lack of distortion* or *precision* of a measuring instrument
- Total obtained variance includes
  - Systematic variance
  - Error variance
- Reliability is the proportion of ‘true’ variance to the total obtained variance of the data yielded by the measurement instrument
- Reliability is the proportion of ‘error’ variance to the total variance yielded by the measurement instrument, subtracted from 1

Reliability

- Attenuation
  - Unreliable measures mask relationships
  - Correction for attenuation
    - Measures correlation supposing perfectly reliable measures: $r_{xy} / v(r_{xx} * r_{yy})$
      - E.g., 0.28 / v(0.3 * 0.4) = 0.81
      - But 0.28 / v(0.9 * 0.9) = 0.31
      - Can help indicate the potential benefit of improving reliability
  - The more items in a measure, the greater the reliability (*ceteris paribus*)

Reliability

- Theory of True and Error Scores
- $V_t = V_8 + V_\varepsilon$ or $V_o = V_T + V_\varepsilon$
- The reliability *coefficient* $r_{tt} = V_t / V_8 = V_o / V_T$ is the square of the correlation between 'true' scores and observed scores
- The correlation itself is sometimes called the *index* of reliability
- Since we don't know the 'true' scores, we can only estimate r_{tt}

Reliability

- One way to estimate it is to compute the correlation between scores for the same subject at different points in time: this is the *test-retest reliability*
- Reliability may also be estimated from the correlation between two alternative but *equivalent* (*parallel*) *forms* of the instrument

Reliability

- Another technique is to split scores on test item into odd and even items, and compute the correlation between the two subsets: the *split-half reliability*
- This is one example of measures based on *internal consistency*
- Since additional items on a test increase reliability, split-half correlations underestimate reliability
- There are a number of alternative ways to correct for this, including the Spearman-Brown Prophecy formula $r_{tt'} = \frac{nr_{tt}}{1+(n-1)r_{tt}}$ or, for split-halves, $r_{tt'} = \frac{2r_{tt}}{1+r_{tt}}$

Reliability

- Two other internal consistency measures of reliability that are well known and have been used extensively are the *Kuder-Richardson formulae* KR-20 and KR-21
- These formulae assume that every test item has the same mean and variance, and that the scoring is binary or dichotomous
- KR-21 is a special case of KR-20 where item difficulties are the same

Reliability

- There is also a more sophisticated theory of reliability called Generalizability Theory (formerly known as Domain Sampling Theory) produced by Cronbach and others
- Out of this theory comes probably the best known of current indicators of reliability, Cronbach's alpha $r_{tt} = a = \frac{k}{k-1}\left(1 - \frac{\sum V_i}{V_t}\right)$ or $r_{tt} = \frac{n\bar{r}_{it}}{1+(n-1)\bar{r}_{it}}$

Reliability

- The Kuder-Richardson formulae can be shown to be special cases of coefficient alpha
- It is equivalent to result of applying the Spearman-Brown formula to the mean of the inter-item correlations from every pair of items
- Note that Lee Cronbach has recently re-considered the value of Coefficient alpha
  - See: Cronbach, L.J. 2004. My current thoughts on Coefficient Alpha and successor procedures. *Educational and Psychological Measurement:* 64, 391-418.

Reliability

- In order to improve reliability
  - Write test items unambiguously
  - Add additional items of equal kind and quality
  - Give clear and standard instructions
- *Item analysis* is used to check that additional items are of equal kind and quality (a validity issue)
  - Item difficulty
  - Item discrimination

Reliability

- How much reliability do we need?
  - Dan Spencer (University of Kansas)
    - 0.6 – 0.8: acceptable
    - 0.8 – 0.9: very good

Reliability

- Designing test instruments
  - Item properties
    - Difficulty
    - Homogeneity
    - Validity
  - Weighting? (arguments against!)
  - Cross-validation against hold-out sample

Reliability

- Scoring
  - For robust measures, raw scores may be meaningful
  - But if departures from Normal too great, may need to normalize for analysis purposes
  - Kolmogorov-Smirnoff test (or others – Shapiro-Wilks, etc.) for normality
  - Percentiles are ordinal only

Reliability

- Combining scores from different samples
  - Measured under same conditions
    - Combine raw scores
  - Measured under different conditions
    - Standard scores (and normalize if desired)
  - When groups differ
    - Combine raw scores
    - Then use standard-score transformation

Reliability

- Standard scores

$$z = \frac{X - \overline{X}}{s_X}$$

- Standardized scores

$$T = \frac{10X}{s_X} + \left(50 - \frac{10\overline{X}}{s_X}\right)$$

- Normalized standard scores
  - Distribution converted to normal (e.g., via percentile scores)

Inter-rater Reliability

- The extent to which two or more individuals (coders or raters) agree
- The degree of stability exhibited when a measurement is repeated under identical conditions by different raters
- Often reported as correlation, or using Cohen's κ statistic
- For a calculator, see www.med-ed-online.org/rating/reliability.html
- For recent work on inter-rater reliability, see, for example, www.staxis.com

Validity

- "The subject of validity is complex, controversial, and peculiarly important" (K&L, 665)
- Are we measuring what we think we are measuring?
- APA, AERA, NCME validities
  - Content
  - Criterion-related
  - Construct

Validity

- Content Validity
  - Representativness or adequacy of the content of a measuring instrument
  - Essentially, this is a judgment call
    - Items included in measurement must be studied
    - Multiple competent judges should be used
    - Subject matter must be clearly defined
    - Some method for pooling judgments will be needed
- Face Validity
  - Immediate informal assessment of content validity

Validity

- Criterion-related Validity
  - Compare scores with one or more external variables (criteria) believed to measure the attribute under study
    - Predictive Validity
      - Future performance of the criterion
    - Concurrent Validity
      - Current performance of the criterion
  - Often used for new tests or instruments
  - Multiple predictors and criteria can be and are used

Validity

- Construct Validity
  - The extent to which constructs really capture what they purport to
  - Are some dimensions omitted?
  - Are irrelevant dimensions incorporated?
  - Convergent Validity
  - Discriminant Validity
  - Multitrait-Multimethod Matrix Method

Validity

- Convergent Validity
  - Instruments purporting to measure the same thing should be highly correlated
- Discriminant Validity
  - Instruments purporting to measure different things should not be highly correlated
- Multitrait-multimethod
  - See K&L, p. 675

Validity

- Other methods of construct validation
  - Correlation with multiple measures
    - Factor analysis
- Variance definition of Validity
  - Illustrates relationship with reliability
  - Introduces *communality* – proportion of common factor variance
- Although reliability and validity are related, we think of reliability as a largely technical matter, whereas validity is more philosophical – are we actually doing what we think we are?

Validity

- Note that Kerlinger (and Lee), in common with their empiricist bent, focus on construct validation from an empirical perspective
- Nevertheless, establishing that what we are measuring is similar to what other measures (of the same or a similar construct) measure, and different from the measure of distinct constructs does not itself prove that what we are measuring corresponds with the definition we have given of our construct in terms of others
- In other words, we need something more to establish the correspondence of our constitutive and our operational definitions of a construct
- Moreover, we may also need to protect others from confusion induced by prior connotations of the name we choose for our construct

Measurement Theory

- Thus, in assessing the measurement we use, we have considered a number of criteria:
  - Reliability
  - Face Validity
  - Content Validity
  - Criterion-related Validity
    - Concurrent
    - Predictive
  - Construct Validity

Research Design

- Research Design is the plan and structure of investigation
  - Framework
  - Organization
  - Configuration of elements
- Research Design has two purposes
  - To answer research questions
  - To control variance
    - Experimental
    - Extraneous
    - Error

Research Design

- Research Design tells us
  - What observations to make
  - How to make them
  - How to analyze their quantitative representations
- Recall that *power* = 1 – *Beta risk* = probability of correctly rejecting a false null hypothesis

Research Design

- Example of detecting admissions discrimination
  - Simple design randomly assigns males or females to colleges and compares admission rates
  - Factorial design crosses Gender with three levels of Ability (in this case both active variables) and tests *interaction*
  - Note that parallel tests at different levels of ability would not be as clear evidence

Research Design

- Maxmincon
  - Maximize systematic variance
  - Minimize error variance
  - Control extraneous variance

- *N.B. Here we are considering the variance of the dependent variable*

Research Design

- Experimental variance
  - Design plan and conduct research so that the experimental conditions are as different as possible
- Extraneous variance
  - Choose participants that are as homogeneous as possible on extraneous independent variables
  - Whenever possible, assign subjects to experimental groups and conditions randomly, and assign conditions and other factors to experimental groups randomly
  - Control extraneous variables by building them into the design
  - Match participants and assign them to experimental groups at random

Research Design

- Error variance
  - Reduce errors
  - Increase reliability of measures

Research Design

- Experiment
  - In an experiment, the researcher manipulates or controls one or more of the independent variables
- Nonexperiments
  - In nonexperimental research the nature of the variables precludes manipulation (e.g., sex, intelligence, occupation)
- “The ideal of science is the controlled experiment” (K&L, p. 467)