

26:010:685

***Social Science Methods
in Accounting Research***

Dr. Peter R. Gillett

Associate Professor

**Department of Accounting & Information Systems
Rutgers Business School – Newark & New Brunswick**

Overview

- Reliability
- Research Design
- Internal Validity
- External Validity
- Research Design Principles
- Validity
- Measurement Theory
- Experimental Validity

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
- * The more items in a measure, the greater the reliability (*ceteris paribus*)

Reliability

- Theory of True and Error Scores
- $V_o = V_T + V_\varepsilon$
- The reliability *coefficient* r_{tt} 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 formula

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 $r_{tt} = \alpha = \frac{k}{k-1} \left(1 - \frac{\sum V_i}{V_t} \right)$
- The Kuder-Richardson formulae can be shown to be special cases of coefficient alpha

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

- ◆ 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) 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

$$z = \frac{X - \bar{X}}{\sigma_X}$$

■ Standardized scores

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

■ Normalized standard scores

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

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 factor 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)

Research Design

■ Four “faulty” designs

* Notation:

- ◆ $X - X$ is manipulated
- ◆ $\sim X - X$ is not manipulated (i.e., subject not given X)
- ◆ $(X) - X$ is not manipulated but measured or imagined

Research Design

■ Design 19.1: One Group

* (a) X Y (Experimental)

* (b) (X) Y (Nonexperimental)

■ “One Shot Case Study”

■ *Scientifically* worthless

Research Design

■ Design 19.2: One Group, Before – After (Pretest – Posttest)

* (a) Y_b X Y_a (Experimental)

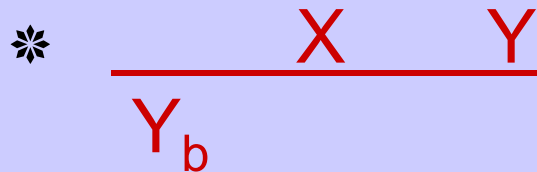
* (b) Y_b (X) Y_a (Nonexperimental)

■ Group is compared to itself

■ Measurement, history, maturation,
regression

Research Design

■ Design 19.3: Simulated Before – After



- Here we cannot tell whether the two groups were equivalent before X

Research Design

■ Design 19.4: Two Groups, No Control

* (a) $\frac{X \quad Y}{\sim X \quad \sim Y}$ (Experimental)

* (b) $\frac{(X) \quad Y}{(\sim X) \quad \sim Y}$ (Nonexperimental)

■ Groups are assumed equal on all other variables

Research Design

■ Criteria

- * Does the design adequately test the hypothesis
- * Does the design adequately control independent variables
 - ◆ Randomize whenever possible
 - ➔ Select participants at random
 - ➔ Assign participants to groups at random
 - ➔ Assign experimental treatments to groups at random
 - ◆ Control independent variables so that extraneous unwanted sources of systematic variance have minimal opportunity to operate
- * Generalizability

Research Design

■ Internal and External Validity

- * Campbell 1957, Campbell and Stanley 1963
- * The primary yardsticks by which the quality of research contributions is judged
- * These goals can and do conflict with each other

Internal Validity

- Did the experimental manipulation really make a significant difference?
- Threats / alternative explanations
 - * Measurement
 - * History
 - * Maturation
 - * Statistical Regression
 - * Instrumentation
 - * Selection
 - * Attrition / experimental mortality
 - * Interaction

Internal Validity

- In a longitudinal study, we take repeated measurements of subjects at different points in time
 - * What are the strengths and weaknesses of such studies as regards internal validity

External Validity

- To what populations can the conclusions from an experiment be generalized

- * Representativeness
- * Variable representativeness

- Threats

- * Reactive / interaction effects of testing
- * Interaction effects of selection biases
- * Reactive effects of experimental arrangements
- * Multiple-treatment interference

Research Design Principles

- Design is data discipline
- A design is formally a subset of the Cartesian product of the independent variable(s) and the dependent variable
- A complete design is based on a cross-partition of the independent variables
- We will not discuss incomplete designs
- Analysis of variance is a statistical technique appropriate for experimental designs that is not appropriate if participants cannot be assigned at random and there are unequal numbers of cases in the cells of the factorial design

Research Design Principles

- Design 20.1: Experimental Group – Control Group: Randomized Participants

* [R] $\frac{X}{\sim X} \quad Y$ (Experimental)
 (Control)

- “Best” for many purposes

Research Design Principles

■ Control group

- * Used to mean exclusively the group that did not receive a treatment
- * This is less obvious when there are multiple levels of treatment
- * More generally, now, it means the particular group against which comparisons are made

Research Design Principles

- ## ■ Design 20.2: Experimental Group – Control Group: Matched Participants

* [M_r] $\frac{X}{\sim X}$ Y (Experimental)
 Y (Control)

- Participants are matched on one or more attributes and randomly assigned to the two groups

Research Design Principles

■ Matching versus Randomization

- * Randomization is preferred
- * In practice, matching may be necessary
 - ◆ Matching by equating participants
 - ◆ Frequency distribution matching method
 - ➔ Can be tricky with multiple variables
 - ◆ Holding variables constant
 - ◆ Incorporating nuisance variables into the research design
 - ◆ Participants acting as own controls
- * Matching is only relevant when the variables are correlated with the dependent variable

Research Design Principles

■ Design 20.3: Before and After Control Group (Pretest – Posttest)

* (a) [R] $\frac{Y_b}{Y_b} \xrightarrow[\sim X]{X} \frac{Y_a}{Y_a}$ (Experimental)
(Control)

* (b) [M_r] $\frac{Y_b}{Y_b} \xrightarrow[\sim X]{X} \frac{Y_a}{Y_a}$ (Experimental)
(Control)

Research Design Principles

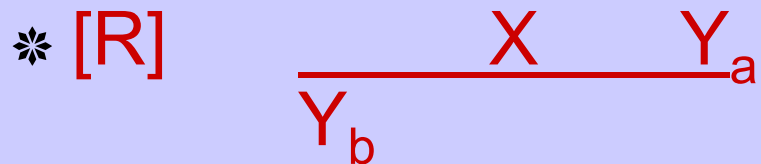
- Design 20.3 supplies a control group against which the difference $Y_a - Y_b$ can be checked
- However, difference scores are problematic unless the experimental effect is strong
- In addition, the pretest can have a sensitizing effect on participants, which decreases both internal and external validity
 - * Pretests should be avoided when the testing procedures are unusual

Research Design Principles

- Difference scores can be problematic - differences may be small compared to the error of measurement
- Residualized scores may be used instead
 - * Predict posttest scores from pretest scores based on correlations (regression)
 - * Subtract predicted posttest scores from actual posttest scores

Research Design Principles

- Design 20.4: Simulated Before – After, Randomized



- Improves Design 19.3 by adding randomization, but is still fairly weak

Research Design Principles

■ Design 20.5: Three Group, Before – After

*	$\underline{Y}_b$	X	$\underline{Y}_a$	(Experimental)
[R]	$\underline{Y}_b$	$\sim X$	$\underline{Y}_a$	(Control 1)
		X	$\underline{Y}_a$	(Control 2)

- Improves on Design 20.3 by adding a way of avoiding confounding effects of pretest

Research Design Principles

■ Design 20.6: Three Group, Before – After

* [R]	$\underline{Y_b}$	X	Y_a	(Experimental)
	$\underline{Y_b}$	$\sim X$	Y_a	(Control 1)
		X	Y_a	(Control 2)
		$\sim X$	Y_a	(Control 3)

- Strong, satisfying, potent controls, combines best designs so far (20.1 and 20.3), somewhat impractical, incomplete

Research Design Principles

- We can think of Design 20.6 as a factorial design, crossing the experimental manipulation with pretest – no pretest
- A factorial design is the structure of research in which two or more independent variables are juxtaposed in order to study their independent and interactive effects on a dependent variable
- For those who want to read more, Chapter 21 gives more examples, and also considers alternative designs such as correlated group designs, repeated trials designs, analysis of covariance

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

- * Representativeness 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

Measurement Theory

- Classic text:

“Measurement Theory for the Behavioral Sciences”

Ghiselli, Campbell & Zedeck

W. H. Freeman & Co, New York.

1981

Experimental Validity

- Internal Validity
- External Validity
- Construct Validity
- Statistical Conclusion Validity

Experimental Validity

- Internal validity: additional threats
 - * Ambiguity regarding causal influence
 - * Diffusion of treatments
 - * Compensatory equalization of treatments
 - * Compensatory rivalry by respondents
 - * Resentful demoralization
 - * Local history effects

Experimental Validity

■ Construct validity: threats

- * Inadequate preoperational explication
- * Mono-operation bias
- * Mono-method bias
- * Hypothesis-guessing
- * Evaluation apprehension
- * Experimenter expectancies
- * Confounding constructs
- * Interactions
- * Restricted generalizability across constructs

Experimental Validity

■ Statistical conclusion validity: threats

- * Low statistical power
- * Violation of assumptions of test
- * “Fishing” and error rates
- * Reliability of measures
- * Reliability of treatment implementation
- * Random irrelevancies
- * Random heterogeneity of respondents

Experimental Validity

- Classic text:

“Quasi-Experimentation”

Cook & Campbell

Houghton Mifflin 1979

- New Edition:

“Experimental and Quasi-Experimental Designs
for Generalized Casual Inference”

Shadish, Cook & Campbell

Houghton Mifflin 2002

Experimental Validity

- Generally trade-offs must be made between the different kinds of validity
- Precedence
 - * **For theory-building**
 - ◆ Internal, Construct, Statistical Conclusion, External
 - * **Applied research**
 - ◆ Internal, External, Construct (effect), Statistical Conclusion, Construct (cause)