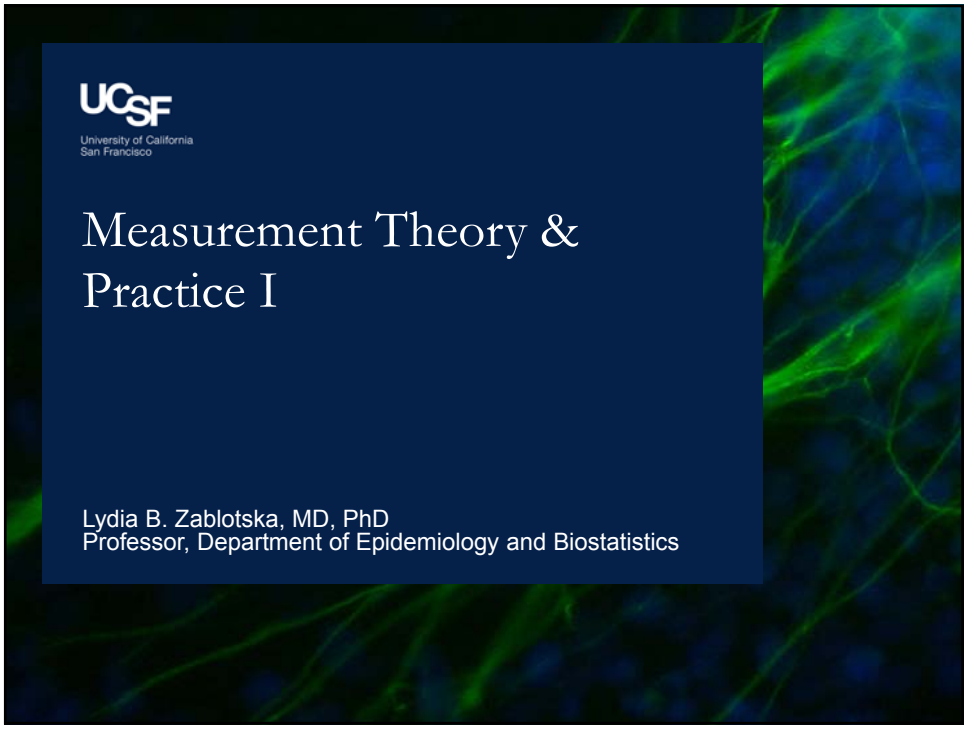


Highlights from Week #5

1. **Case-only design**
    1. Case-specular
    2. Case-crossover
2. **Assumptions of these methods**
3. **Advantages and limitations**

Highlights from Week #5

1. **Study designs used in pharmacoepidemiology**
   1. Case-crossover (CXO)
   2. Self-controlled case series (SCCS)
2. **Study designs used in genetic epidemiology**
   1. Family-based association studies
   2. Case-only Gene x Environment Interaction studies
3. **Assumptions**
   - **Monotonicity**
4. **Advantages and limitations**

3

UCSF

Highlights from Week #5

1. **Cross-sectional study designs**
   1. Standard, case / control
   2. Repeated survey, survey follow-up, intervention follow-up
   3. Proportionate mortality
2. **Ecological Studies**
   - Ecologic measures
   - Ecologic effect
   - Ecologic fallacy/ Atomistic fallacy
3. **Comparison of study designs**
4. **Levels of Evidence and their role in Evidence-Based Medicine**

4

UCSF

Learning Objectives

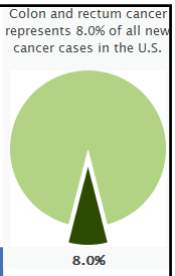
- Define measurement, phenomena, domains, concepts/conceptual models/conceptual frameworks
- Introduce classical measurement theory (CTT)
  - Give examples of modern measurement theories (Generalizability Theory, Item Response Theory)
  - ➢ **Additional slides and examples of modern measurement theories are at the end of the slide deck**
- Validity:
  - Discuss traditional definitions of validity of instruments (three C's: content, criterion and construct)
  - Describe the meaning of measurement validity and recommendations for its estimation using quantitative methods
  - Introduce consequential validity and discuss its importance for health research
- Reliability:
  - Review the relationship between reliability and validity
  - Define main types of reliability measurements and associated statistical tests:
    1. Internal consistency
    2. Parallel/ alternate forms
    3. Inter-rater agreement
    4. Test-retest

5

UCSF

Example: "The Fiber Yarn"

Study of the association between fiber intake and risk of colorectal cancer



Incidence rates of colorectal cancer per year in the U.S.			
	SEER 2000	SEER 2010	SEER 2013
Males	64 per 100,000	47 per 100,000	44 per 100,000
Females	47 per 100,000	36 per 100,000	34 per 100,000

- 4.4% of men and women born today will be diagnosed with cancer of the colon and rectum at some time during their lifetime
- 4th leading cause of cancer (after prostate, breast and lung)
- 2nd leading cause of death from cancer (after lung)

6 Howlader et al. 2016

UCSF

UCSF

… but wait, this seems to be all about **measurement!**

Knudsen MD et al. Am J Epidemiol.2014;179(10):1188-96. Self-reported whole-grain intake and plasma alkylresorcinol concentrations in combination in relation to the incidence of colorectal cancer.

RESULTS: We investigated this association using intakes of whole grains and whole-grain products measured via **FFQs** and **plasma alkylresorcinol concentrations, a biomarker of whole-grain wheat and rye intake**... Plasma alkylresorcinol concentration … (was) inversely associated with the incidence of distal colon cancer when the highest quartile was compared with the lowest (IRR= 0.34, 95% CI: 0.13, 0.92). No association was observed between whole-grain intake and any colorectal cancer when using an FFQ as the measure/ exposure variable for whole-grain intake.

CONCLUSIONS: … Assessing whole-grain intake using a combination of FFQs and biomarkers slightly increases the precision in estimating the risk of colon or rectal cancer...

7 UCSF

… Saga Continues …
a.k.a. "Untangling *the fiber yarn*"

Summary:

- Changing definitions of "fiber consumption" have an effect on the findings of studies looking for causes of colorectal cancer
- Questionable validity of some of the measures of "fiber consumption"
- Study methods, i.e. "study design" and data collection instruments, have a profound effect on the study findings

8 UCSF

Back to basics:

What is measurement?

"Measurement of some attribute of a set of things is the process of assigning numbers or other symbols to the things in such a way that relationships of the numbers or symbols reflect relationships of the attributes of the things being measured. A particular way of assigning numbers or symbols to measure something is called a scale of measurement."

Stevens, 1951

9

UCSF

Back to basics:

What is measurement theory?

"A branch of applied statistics that attempts to describe, categorize, and evaluate the quality of measurements, improve the usefulness, accuracy, and meaningfulness of measurements, and propose methods for developing new and better measurement instruments."

Allen, *Introduction to Measurement Theory*, 2001

Origins in psychological measurements, i.e. psychometrics, in the early 1930s.

10

UCSF

UCSF

Back to basics:

What are the building blocks of measurement?

Phenomenon ← domains ← concepts } conceptual framework

- "**Phenomena**" are observable facts or events.
- **Concepts** define the content of interest in measuring phenomena and systematically guide the measuring process. Concepts of interest in health research are usually difficult to operationalize, i.e. render measurable.
- **Conceptual frameworks** are models which are formed after a conceptualization or generalization process, they unify our understanding of how the concepts interact and help us organize our ideas. Sometimes, frameworks could be clearly subdivided into **domains** or spheres of knowledge, e.g., cognitive, affective and psychomotor domains of learning, each of which has its unique concepts and conceptual frameworks.

11 UCSF

Back to basics:

What are the steps of measurement process?

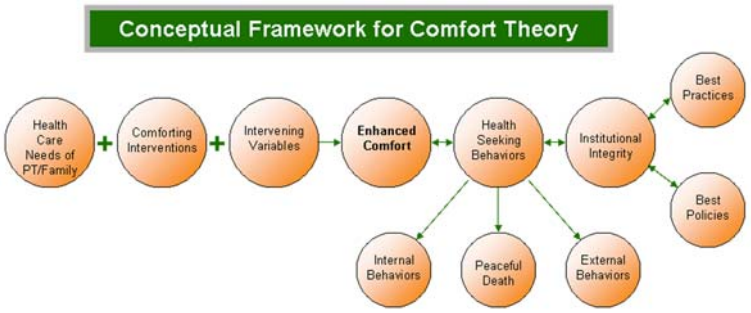
- A process of operationalizing concepts/ conceptual models/ conceptual frameworks through instrumentation

1. Start with **conceptualization** (a.k.a., 'theorization', 'generalization' process): Determine what constitutes phenomenon of interest? What are the main concepts?
2. "**Operationalization**" means translating an abstract concept into concrete observable events of phenomena**, i.e., creating operational definitions.**
3. "**Instrumentation**" means selecting or developing tools and methods appropriate for measuring an attribute or characteristic of interest, e.g., standardized or study-specific questionnaires.

12 UCSF

From *Phenomenon* to *Conceptual Framework* Example 1:

- We want to examine
  - Comfort
  - ➢ Conceptual framework developed by Kolcaba 2007



13 © Kolcaba (2007)

From *Phenomenon* to *Concept* to *Instrumentation* Examples 2 & 3:

- Phenomenon: Functional status covers both the individual carrying out activities of daily living and the individual participating in life situations and society.
  - Concept: The Functional Health Status
  - Instrumentation: Inventory of Functional Status in the Elderly Sickness Impact Profile
- Phenomenon: Decreased physical activity and discomfort resulting from inadequate sleep
  - Concept: Sleep Pattern Disturbance
  - Instrumentation: Wrist actigraph with scoring of "sleep efficiency, time in bed, total sleep time, awake time, after sleep onset, and number of nocturnal awakenings"

14 Ulrich and Canale *Nursing Care Planning Guides*, 7th ed. 2010

From *Research Question* to *Phenomenon* to *Domains* to *Concepts* to *Instrumentation* Example 4:

- Research Question: Effects of introduction of EMR on patient safety
- Phenomenon: Patient safety
- Domains: Organizational and Individual (Patient safety culture has been described as a product of individual and organizational factors and includes individual and group values, attitudes, perceptions, competencies, and patterns of behavior)
- Concepts: teamwork climate, organizational climate (including perceptions of management, working conditions, job satisfaction, and stress recognition), communication/communication openness, and organizational learning .
- Instrumentation: The AHRQ Hospital Survey on Patient Safety and the Safety Attitudes Questionnaire

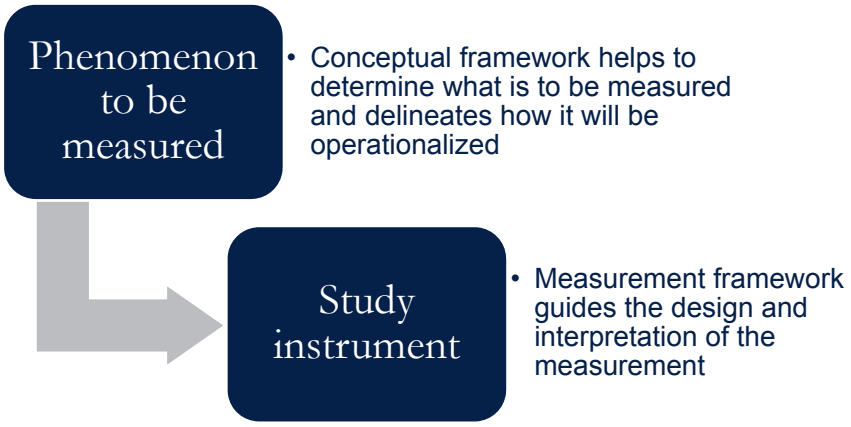
15 UCSF

From *Phenomenon* to *Concept* to *Operational Definition*

There are, however, six stages and several processes that are useful in engaging in the whole activity of theorization, whether theorization is used as a framework for research, for data interpretation, for concept development, for statistical model building, or for the development of a theory. The stages are: (1) sensing and taking in a phenomenon, (2) describing a phenomenon, (3) labeling, (4) concept development, (5) statement development, (6) explicating assumptions, and (7) sharing and communicating. Although these stages and processes are presented here linearly and sequentially, they could occur simultaneously, out of sequence, or in conjunction with other, yet undelineated, stages. It is useful for students of theory to deliberately and consciously experience each of these stages, even when such experiences are based on only a rehearsal of what they would use in the development of theory.

16 Meleis A.I. 2011 UCSF

Measurement theory



17

UCSF

Types of Measures

- **Direct** – values of concrete factors such as *weight, height, blood pressure, etc.*
- **Indirect** – measures or indicators of abstract ideas, characteristics or concepts such as *pain, stress, caring, coping, etc.*

18

UCSF

UCSF

Measurement Frameworks

- **Norm-referenced** – evaluate a subject relative to other subjects in a well-defined norm or comparison group
  - Eg., The Stress of Discharge Assessment Tool (SDAT-2) for patients with acute MI discharged from the hospital. Patient's score is compared to the scores obtained by other patients who responded to the same tool.
- **Criterion-referenced** – determine whether a subject has acquired a predetermined set of characteristics or behaviors
  - Eg., standards defined by the Guidelines for the Management of Patients with Chronic Stable Angina, by the AHA
  - Distribution has less variance and is typically skewed
- **Could be quantitative or qualitative**

Types of quantitative measurement frameworks

- **Objective** – construct the latitude of respondents and specify very clear criteria for scoring
  - Eg., multiple-choice, physiologic measures
- **Subjective** – allow respondents latitude in constructing responses and require individual judgment by scorers/ raters
  - Eg., description of feelings and emotions; there is a high probability that different scorers apply different criteria; the same rater could assign different scores to the same response on different occasions

Example: Types of quantitative measurement frameworks of organizational culture in health care settings

- Cognitive measures
- Affective measures (to determine interests, values, and attitudes);

HSR: Health Services Research 38:3 (June 2003)

The Quantitative Measurement of Organizational Culture in Health Care: A Review of the Available Instruments

Tim Scott, Russell Mannion, Huw Davies, and Martin Marshall

20

13 instruments that could be used to quantify culture in health care settings

Table 1: Organizational Culture Measurement Instruments that Have a Track Record in Health Care Settings

Name and Key References	*Cultural Dimensions and Outcome Measures*	*Number of Items*	*Nature of Scale*	*Examples of Use in Health Settings*	*Scientific Properties in Health Settings*	*Strengths*	*Limitations*	*Comments*
Competing Values Framework (Cameron and Freeman 1991; Gerowitz et al. 1996; Gerowitz 1998)	Key dimensions are staff climate, leadership style, bonding systems, prioritization of goals. Assessment results in four different culture types, described as: clan, adhocracy, hierarchy, and market types. Each organization usually has more than one of these types.	16	Brief scenarios describe dominant characteristics of each type. Respondents divide 100 points between these scenarios depending on how similar each scenario is to own organization.	Applied to top managers in 265 hospitals in U.K., U.S., and Canada (Gerowitz et al. 1996; Gerowitz 1998).	No details provided.	Simple and quick to complete, high face validity, used in several studies in health settings, strong theoretical basis, assesses both congruence and strength of culture.	Narrow classification of organizational types.	Originally developed for use in educational organizations.
Quality Improvement Implementation Survey (Shortell et al. 2000)	Key dimensions are: character of organization, managers' style, cohesion, prioritization of goals, and rewards. Assessment results in four different culture types, described	20	Brief scenarios describe dominant characteristics of each type. Respondents divide 100 points between these scenarios depending on how similar each scenario is	Used to assess relationship between culture and implementation of TQM in 16 hospitals (Shortell et al. 2000), and examination of relationship between implementation	Validity unknown. Internal consistency for one of the scales 0.79 (Shortell et al. 2000)	Simple and quick to complete. High face validity. Used in health settings by one of the leading teams in the field Adds extra dimension	Narrow classification of organizational types.	Based closely on the CVF but some terms modified to increase relevance to health organizations.

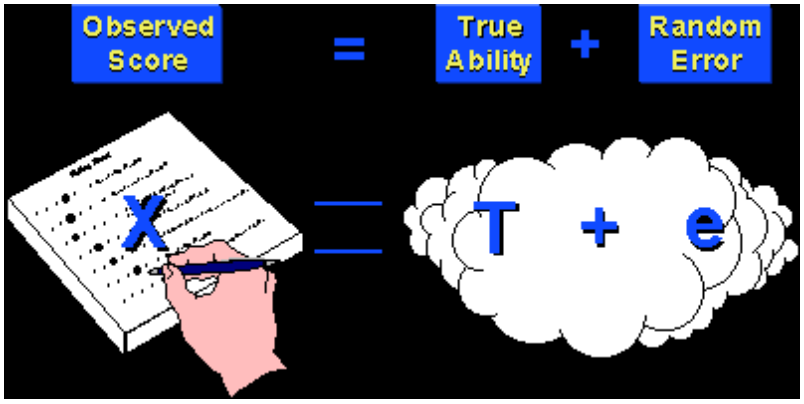
22 Scott et al. 2003

CLASSICAL MEASUREMENT THEORY

23

UCSF

Classical measurement theory



24

UCSF

UCSF

Classical measurement theory ('classical test theory' (CTT))

- Provides a model for assessing **random measurement error**
- Describes ways in which errors of measurement (due to falliable measures) can influence **observed** scores
- Describes **unsystematic**, or random, deviation of an individual's observed score from a theoretically expected or "true" score
- Effects of random error in the model are thought to cancel one another out if sample size is large

25 UCSF

Classical measurement theory ('classical test theory' (CTT))

- If the assumptions are reasonable, and the model is useful, then we can assume that the conclusions we draw from it are likely to be valid.
- Classical measurement theory is a relatively simple, quite useful method that describes ways in which errors of measurement (due to ***fallible*** measures) can influence ***observed*** scores.

26 UCSF

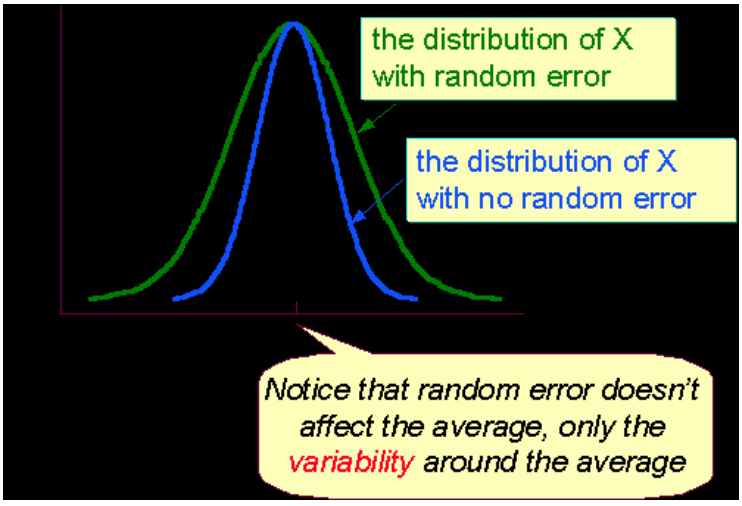
Examples of modern measurement theories

- Generalizability theory – acknowledges the multiple sources of measurement error by deriving estimates of each source separately
- Item response theory – in contrast to classical measurement theory, which focuses on test performance, this theory focuses on item's performance and the examinee's ability; eg., examinee's performance could be predicted based on a set of factors referred to as traits, latent traits, abilities or proficiencies

27

UCSF

Type of error?

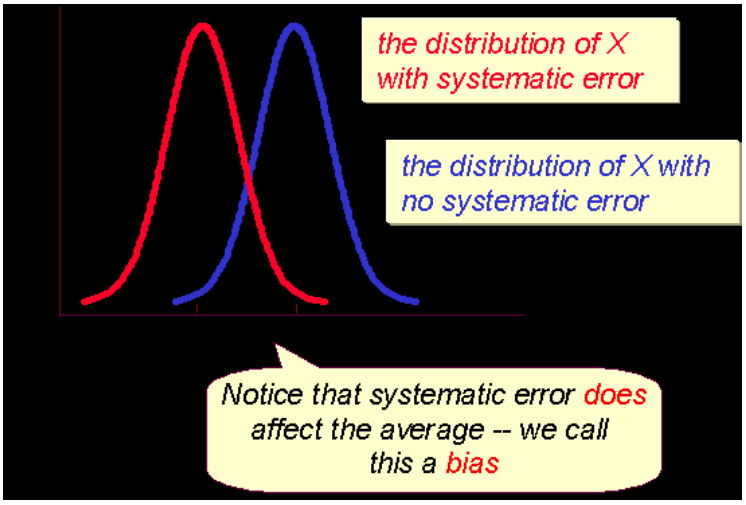


28

UCSF

UCSF

Type of error?



29

UCSF

Classical measurement theory: Assumptions 1.

$$X = T + E$$

X = Observed score
T = *True* score
E = Error of measurement

For a given individual, we assume that ***T*** is a fixed value, while ***X*** and ***E*** can vary from one measurement occasion to another.

30

UCSF

UCSF

Classical measurement theory: Assumptions 2.

$$\varepsilon(X) = T$$

The *expected value* (population mean) of X is T.

T is the mean of the theoretical distribution of X scores that would be found in repeated **independent** measurements of the same person with the same instrument.

31

UCSF

Classical measurement theory: Assumptions 3.

$$\rho_{ET} = 0$$

Error scores and true scores obtained by a population of individuals on one instrument are **uncorrelated**.

This assumption implies that people who score high on an instrument do not have systematically more positive or negative errors of measurement than people with low scores.

32

UCSF

UCSF

Classical measurement theory: Assumptions
4.

$$\rho_{E_1 E_2} = 0$$

E_1 = Error score for Measure 1

E_2 = Error score for Measure 2

$\therefore$ errors of measurement on two different measures are uncorrelated.

33

UCSF

Classical measurement theory: Assumptions
5.

$$\rho_{E_1 T_2} = 0$$

Errors of measurement on one measure are uncorrelated with true scores on *another* measure.

34

UCSF

UCSF

Summary

- Classical test theory (CTT) has been widely used in the development, characterization, and sometimes selection of outcome measures in health research. That is, qualities of outcomes, whether administered by clinicians or representing patient reports, are often describe in terms of “validity” and “reliability.”
- The fundamental feature of CTT is the formulation of every observed score (X) as a function of the individual’s true score (T) and random measurement error (e)
- CTT focuses on total test score – classical test theoretic constructs operate on the summary (sum of responses, average response) of items, individual items are not considered.
- CTT-based characterizations are ‘best’ when a single factor underlies the total score (break up the whole assessment into unidimensional bits, each of which has some reliability estimate)

35

UCSF

CTT’s limitations

- **X = T + E**
- Although E could represent many different types of error, CTT only permits the estimation of one general error term, which incorporates all sources of error but emphasizes certain types of error over others depending on the reliability study conducted:
  - Inter–rater reliability focusing on rater error
  - Test–retest reliability focusing on temporal error (occasion facet)
  - Internal consistency reliability focusing on how well items fit together to measure the phenomenon consistently
- Recognizes only two sources of variance
  - Test -retest (**stability over time**)
  - Parallel/ alternate forms (**equivalence in item sampling**)
- **Cannot adequately estimate individual sources of error influencing a measurement!**

36

UCSF

UCSF

Sources of Error

- **Random**
  - Fluctuations in the measurement based purely on chance
  - Scoring, rater drift, subject state, administration of measure, environment
  - ➢ **Primarily affects reliability (consistency or stability of measures)**
- **Systematic**
  - Measurement error that affect a score because of some particular characteristic of the person or the test that has nothing to due with the construct being measured
  - Characteristics of subject, measurement tool, measuring process, proxy response
  - ➢ **Primarily affects validity**

37 UCSF

QUANTITATIVE APPROACHES TO VALIDITY

38 UCSF

Classical validity theory

- Developed during World War II in the process of development of a wide array of measures, including personality screening tests for prospective fighter pilots, psychophysical measures to test reaction times, etc.
- Cronbach and Meehl (1955) categorized validity into:
  - Content
  - Criterion (including concurrent and predictive)
  - Construct

39 UCSF

Criticism of the classical validity theory

- Construct validity encompasses all other types of validity
- Fractional approach to validity is not helpful in practice
- Validity is not a property of the data collection instrument/ test, but of the data obtained using them when used for a specific study and with a particular population
- Many different types of evidence is necessary to evaluate validity of the instrument for data collection

40 UCSF

Instrument Validity

- A characteristic of an instrument that exemplifies the degree to which it measures what it purports to measure.

  Consequently…it is virtually impossible to state unequivocally that a measure is ***"valid"***

  Instead, *one can only build a case for a satisfactory degree of validity in the current study*

41

Instrument Validity

- One does not assess the validity of a measure per se... rather, one validates the use of the measure for a given situation.
- This distinction separates the characteristic of validity from the property of reliability.

42

Validation

- Cronbach (1971) described validation as the process by which a test developer or test user collects evidence to support the type of inferences that are to be drawn from test score
- “The correlation between a test score and some outside criterion.”

UCSF

Main steps in the evaluation of validity

1. **Requires empirical evidence** - A measurement method cannot be declared valid or invalid before it has ever been used and the resulting scores have been thoroughly analyzed.
2. **An ongoing process** - The conclusion that a measurement method is valid depends on the results of many studies done over a period of years. Every new study using that measurement method provides additional evidence for or against its validity.
3. **Is not an all-or-none property of a measurement method** - It is possible for a measurement method to be judged "somewhat valid" or for one measure to be considered "more valid" than another. A measurement method might be valid for one subject population or in one context, but not in another.

UCSF

UCSF

Classical validity theory: 3 C's
Methods of evaluation/assessment of validity

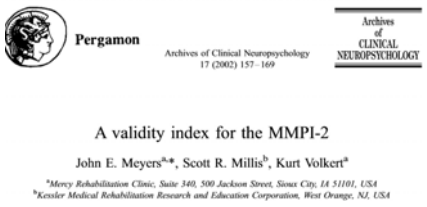
- **Content**
  - Face validity
  - Logical/Domain/Sampling-related Validity
- **Criterion-related**
  - Concurrent validity
  - Predictive validity
- **Construct**
  - Convergent validity
  - Discriminate validity

45

Content Validity – Face Validity

- The degree to which a measure *looks like* it's measuring the right thing
- In **face validity**, you look at the operationalization and see whether "on its face" it seems like a good translation of the construct.
- Often important, though not essential, face validity is usually **assessed by "expert" consensus**. Some very useful measures, however, have very little face validity (e.g., MMPI)

Example: There have been numerous methods developed for the detection of valid profiles on the **Minnesota Multiphasic Personality Inventory (MMPI)**. Chronic pain patients in litigation produce a different validity profile than do non-litigants.



46

Content validity –

- Refers to how well a test measures the risk factor/ trait/ behavior for which it is intended
- Established through the assessment of the actual content of a measure

47

UCSF

Content validity – Logical/Domain/Sampling-related Validity

- It addresses whether items on an instrument adequately measure a desired domain of content
- The degree to which a measure was designed to contain a representative sample of the domain of items/questions/areas/etc., from the universe of content.

48

UCSF

UCSF

Content validity – Steps for evaluation

- Content validity consists of a two-stage process (Lynn, 1986)

1) **Development**
   a) Domain identification
   b) Item generation
   c) Instrument construction
2) **Judgment-qualification**
   a) Ask a specific number of experts to evaluate the validity of items individually and as a set.
   b) Content experts should have relevant training, experience, and qualifications.
      – May include 2-20 content **experts**

49 UCSF

Content validity – Structural elements in content reviews

- **Item content**
  - Judge how representative individual items are of the content domain and whether the content domain adequately measures all dimensions of the construct.
- **Item style**
  - Judge the clarity of item construction and wording
- **Comprehensiveness**
  - Judge whether the complete set of instrument items is sufficient to represent to total content domain

50 UCSF

Caregiving Burden Items	Representativeness	Burden Dimension
Conceptual/Theoretical Definitions: Caregiving Burden-the demands of and reaction to the experience of caregiving within a family. Objective Caregiving Burden-the degree of disruption or change in the caregiver's life and household as a result of the caregiving experience. Subjective Caregiving Burden-the caregiver's attitude toward or psychological reaction to the caregiving experience.	1 = the item is not representative of caregiving burden 2 = the item needs major revisions to be representative of caregiving burden 3 = the item needs minor revisions to be representative of caregiving burden 4 = the item is representative of caregiving burden	1= subjective 2= objective 3= unable to classify
1. providing personal care (e.g., bathing, feeding) to my family member	1 2 3 4 Comments:	1 2 3 Comments:
2. communicating with my family member	1 2 3 4 Comments:	1 2 3 Comments:
3. managing finances, bills, etc., for my family member	1 2 3 4 Comments:	1 2 3 Comments:

Clarity: Are the caregiving burden items well written, distinct, and at an appropriate reading level for individuals who have primary home care responsibility for the psychosocial and physical care of a family member?

__Yes, the following items are clear (in the space below, indicate which items are clear):

__No, some of the items are unclear (in the space below, indicate which items are unclear):

Suggestions for making the items clear:

51

SF

Content validity – Quantitative approach

1. Content validity index (CVI):

- A CVI value can be computed for each item on a scale (I-CVI) as well as for the overall scale (S-CVI)
- To calculate an item-level CVI (I-CVI), experts are asked to rate the relevance of each item, usually on a 4-point scale
  - for each item, the I-CVI is computed as the number of experts giving a rating of either 3 or 4, divided by the number of experts—that is, the proportion in agreement about relevance.

52

UCSF

UCSF

CVI Example

Table 1. Fictitious Ratings on a 10-Item Scale by Three Experts: Items Rated 3 or 4 on 4-Point Relevance Scale

Item	Expert 1	Expert 2	Expert 3	Experts in Agreement	Item CVI
1	√	√	√	3	1.00
2	√	√	√	3	1.00
3	√	√	√	3	1.00
4	√	√	√	3	1.00
5	√	√	√	3	1.00
6	√	√	√	3	1.00
7	√	√	√	3	1.00
8	—	√	√	2	.67
9	—	√	√	2	.67
10	√	—	√	2	.67
				Average I-CVI =	.90
Proportion relevant	.80	.90	1.00		

I-CVI, item-level content validity index; scale-level content validity index, universal agreement method (S-CVI/UA) = .70; scale-level content validity index, averaging method (I-CVI/Ave) = .90; average proportion of items judged relevant across the three experts = .90.

53

UCSF

Content validity – Quantitative approach

2. Index of Item-Objective Congruence

1. Content experts rate items regarding how well they do (or do not) tap the established objectives
2. **The ratings are:**
   - 1: item clearly taps objective
   - 0: unsure/unclear
   - -1: item clearly does not tap objective
3. Several competing objectives are provided for each item
4. A statistical formula is then applied to the ratings of each item across raters
5. The result is an index ranging from '–1' to '+1'

54

UCSF

UCSF

Content validity – Quantitative approach

Crocker and Algina (1986, p. 221) provided a simplified version of the formula in their *Introduction to Classical and Modern Test Theory* text:

$$I_{ik} = \frac{N}{2N-2}(\mu_{ik} - \mu)$$

where I_{ik} is the index of item-objective congruence for item i on objective k, N = the number of objectives, μ_{ik} = the judges' mean rating of item i on objective k, and μ = the judges' mean rating of item i on all objectives.

2. Index of Item-Objective Congruence (IOC)

(Rovinelli & Hambleton 1977)

$$I_{ik} = \frac{(N-1)\sum_{j=1}^{n} X_{ijk} - \sum_{i=1}^{N}\sum_{j=1}^{n} X_{ijk} + \sum_{j=1}^{n} X_{ijk}}{2(N-1)n}$$

where I_{ik} is the index of item-objective congruence for item k on objective i,
N is the number of objectives (i = 1, 2, …, N),
n is the number of content specialists (j = 1, 2, …, n),
and
X_{ijk} is the rating (1, 0, -1) of item k as a measure of objective i by content specialist j (Berk, 1984, p.209).

- An index of '–1' can be interpreted as complete agreement by all experts that the item is measuring all the wrong objectives
- An index of '+1' can be interpreted as complete agreement by all experts that the item is only measuring the correct objective

55

UCSF

IOC Example

Content validity: Quantitative approach

INTERNATIONAL JOURNAL OF TESTING, 3(2), 163–171
Copyright © 2003, Lawrence Erlbaum Associates, Inc.

Indexes of Item-Objective Congruence for Multidimensional Items

Ronna C. Turner
Department of Educational Leadership, Counseling, and Foundations University of Arkansas

Laurie Carlson
Department of Counseling and Career Development Colorado State University

- Five items, five objectives, and ratings provided from four content experts
- The experts agree that Item #2 is clearly measuring Objective 1 and not measuring Objectives 3, 4 and 5
- Only one of four experts

Traditional Index of Item–Objective Congruence Values and the Associated Average Judges' Ratings for Each Objective

Item	Index of Item–Objective Congruence	Objectives 1	2	3	4	5
1	1.00	1.00*	−1.00	−1.00	−1.00	−1.00
2	0.97	1.00*	−0.75	−1.00	−1.00	−1.00
3	0.75	−0.75	0.75*	−0.75	−0.75	−0.75
4	0.69	−0.50	−1.00	0.50*	−1.00	−1.00
5	0.75	1.00*	1.00	−1.00	−1.00	−1.00

*Indicates the valid objective defined to be measured by the item.

believed t it is unclear whether the item is a measure of Objective 2

- There are no statistical tests for assessing significance of the measure
- One expert provided the rating of unsure for each of the 5 objectives of item #3
- Item #4 has a lower index value than the other items (either 2 of the 4 experts are unsure or 1 of the 4 experts believes the item is clearly not measuring Objective 3
- Items #1-3 would fall within the acceptable range of values using 4 content experts

56

UCSF

UCSF

Content validity - Steps for evaluation

Summary

1. Define the performance domain of interest
2. Select a panel of experts
3. Provide a structured framework for the process of matching items to the performance domain
4. Collect and summarize the data
5. Estimate numeric values for content validity measures

Types of validity

- **Content**
- **Criterion-related**
  - Concurrent validity
  - Predictive validity
- **Construct**
  - Convergent validity
  - Discriminate validity

Criterion-related validity –

- Refers to the extent that a test performance corresponds to an accurate measure of interest
- This type of validity is established by relating scores on the measure to some external criterion, and is usually expressed as a correlation coefficient of some sort, that is often referred to as the validity coefficient

59

UCSF

Criterion-related validity –
Steps for evaluation

1. Identify a criterion and a method for measuring it
2. Identify an appropriate sample of examinees
3. Administer the test
4. Obtain data on the criterion
5. Determine the strength of the relationship between test scores and criterion performance

60

UCSF

UCSF

Criterion-related validity – Concurrent validity

- A correlation between a measure and a criterion when both measurements are obtained at the same point in time.
- Is the test effective in predicting performance on a simultaneously administered criterion measure?
- This type of validity is established by relating scores on the measure to some external criterion, and is usually expressed as a correlation coefficient of some sort, that is often referred to as the validity coefficient.

61 UCSF

Criterion-related validity – Concurrent validity

- In concurrent validity, we assess the operationalization's ability to distinguish between groups that it should theoretically be able to distinguish between.
  - **Example:** if we come up with a way of assessing manic-depression, our measure should be able to distinguish between people who are diagnosed as manic-depression and those diagnosed paranoid schizophrenic.
- It is also demonstrated where a test correlates well with a measure that has previously been validated

62 UCSF

Criterion-related validity – Predictive validity

- A correlation between a measure and a criterion when there is a *reasonable* period of time between measurements.
- Is the test effective in forecasting future behavior?

63

UCSF

Criterion-related validity – Predictive validity

- In **predictive validity**, we assess the operationalization's *ability to* **predict something** it should theoretically be able to predict.
  - **Example:** we might theorize that a measure of math ability should be able to predict how well a person will do in an science-based profession.

64

UCSF

UCSF

Criterion-related validity – The Criterion Problem

- **The enigma**
  - If you have a good criterion, then why on earth are you developing a new measure!
- **The reality**
  - Most external criteria are inherently flawed… either they aren't completely appropriate, or they have bad statistical properties like restriction of range

65

UCSF

Criterion-related validity – Quantitative approach

NONE!

66

UCSF

UCSF

Types of validity

- **Content**
- **Criterion-related**
- **Construct**
  - Convergent validity - measures that are theoretically supposed to be highly interrelated are, in practice, highly interrelated
  - Discriminate validity - measures that shouldn't be related to each other in fact were not

67

Construct Validity –

- The degree to which an instrument measures the theoretical construct or trait/characteristic it was designed to measure
- In order to establish a measure's construct validity, the developer makes certain predictions that relate the measure to the theory that underlies it. These predictions can take a number of forms.

68

Construct validity – Quantitative approach

- Thorough and appropriate specification and sampling of the domain
- Multitrait-Multimethod Matrix (MTMM)
- Factor Analysis
  - Allows one to examine the congruence between the empirical data and the theoretical domain
    - ➢ **NEXT WEEK!**

69 UCSF

Construct validity – Quantitative approach

Multitrait-multimethod validation
(Cambell & Fiske, 1959)

- In a multi-method, multi-trait validation strategy, the researcher not only uses multiple indicators per concept, but also gathers data for each indicator by multiple methods and/or from multiple sources.

70 UCSF

Construct validity – Quantitative approach

Multitrait-multimethod validition

- **Convergent validity**
  - Demonstrated by ***high*** correlations between scores on instruments measuring the ***same*** trait by ***different*** methods (*Monotrait-Hetero method*)
- **Divergent (or discriminant) validity**
  - Demonstrated by ***low*** correlations between scores on instruments measuring ***different*** traits by the ***same*** method (*Heterotrait-Mono method*)
- **Use reliability indices to assess validity?!**

71

UCSF

Example:

A study of sixth grade students

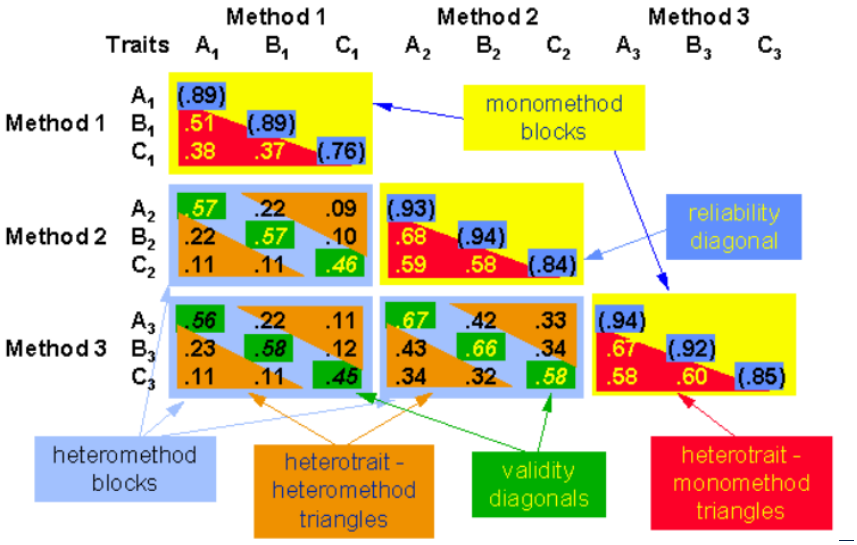
- **We want to measure three traits or concepts:**
  - Self Esteem (SE)
  - Self Disclosure (SD)
  - Locus of Control (LC)
- **Measure each of these three different ways:**
  - Student self assessment via Paper-and-Pencil (P&P)
  - Teacher rating
  - Parent rating
- **Measure reliability by:**
  - test-retest, internal consistency
- **Look at correlations between these reliability indices to assess validity**

72

UCSF

UCSF

Multitrait-multimethod validation

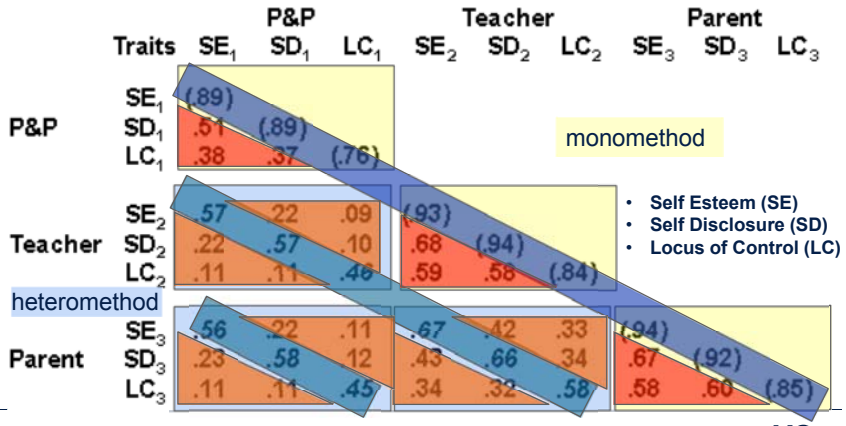


UCSF

Example:
A study of sixth grade students

Monotrait-Hetero method:

Construct validity demonstrated by ***high*** correlations between scores on instruments measuring the ***same*** trait by ***different*** method

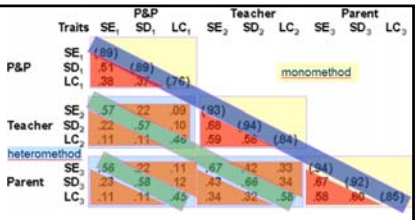


UCSF

UCSF

Example: A study of sixth grade students

Interpretation



- Coefficients in the **reliability diagonal** should consistently be the highest in the matrix (test-retest)
- Coefficients in the **validity diagonals** should be significantly different from zero and high enough to warrant further investigation (→ convergent validity)
- **A validity coefficient** should be higher than values lying in its column and row in the same **heteromethod block**
  - (SE P&P)-(SE Teacher)=0.57 should be greater than (SE P&P)-(SD Teacher)=0.22, (SE P&P)-(LC Teacher)=0.11, (SE Teacher)-(SD P&P)=0.22 and (SE Teacher)-(LC P&P)=0.09
- **A validity coefficient** should be higher than all coefficients in the **heterotrait-monomethod triangles** (trait factors should be stronger than methods factors)
  - However, (LC P&P)-(LC Teacher)=0.46 **is less** than (SE Teacher)-(SD Teacher)=0.68, (SE Teacher)-(LC Teacher)=0.59, and (SD Teacher)-(LC Teacher)=0.58
- The same *pattern* of trait interrelationship should be seen in all triangles
  - in all triangles the SE-SD relationship is approximately twice as large as the relationships that involve LC

75 **Conclusion: fairly strong construct validity!** UCSF

Construct Validity Example:

Journal of Clinical Nursing, 20, 2591–2599
CLINICAL ISSUES
Journal of Clinical Nursing

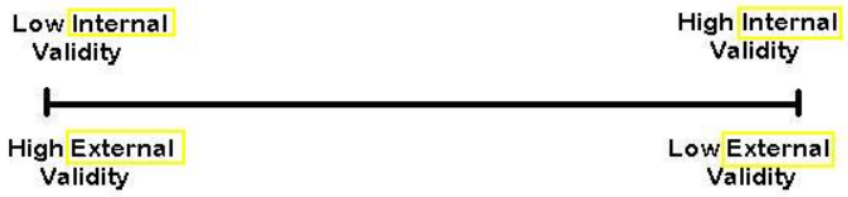
Psychometric evaluation of the Turkish version of the Pediatric Symptom Checklist-17 for detecting psychosocial problems in low-income children

Semra Erdogan and Meryem Ozturk

- N= 306 parents with children between 6–16 years in Turkey
- Parents completed PSC-17 and Child Behavior Checklist (CBCL)
- Content validity was ascertained by an expert panel of 8 academicians specializing in child and adolescence psychiatry
- The content validity index (CVI) score was computed by summing the percentage agreement scores of all items that were given by the experts a rating of '3' or '4'. The criterion for retaining an item was at least 80% agreement among the experts at the agree or strongly agree level of relevance to the construct (CVI=98.5%)
- The convergent validity was assessed by the Spearman correlation between PSC-17 and CBCL scores (r=0.82 for Total scores)
- The discriminant validity of the PSC-17 was assessed by examining whether the PSC-17 could differentiate among the normal, borderline and clinical range groups using ANOVA test (p<0.001 between the groups)
- The construct validity was assessed by factor analysis
- CBCL was used to estimate the optimal cut-off score of PSC-17 by using a Receiver operating characteristic analysis

76 UCSF

The Continuum of Validity



77

UCSF

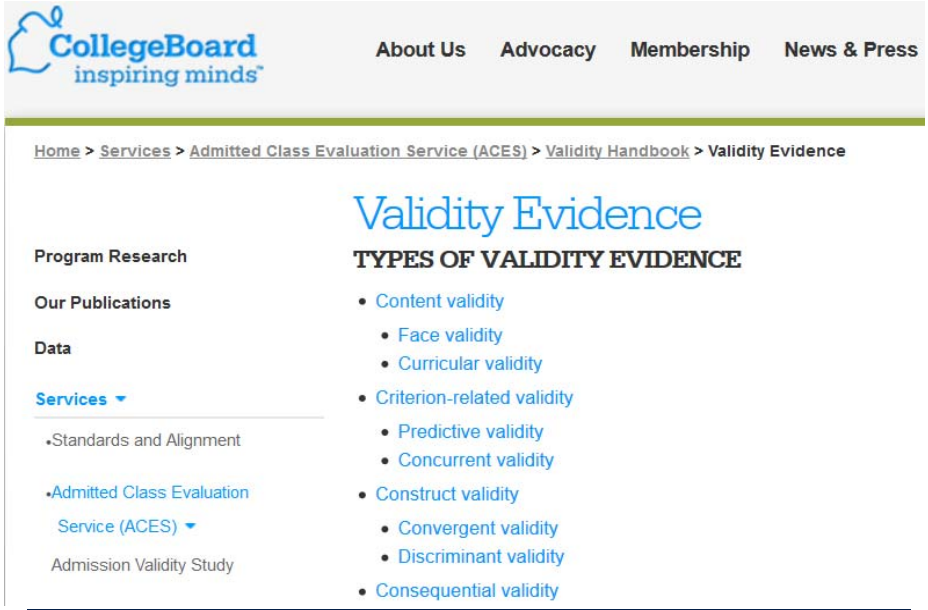
Types of validity evidence (Goodwin 2002)

- 24 validity standards to evaluate 5 types of validity evidence:
  - Evidence based on test content
  - Evidence based on response process
  - Evidence based on internal structure
  - Evidence based on relations to other variables
  - Evidence based on the consequences of testing

78

UCSF

UCSF



79

UCSF

What is Consequential Validity?

- Messick (1989) originally introduced consequences to the validity argument. Later, Shepard (1993, 1997) broadened the definition by arguing one must investigate both positive/negative and intended/ unintended consequences of score-based inferences to properly evaluate the validity of the assessment system.

80

UCSF

UCSF

Consequential Validity – Example:

APPLIED MEASUREMENT IN EDUCATION, 16(1), 1–26
Copyright © 2003, Lawrence Erlbaum Associates, Inc.

Consequences of a State Accountability Program: Examining Relationships Between School Performance Gains and Teacher, Student, and School Variables

Clement A. Stone and Suzanne Lane
Department of Psychology in Education
University of Pittsburgh

- Objectives: Explore the relationship between (a) *changes in the scores* from the Maryland State Performance Assessment Program *from 1993 to 1998* and (b) *classroom instruction and assessment practices*, student learning and motivation, students' and teachers' beliefs about and attitudes toward the assessment, and a school characteristic
- Intended Consequences - state assessments are intended to impact:
  - Student, teacher, and administrator motivation and effort;
  - Curriculum and instructional content and strategies;
  - Content and format of classroom assessments;
  - Professional development support;
  - Use and nature of test preparation activities; and
  - Student, teacher, administrator, and public awareness and beliefs about the assessment, criteria for judging performance, and the use of assessment results.

Consequential Validity – Example:

- **At times, however, *unintended consequences* are possible such as:**
  - Narrowing of curriculum and instruction to focus only on the specific learning outcomes assessed
  - Use of test preparation materials that are closely linked to the assessment without making changes to the curriculum and instruction
  - Use of unethical test preparation materials
  - Inappropriate use of test scores by administrators

Consequential validity – Impact on future research

- **Research questions on the consequential validity of alternate assessments from the perspective of:**
  - Students/Parents
    - What benefits to students have accrued from the participation in AA-AAS?
    - What is the relationship between student performance in AA-AAS and post-school life outcomes?
  - Teachers
    - What is the extent to which alternate assessments are a part of daily classroom routine?
  - School
    - Is there any relationship between student performance in the AA-AAS and student performance in the general assessment?

UCSF

Advantages of the new classification with consequential validity over the classical validity theory

- Clearly defined validation activities
- New emphasis on the consequences of testing
  - Social consequences of using a particular test for a particular purpose
  - How are the scores really being used?
  - Does this testing lead to benefits for participants?
  - Are there negative spin-offs?
  - Kreiter et al. 2013: Testing within medical education makes a strong positive contribution to learning
- Better understanding of the meaning of 'valid instrument' and 'validity of measure'

So, should we stop using the old classical validity theory?

➢ **Both continue to co-exist!**

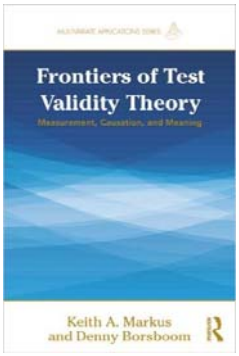
UCSF

UCSF

New Validity Theories

Table 12.2 Causal Interpretations in Measurement Models.

Type of causation	*Robustness to change of situation*	*Model interpretation*	*Required evidence*
No causation	Items will show similar covariances with each other and other measures in samples from the same population (as long as the world does not change).	Measurement models as merely descriptive of statistical associations.	Sampling of persons and indicators is adequate; backup for stability assumption.
Regularity causation	Causal models will remain robust to causally inert control variables. Psychometric models will therefore generalize to various subpopulations. Manipulation of constructs will impact items (reflective) or vice versa (formative) so long as intervention does not disrupt causal regularities.	Measurement models with generalization over inert variables based on statistical induction.	As in (a), plus support for transfer over inert variables (replication of results in new populations and situations) and support for a causal link through model fitting or experimentation.
Counterfactual causation	Causal models will remain robust to interventions that change the values of control variables outside the measurement model, so long as the values remain in keeping with the scope of the counterfactuals.	Measurement models with more readily generalizable causal links, but purely nomothetic.	As in (a) and (b), plus support for a causal link that is robust against changing control variables (e.g., measurement invariance over their levels).
Process causation	Causal models will remain robust to any intervention that does not alter the response process. This response process can be determined independently of the item responses.	Measurement models with elaborated causal mechanisms that explain nomothetic relationships and aid generalization.	As in (a), (b), and (c), plus evidence that the proposed process indeed establishes the relevant causal link.



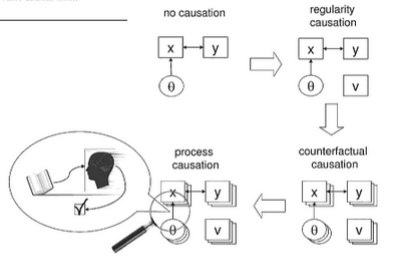


Figure 12.2 A Graphical Representation of Progressively Stronger Causal Claims in Validity Theory. Theta is the focal construct; X is the focal test score; Y is an external variable (e.g., a criterion); V is a causally inert control variable. A single box or circle represents a variable across actual cases, whereas stacked boxes or circles vary over possible cases as well.

85

Under regularity causation, such inferences across inert factors can be justified without saying anything specific about a causal link between the construct of interest and the indicators. Of course, the notion of causally inert factors contrasts with factors that are not causally inert (factors that cause differences in test scores), but one need not specify or name these to establish generalizability across inert factors.

Evidence for regularity causation typically arises from studies that establish generalizability of main results to novel populations (e.g., inhabitants of different countries), or to different testing contexts (e.g., computerized assessment). The concept of causally inert factors also helps bridge the gap from passive observation to intervention. Suppose that an intervention is hypothesized to affect the focal construct but not to affect the test scores directly. For example, consider an after-school enrichment program that addresses the full mathematics curriculum for a given grade and does not merely teach to the test. Further, assume that the construct is sufficiently delineated to be manipulated in (quasi-) experimental research. Under these circumstances, regularity causation implies that manipulation of the construct should be followed by changes in the test scores (reflective model) or the other way around (formative model). Establishing this effect is the goal of several standard validation research procedures (Borsboom & Mellenbergh, 2007).

12.3.3. *Counterfactual Causation*

Regularity causation allows one to list a set of inert variables over which generalizations hold. Thus, it would typically identify the factors that do make a difference negatively (i.e., as not occurring in the list). Among the factors that do make a difference, one may or may not find the focal construct one intended to assess. If one does, one knows one is on the right track. However, each new context, distinguished by a change in background variables, or an intervention in variables previously studied through passive observation, requires extensive new empirical study to support the generalization of the regularities to the new context. This limits the usefulness of regularity causation in facilitating applied testing programs.

A counterfactual model strengthens the claims that may be inferred. It does so by bolstering the measurement system by adding a primarily positive account that specifies the counterfactual dependence of the test

7.3.3. *Process Theories of Causation*

Both regularity theories and counterfactual theories share a general strategy of attempting to provide an account of causation that introduces as little as possible beyond the causes and the effects in question. Process theories go in the other direction by accepting a more robust explanatory base in order to provide a more successful account of causation (Markus, 2010). The present chapter uses the term 'process' to cover a range of theories that include processes (Dowe, 2007; Salmon, 1998), mechanisms (Bechtel & Richardson, 1993; Craver, 2007; Glennan, 2002), and natural or invented machines understood in a broad metaphorical sense (Cartwright, 1999). The critical element is that processes are understood as something more than just a chain of nomological connections between variables. Instead, processes exist independently and maintain such connections. Process theories take as their basic idea that the existence of a corresponding causal process separates causal regularities from noncausal regularities. For example, moving a light across the surface of a wall produces a regularity from illumination of one region to the next. Nonetheless, the illumination of the previous region does not cause the illumination of the next. Process theories explain this with the fact that no causal process follows the trajectory of the light along the wall (Salmon, 1998). However, a causal process does link the position of the light to the illumination, and this produces a genuine causal regularity.

Markus and Borsboom 2013

Summary (and practical advice)

- Appreciate that assessment of validity is not simple and requires time and multiple steps
- Understand the caveats of classical validity theory
- For building your own tests, **think content validity**
- For judging externally prepared tests/instruments, start with a clear definition of what's to be covered
- For criterion-related validity, take into account group variability; and think about validity of the criterion
- For your own assessments, try to eliminate the influence of any factors not related to what you want to measure
- To better understand the overall research questions, **think about consequential validity while developing your instrument (questionnaire)**

87

UCSF

RELIABILITY IN QUANTITATIVE MEASUREMENT

88

UCSF

UCSF

CTT in measuring reliability of measurements

- CTT is used to determine the degree to which variations in the conditions of measurement (e.g., different raters and different measurement occasions) affect the consistency with which a construct is measured
- CTT ***reliability coefficients are computed using the Pearson product–moment correlation***; only useful where we can identify the X and Y scores in the pair (X and Y).
- Intra-class correlation (ICC) to assess relations among elements within a class or a group (e.g., relations among the characteristics of siblings in a family)

89

UCSF

Why is reliability important?

- Reliability is a psychometric property that all measurement instruments must satisfy. (Stanley, 1971)
- A construct that enables researchers and instrument developers to examine the degree to which an instrument generates reproducible data
- Reliability is ***a measure of precision***
- Without reliability, measures are subject to large amounts of random variation that make measurements different from one occasion to the next. (i.e., measurements contain error)
- Researchers must demonstrate instruments are reliable since without reliability, research results are not replicable

90

UCSF

UCSF

Theoretical definition of reliability

- Most commonly, reliability is defined as the *squared* correlation between observed scores (*X*) and [hypothetical] *true* scores (*T*)

- Another common theoretical definition of reliability is expressed as the correlation between observed scores on two *parallel* measures (X- test and Y- criterion).

91

UCSF

Relationship between Reliability and Validity

- A *validity coefficient* can be no larger than the square root of the measure's reliability.

$$r_{xy} \leq \sqrt{(r_{xx})(r_{yy})}$$

  - where r_{xy} is the validity (correlation) coefficient, r_{xx} is the reliability of the x variable (test), and r_{yy} is the reliability of the y variable (criterion).
  - the **criterion validity coefficient** cannot be higher than the square root of the product of the reliabilities. It can, however, be lower than the product, because even if the x and y variables each had perfect reliability, the two variables may be completely unrelated to one another.

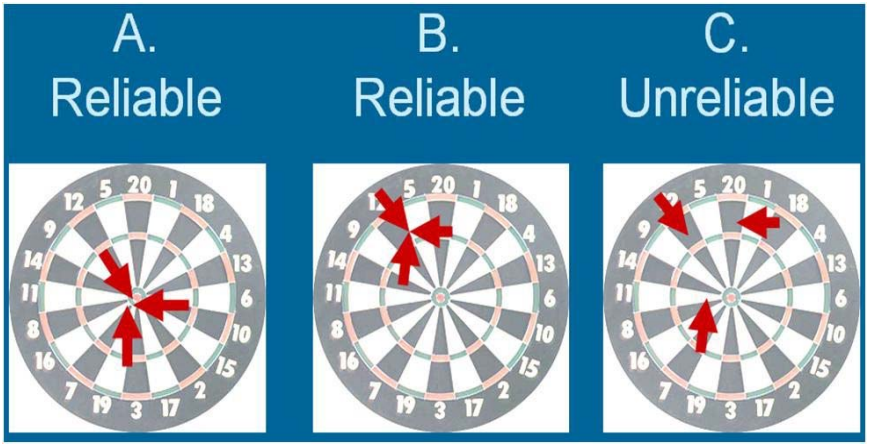
- *The validity of a test is constrained by how reliable it is.*
  - A test must do what it does consistently before we are sure it does what it says it does.

92

UCSF

UCSF

Reliability vs. Validity



93

UCSF

Reliability vs. Validity

- If an instrument is deemed “reliable”, it will be reliable across virtually any *normal* situation
- Conversely, validity is not assessed unequivocally… instead, one validates the *use* of a measure for a given situation. Consequently, if an instrument is felt to be valid for measuring certain phenomena, it might well *not be valid* for the measurement of related phenomena, or even the same phenomena under different circumstances.

94

UCSF

UCSF

Relationship Between Reliability and Validity

- Reliability is a *necessary* but NOT *sufficient* condition for validity.
- All "valid" measures are reliable—but not all reliable measures are valid.
  - Instruments could not be called valid, they are valid for specific studies and populations
  - Instruments could not be called reliable. Reliability comes from interaction of the instrument, people taking the test and the situation

95 UCSF

I. Classifications of methods for estimating reliability

(in relation to the measurement theory)

- Classical:

Based on the CTT, cannot partition error term

  - Internal consistency
    - Coefficient alpha
  - Parallel-alternate forms
    - Pearson correlation
    - Spearman Brown
  - Test-retest
    - Cohen's kappa
    - ICC
    - T-tests, Pearson correlation

- Additional:

Departs from CTT, could partition error terms into parts due to cases and raters

  - Inter-rater reliability
    - Cohen's kappa
    - ICC

96 UCSF

II. Classifications of methods for estimating reliability

(in relation to a number of tests)

Methods requiring a single test administration

- Internal consistency of the items within a measure (internal consistency reliability)

Methods requiring 2 test administrations

- Consistency across alternative forms of a measure (parallel forms reliability)
- Test-retest reliability: estimation based on the correlation between two administrations of the same scale at different
- Test-retest with alternate forms

97

UCSF

III. Classifications of methods for estimating reliability

(in relation to the major error source)

TABLE 7.4. Summary of Approaches to Reliability Estimation

Major Error Source	Reliability Coefficient	Data-Collection Procedure	Statistical Treatment of Data
1. Change in examinee's overtime	1. Stability coefficient	1. Test, wait, retest	1. Compute Pearson product moment coefficient, $\hat{\rho}_{12}$
2. Content sampling from form to form	2. Equivalence coefficient	2. Give form 1, give form 2	2. Compute Pearson product moment coefficient, $\hat{\rho}_{12}$
3. Content sampling, or flawed items	3. Internal consistency coefficient	3. Give one form on one occasion	3a. Divide test into halves; correlate half-test, $\hat{\rho}_{AB}$; use Spearman Brown correction: $\hat{\rho}_{XX'} = \frac{2\hat{\rho}_{AB}}{1+\hat{\rho}_{AB}}$
			b. Divide test into halves; use Guttman's or Rulon's formula: $\hat{\rho}_{XX'} = 1 - \frac{\hat{\sigma}_D^2}{\hat{\sigma}_X^2}$
			c. Compute item variances; compute coefficient alpha: $\hat{\alpha} \simeq \frac{k}{k-1}\left(1 - \frac{\Sigma\hat{\sigma}_i^2}{\hat{\sigma}_X^2}\right)$

98

UCSF

UCSF

Main types of reliability measurements and associated statistical measures

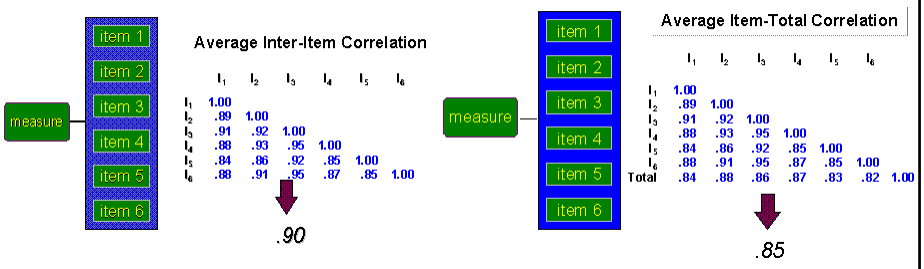
1. Internal consistency
2. Parallel/ alternate forms
3. Inter-rater agreement
4. Test-retest

99

UCSF

1. Internal Consistency Reliability

- Internal consistency of the items within a measure (internal consistency reliability)
- Requires a single test administration
- Main statistical methods:



100

UCSF

UCSF

1. Internal Consistency Reliability

- **Main statistical measures:**
  - Split-half technique:
    - One of the first methods for estimating internal consistency reliability. With this method, an instrument is divided into two parts. The two forms must be considered parallel or essentially equivalent. The two halves are then correlated. Because each "form" is only half as long as the original instrument, the correlation must be adjusted so that it provides a better estimate of the reliability of the "full length" instrument.
    - Spearman-Brown coefficient: estimation based on the correlation of two half-tests
    - Rulon or Guttman methods to estimate reliability of the full-length test from the scores of the two half-tests
    - Limitation: different methods of splitting the test yield different reliability estimates

UCSF

1. Internal Consistency Reliability

Split-Half Reliability

1. Divide items into 2 sets
   - E.g., odd versus even
2. Estimate alpha coefficient for each set
3. Create total scores for each set
4. Correlate totals
5. Calculate split-half formula using correlation

UCSF

UCSF

1. Internal Consistency Reliability

- **Main statistical measures (continued):**
  - Coefficient alpha methods based on item covariances
    - The average of all the split-half coefficients that would be obtained if the test were divided into all possible split-half combinations
      - Note that if the two halves are not even considered essentially equivalent, coefficient α still gives a *lower bound* to reliability, which is to say that instrument reliability will be greater than or equal to α.

UCSF

1. Internal Consistency Reliability Estimation

- **Types of Coefficient Alpha indices:**
  - **Cronbach's alpha:** estimation based on the correlation among the variables comprising the set
  - **Kuder-Richardson 20 or 21:** can be used only with dichotomously scored items
    - When items in the scale differ in difficulty, KR21 will be systematically lower than KR20

UCSF

UCSF

1. Internal Consistency Reliability
Coefficient Alpha

- Coefficient α is the average of all possible split half reliability estimates.
- If items are scored dichotomously, then coefficient α is equivalent to KR-20, the reliability coefficient proposed by G. Frederic Kuder and Marion Richardson (1937).
- Nunnaly (1978, p. 230) says, “Coefficient α provides a good estimate of reliability in most situations since the major source of measurement error is because of the sampling of content."

UCSF

1. Internal Consistency Reliability
Coefficient Alpha

Relationship Between Test Length and Reliability:

- Changes in test length can affect reliability. The generalized form of the Spearman-Brown "prophesy" formula is given by:

$$\rho_{XX'} = \frac{N\,\rho_{YY'}}{1 + (N-1)\,\rho_{YY'}}$$

- and can be estimated in the sample by:

$$\hat{\rho}_{XX'} \doteq r_{XX'} = \frac{N\,r_{YY'}}{1 + (N-1)\,r_{YY'}}$$. ***Note***: In this context, N is a scaling factor.

UCSF

UCSF

1. Internal Consistency Reliability
Coefficient Alpha

- If, for example, a 30-item instrument has a reliability of .70, what would its estimated reliability be if we added 15 more items?

$$r_{XX} = \frac{(1.5)(.70)}{1+(1.5-1)(.70)} = .78$$

107

UCSF

1. Internal Consistency Reliability
Coefficient Alpha

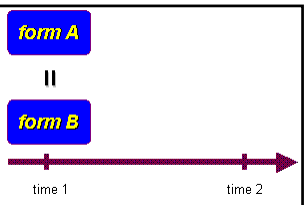
- Represents the extent to which performance on one item indicates performance on any other item in the measure
- Can expect higher alpha when there are more items, greater variance, a normal distribution and a heterogeneous sample
- The major gains in additional items up to approximately 10, when the increase in reliability for each additional item levels off.

108

UCSF

UCSF

2. Parallel/Alternate-Forms Reliability



- Designed to minimize the problems inherent in test-retest reliability estimation
- Defined as the correlation between *observed* scores on two parallel instruments
- Because it is virtually impossible to verify the parallelism between two measures, *alternate forms* are often used as a substitute for parallel forms
- *Alternate forms* are defined as instruments that were designed to be parallel, and have equal (or nearly equal) observed score means and variances
- Because it is impossible to prove that they produce identical true-scores, however, they must be considered alternate forms.

UCSF

2. Parallel/Alternate-Forms Reliability

Methods requiring 2 test administrations

- Consistency across alternative forms of a measure (parallel forms reliability)
  - Coefficient of **equivalence** (range of 0.80-0.90)
- Test-retest with alternate forms
  - Coefficient of **equivalence** and **stability** (the range is lower than above)

Coefficients: Pearson's correlation or Spearman rho for each of the scenarios above

UCSF

UCSF

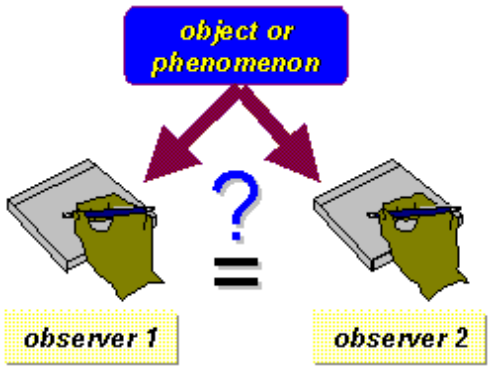
2. Parallel/Alternate-Forms Reliability

- Limitations:
  - Subject to carry-over and administration interval effects, though usually to a lesser extent. It is clearly inappropriate for characteristics that tend to change over time.
  - The other major problem with this method is the unavailability of alternate forms.

111

UCSF

3. Inter-Rater Reliability (inter-observer agreement)



112

UCSF

UCSF

Reliability of Ratings

- Sometimes, it is not possible, practical, or desirable to obtain direct measurement of a characteristic from individuals. Perhaps, for example, the subjects are incapable of responding (e.g., in a coma, psychotic, or experiencing severe pain). In cases such as these, it is often desirable to obtain ratings of the studied characteristic from trained observers. Unfortunately, many observed characteristics are susceptible to subjectivity on the part of the observers. For that reason, it is customary to establish the reliability of the ratings.

Correlation vs. Agreement

- In discussing the reliability of ratings, it is necessary to look at the difference between the concepts of correlation (or association) and agreement. Two raters could **consistently** rate a characteristic, but never agree. If, for instance, one rater was **always** two scale points above the other rater, the correlation would be 1.0, but they would **never** agree! Are they reliable?

Correlation vs. Agreement

- One way to get around this problem is to use rank-order correlations such as Spearman's ρ. If there are more than two raters, one could use Kendall's Coefficient of Concordance to estimate "reliability." While this seems to be an appropriate way of estimating reliability, all it really does is "cover up" possible disagreements between raters.
- If one is really more interested in **agreement** than correlation (as a measure of interrater reliability, there are general methods that are recommended: various forms of **Cohen's Kappa (κ)** and the **Intraclass Correlation Coefficient** (ICC).

115 UCSF

3. Inter-Rater Reliability

- Consistency across different raters or judges in use of the measure (inter-rater reliability
  - **Intra-class correlation coefficients** (Pearson Coefficients, t-tests) - only when you have meaningful pairings between two and only two raters
    - Typically, for norm-referenced measures
  - **Inter-rater reliability** (Kappa or ICC): estimation based on the correlation of scores between/among two or more raters
    - Typically, for criterion-referenced measures

116 UCSF

3. Inter-Rater Reliability
Going back to CTT…

$$Observed\ Score = True\ Score + Measurement\ Error \qquad X = T + E$$

$$Var(X) = Var(T) + Var(E)$$

$$Reliability = \frac{Var(T)}{Var(X)} = \frac{Var(X) - Var(E)}{Var(X)} = \frac{Var(T)}{Var(T) + Var(E)}$$

IRR aims to determine how much of the variance in the observed scores is due to variance in the true scores after the variance due to measurement error between coders has been removed

an IRR estimate of 0.80 would indicate that 80% of the observed variance is due to true score variance or similarity in ratings between coders, and 20% is due to error variance or differences in ratings between coders

117

UCSF

3. Inter-Rater Reliability
Cohen's Kappa

- Measures agreement between two raters or coders
- Measures percentage of agreement in the main diagonal of a 2x2 table and then adjusts the values for chance

118

UCSF

UCSF

3. Inter-Rater Reliability
Cohen's Kappa (κ)

- Defined as a chance-corrected measure of percent agreement with a statistical base, and given by:

$$K = \frac{P(a) - P(e)}{1 - P(e)}$$

where P(a) denotes the observed percentage of agreement, and P(e) denotes the probability of expected agreement due to chance

- test the statistical significance of κ by:

$$Z = \frac{\kappa}{SE_\kappa}$$

where SE_k denotes standard error of K

119

UCSF

http://epiville.ccnmtl.columbia.edu/popup/how_to_calculate_kappa.html

3. Inter-Rater Reliability
Cohen's Kappa (κ)

$$K = \frac{P(a) - P(e)}{1 - P(e)}$$

$$\text{Observed Agreement} = \frac{(a + d)}{N} \qquad \text{Expected Agreement} = \frac{(\text{Expected (a)} + \text{Expected (d)})}{N}$$

- P(a) – observed agreement=0.7
- P(e) – expected agreement=0.5
- K = (0.7 – 0.5) / (1 – 0.5)
- K=0.4

Example:

Rater 2	Rater 1		TOTAL
	Yes	No	
Yes	40	20	60
No	10	30	40
TOTAL	50	50	100

$$Z = \frac{\kappa}{SE_\kappa}$$

$$Z = \frac{0.4}{0.098} = 4.08 \quad (p < 0.0001)$$

120

UCSF

UCSF

3. Inter-Rater Reliability
Cohen's Kappa (κ)

Landis and Koch (1977, p.165) suggested the following guideline for the evaluation of Kappa:

kappa	interpretation
< 0	poor agreement
0 - 0.20	slight agreement
0.21 - 0.40	fair agreement
0.41 - 0.60	moderate agreement
0.61 - 0.80	substantial agreement
0.81 - 1.00	almost perfect agreement

Kappa is scaled such that 0 is the amount of agreement that would be expected by chance, and 1 is perfect agreement.

121

UCSF

3. Inter-Rater Reliability
Cohen's Kappa (κ)

- Actually, κ ranges from -1 to +1.0. When κ is negative, agreement is less than chance. κ will equal 1 only when agreement is perfect *and* marginal probabilities are equal. It should also be noted that the statistical significance test for κ is often criticized as appearing inordinately high and out of step with its raw value. That is, relatively small κs often yield large Z statistics. You should also be aware that κ is designed for nominal categories. If your categories are ordered, you should use a special form of κ called "weighted κ."

- Rater drift can occur when an individual rater alters the way he or she applies the scoring criteria (i.e., becoming more lenient or stringent) over time.

122

UCSF

UCSF

3. Inter-Rater Reliability
Cohen's Kappa (κ) – different types

- The Kappa statistic is the most widely used reliability coefficient when the variables being compared for agreement are unordered categorical (nominal level of measurement).
- Cohen's original (1960) kappa is subject to biases in some instances and is only suitable for fully-crossed designs with exactly two coders.
- For ordered categorical data (ordinal level of measurement), the weighted kappa is used.

PABAK (Prevalence and Bias Adjusted Kappa)

- When the observed cells counts in a cross-tabulation table of two raters bunch up in any corner of the table, the kappa coefficient is known to produce paradoxical results. An alternative form of kappa, called PABAK, has been proposed as a solution to this problem.

123 UCSF

3. Inter-Rater Reliability
Intraclass Correlation Coefficient

- The intraclass correlation coefficient (ICC) is an excellent measure of agreement in virtually all situations.
- ICC can be thought of as an "average" correlation between several sets of ratings. The Pearson correlation coefficient is an interclass correlation.
- The ICC is appropriate for both nominal and ordinal data, but is especially good for continuous rating (interval scale). In the latter case, the Repeated Measures ANOVA can be used to compute the ICC.
- ICC compares the variability of different ratings of the same subject to the total variation across all ratings and all subjects.
- Also called the *intracluster correlation coefficient (ICC)*, or the *reliability coefficient* (*r* or *rho*).

124 UCSF

3. Inter-rater Reliability
Intraclass Correlation Coefficient

Subject	Rater 1	Rater 2	Rater 3
1	10	11	12
2	5	6	
3	0	0	0
4	15	15	15
5		13	13

- there are incomplete ratings for two subjects

UCSF

3. Inter-rater Reliability
Intraclass Correlation Coefficient

Major advantages:

- Computation of the ICC, unlike most other interrater reliability coefficients, does not require a complete set of ratings for each subject. This is a major advantage.
- presumes that one has the results of a Repeated Measures ANOVA (Analysis of Variance)

UCSF

UCSF

ANOVA MODELS

Case 1	Raters for each subject are selected at random
Case 2	The same raters rate each case. These are a random sample.
Case 3	The same raters rate each case. These are the only raters.

- One-Way Random
  - Raters are not consistent for all cases
  - No attempt to look at separate effects of raters and cases – examines only one effect
  - Assumes raters are randomly drawn from a population of possible raters
  - Also called ICC-1, usually the smallest of ICCS

The patient is treated as a random effect and observer as a fixed effect. Thus, we are making an inference to the population of all patients, but only to these specific observers.

The interpretation is that x% of the variance in the observations was the result of "true" background variability among patients, or patient-to-patient differences, rather than due to lack of agreement among the raters or measurement error. It is a round-about way of describing the agreement among the raters.

127

UCSF

ANOVA MODELS

Case 1	Raters for each subject are selected at random
Case 2	The same raters rate each case. These are a random sample.
Case 3	The same raters rate each case. These are the only raters.

- One-Way Random
- Two-Way Random
  - Raters are the same for all cases
  - Models a separate effect for raters and for cases (2 Way)
  - Assumes raters are randomly drawn from a population of possible raters
  - Treats the rater and rate effects as random
  - Also called ICC-2

128

UCSF

UCSF

ANOVA MODELS

Case 1	Raters for each subject are selected at random
Case 2	The same raters rate each case. These are a random sample.
Case 3	The same raters rate each case. These are the only raters.

- One-Way Random
- Two-Way Random
- Two-Way Fixed
  - Raters are the same for all cases
  - Models a separate effect for raters and for cases (2 Way)
    - The only raters of interest (fixed)
  - Also called ICC-3, usually the largest of ICCs

Assumes that raters are not random but are fixed, i.e., the raters are the only raters anyone would be interested in. (This is uncommon in coding, because theoretically your research assistants are only a few of an unlimited number of people that could make these ratings. This means ICC(3) will also always be larger than ICC(1) and typically larger than ICC(2), because 1) it models both an effect of rater and of ratee (i.e. two effects) and 2) assumes a random effect of ratee but a fixed effect of rater (i.e. a mixed effects model).

129 UCSF

3. Inter-rater Reliability
Intraclass Correlation Coefficient

Determining which ICC you need:

1. Do you have consistent raters for all ratees? For example, do the exact same 8 raters make ratings on every ratee?

No – ICC-1, Yes →Q2

2. Do you have a sample or population of raters?

Sample – ICC-2, Population – ICC-3

3. Do you need a measure of absolute agreement or consistency?

Example: Variable 1 with values 1, 2, 3 and Variable 2 with values 7, 8, 9. Even though these scores are very different, the correlation between them is 1 – so they are highly consistent but don't agree.

4. Are you interested in the reliability of a single rater, or of their mean?

Agreement between the averages of ratings over several raters (average ICC) are typically used when individual ratings are deemed unreliable.

130 UCSF

Tutor Quant Methods Psychol. 2012 ; 8(1): 23–34.

Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial

Kevin A. Hallgren

Summary of ICC statistic parameters

Statistical Family	Parameter	Uses
Intraclass Correlation		Agreement for ordinal, interval, or ratio variables
	Model	*One-way*: Raters randomly sampled for each subject *Two-way*: Same raters across subjects
	Type	*Absolute agreement*: IRR characterized by agreement in absolute value across raters *Consistency*: IRR characterized by correlation in scores across raters
	Unit	*Average-measures*: All subjects in study rated by multiple raters *Single-measures*: Subset of subjects in study rated by multiple raters
	Effect	*Random*: Raters in study randomly sampled and generalize to population of raters *Mixed*: Raters in study not randomly sampled, do not generalize beyond study

UCSF

3. Inter-rater Reliability
Intraclass Correlation Coefficient

Cicchetti and Sparrow (1981) suggested the following guideline for the evaluation of ICC used to measure inter-rater agreement:

ICC	interpretation
<0.40	Poor agreement
0.40 - 0.59	fair to moderate agreement
0.60 - 0.74	good agreement
0.75 - 1.00	excellent agreement

UCSF

UCSF

Equivalence of Kappa and ICC

Unweighted Kappa

- For dichotomous ratings, kappa and ICC are equivalent.

Weighted Kappa

- For order categorical (ordinal scale) ratings, the weighted kappa and ICC are equivalent, provided the weights are taken as

$$w_{ij} = 1 - \frac{(i-j)^2}{(k-1)^2}$$

133

UCSF

3. Inter-Rater Reliability
Limitations of ICCs

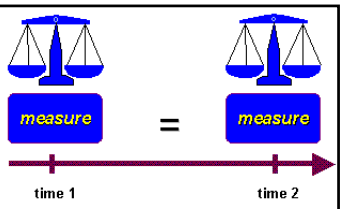
- The ICC is strongly influenced by the variance of the trait in the sample/population in which it is assessed. ICCs measured for different populations might not be comparable.
  - For example, suppose one has a depression rating scale. When applied to a random sample of the adult population the scale might have a high ICC. However, if the scale is applied to a very homogeneous population--such as patients hospitalized for acute depression--it might have a low ICC.
- Assumes equal spacing between categories
- Association vs. bias
  - The ICC combines, or some might say, confounds, two ways in which raters differ: (1) *association*, which concerns whether the raters understand the meaning of the trait in the same way, and (2) *bias*, which concerns whether some raters' mean ratings are higher or lower than others.

134

UCSF

UCSF

4. Test-Retest Reliability



- Estimates reliability by generating the correlation between two administrations of the same test.
- If all individuals score exactly the same on both administrations, then the instrument is said to be *perfectly* reliable. This is the same as saying that there is a perfect linear relationship between both sets of *observed* scores. The degree to which this relationship is not perfect estimates the amount of unreliability.

135

UCSF

4. Test-Retest Reliability

- Test-retest reliability estimation is most appropriate in situations where the measured characteristic is not susceptible to carry-over and time interval effects or where the investigator wishes to show the temporal stability of a measure.
- Example: physiological or sensory characteristics, although even these characteristics can be somewhat affected -- a person having his blood pressure taken a second time may be less apprehensive than they were the first time.

136

UCSF

UCSF

4. Test-Retest Reliability

Limitations:

- *Carry-over effect* – The first administration of the instrument may influence individuals' responses on the second administration. This is also commonly referred to as a *practice* effect. Because this effect manifests itself in many different ways, it is hard to predict whether it will produce an underestimate or an overestimate of reliability. Worse yet, there is likely to be much individual variation, where some people might remember their answers, others might get bored and misread or mismark items, while others might seek advice or knowledge between administrations.
- *Administration interval* – The length of time between administrations is likely to cause problems. A short interval usually results in greater carry-over effects. A long interval may increase the effects of history, maturation or affective state.

UCSF

4. Test-Retest Reliability

- Measures stability of the measure over time
  - Norm-referenced measures: Pearson correlations
  - Criterion-referenced measures: Cohen's kappa

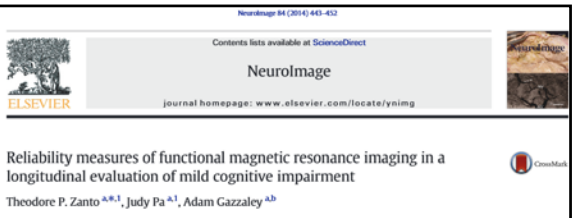
- **Test-retest reliability: estimation based on the correlation between two administrations of the same scale at different times**
  - Coefficient of stability (range of 0.70-0.90)
- **Test-retest with alternate forms**
  - Coefficient of equivalence and stability (lower range)

Coefficients: **Pearson's correlation or Spearman rho for each of the scenarios above**

UCSF

UCSF

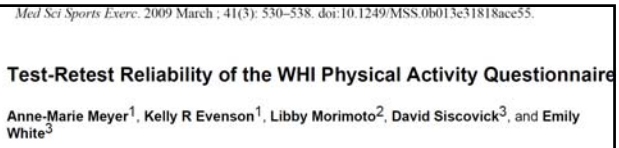
Example 1:



- To characterize amnestic mild cognitive impairment (aMCI), a preclinical stage of Alzheimer's disease (AD) using Functional magnetic resonance imaging (fMRI)
- aMCI participants completed two visual working tasks, a Delayed-Recognition task and a One-Back task, on three separate scanning sessions over a three-month period
- Reliability measures: fMRI blood oxygen level dependent (BOLD) activity test-retest reliability using ICC
- Results: frontal lobe, medial temporal lobe, and subcortical structures exhibited fair to moderate reliability (ICC = 0.3–0.6),while temporal, parietal, and occipital regions exhibited moderate to good reliability (ICC = 0.4–0.7)
- Reliability across brain regions was more stable when three fMRI sessions were used in the ICC calculation relative to two fMRI sessions

139

Example 2:



- Few physical activity (PA) questionnaires were designed to measure the lifestyles and activities of women. They do not collect detailed information on types of activities and use terminology many women do not identify with.
- PA in the Women's Health Initiative (WHI) Study: usual frequency, duration, and pace of recreational walking, frequency and duration of other recreational activities or exercises (mild, moderate and strenuous), household, and yard activities.
- Test-retest reliability of a PA questionnaire (3 months): 569 women repeated questions on recreational PA and 523 women repeated questions related to household and yard activities

140

Table 4
Kappa statistics and 95% confidence intervals (CI) of the physical activity components measures overall and among women who reported recreational activity in the WHI Measurement and Precision Study

	Entire sample			Women with ≥1 episode of recreational activity[#]		
	N	Weighted Kappa	95% CI	N	Weighted Kappa	95% CI
Form 34: Exercise or recreational physical activity						
Mild physical activity, days per week	548	0.36	0.27, 0.45	303	0.40	0.28, 0.51
Mild physical activity, minutes per session	528	0.52	0.43, 0.61	295	0.51	0.40, 0.62
Moderate physical activity, days per week	563	0.53	0.47, 0.59	314	0.54	0.47, 0.47
Moderate physical activity, minutes per session	544	0.48	0.41, 0.54	297	0.44	0.36, 0.53
Strenuous physical activity, days per week	555	0.62	0.55, 0.68	314	0.62	0.55, 0.69
Strenuous physical activity, minutes per session	546	0.61	0.54, 0.68	306	0.60	0.52, 0.68
Number of walks per week >= 10 minutes	567	0.60	0.55, 0.65	314	0.55	0.48, 0.61
Minutes per walk	555	0.59	0.54, 0.65	307	0.59	0.52, 0.66
Usual speed of walk	556	0.60	0.54, 0.65	306	0.58	0.50, 0.66
Total exercise and recreational activity exposure	569	0.61	0.56, 0.66	314	0.51	0.42, 0.59
Form 42: Household and yard physical activity						
Heavy indoor chores hours per week	517	0.52	0.45, 0.58	N/A[*]		
Yard work, months per year	511	0.67	0.62, 0.71			
Yard work, hours per week	509	0.64	0.59, 0.70			
Historical strenuous physical activity						
Strenuous physical activity at age 18 years	527	0.55[**]	0.48, 0.63	288	0.57[**]	0.47, 0.66
Strenuous physical activity at age 35 years	526	0.55[**]	0.48, 0.63	294	0.55[**]	0.45, 0.65
Strenuous physical activity at age 50 years	535	0.53[**]	0.46, 0.60	301	0.53[**]	0.44, 0.63

[*] Only applicable for women who completed the recreational physical activity form

[#] One episode of any recreational physical activity, regardless of intensity or duration

[**] Simple kappa statistic

Med Sci Sports Exerc. 2009 March ; 41(3): 530–538. doi:10.1249/MSS.0b013e31818ace55.

Example 2:

Test-Retest Reliability of the WHI Physical Activity Questionnaire

Anne-Marie Meyer[1], Kelly R Evenson[1], Libby Morimoto[2], David Siscovick[3], and Emily White[3]

Results:

- No meaningful differences were observed by race/ethnicity, age, time between test and retest, and amount of reported PA.
- Questions on recreational walking, moderate recreational PA, and strenuous recreational PA had higher test-retest reliability (weighted kappa: 0.50-0.60), than questions on mild recreational PA (weighted kappa: 0.35-0.50).

Conclusions—The WHI PA questionnaire demonstrated moderate to substantial test-retest reliability in a diverse sample of post-menopausal women.

Agree??

142

Summary
Sources of Random Error Which Affect Reliability

- Imprecision in the measure itself
  - Error in instrument construction
- Differences in measurement time or context
  - Environmental or biological factors
- Differences in administration or scoring
  - Differential standards when rules unclear
  - Rating wrong characteristic
  - Central tendency error

143

UCSF

Summary
Which Reliability Estimate Should One Use?

- If a sample of examinees is highly homogeneous on the trait being measured, the reliability estimate will be lower than if the sample were more heterogeneous
- Longer tests are more reliable than shorter tests
- Pearson's correlation coefficient is an inappropriate measure of reliability because the strength of linear association, and not agreement, is measured (it is possible to have a high degree of correlation when agreement is poor)

144

UCSF

UCSF

Summary

- Reliability conducted on one occasion or with one set of subjects does not assure reliability on other occasions or with other populations
- Ideally, reliability should be examined each time a measure is employed

145

UCSF

TULIPMANIA AT PIER 39

FEBRUARY 11 - FEBRUARY 19

65th Pacific Orchid & Garden Exposition

Big Ideas for Small Gardens

February 24–26, 2017

Fri. & Sat. 9am–6pm • Sun. 10am–5pm

146

UCSF

UCSF

ADDITIONAL SLIDES

EXAMPLES OF MODERN MEASUREMENT THEORIES: Generalizability Theory

Examples of modern measurement theories

- Generalizability theory – acknowledges the multiple sources of measurement error by deriving estimates of each source separately
- Item response theory – in contrast to classical measurement theory, which focuses on test performance, this theory focuses on item's performance and the examinee's ability; eg., examinee's performance could be predicted based on a set of factors referred to as traits, latent traits, abilities or proficiencies

Generalizability theory (GT), Cronbach, 1963

- **Classical true score model considers only persons and items/scales**
- GT takes into account the various sources of error that affect measurement
  - assumes that an observed score is a linear combination of true score and error
  - error is multi-faceted and can be systematic
  - Population, Time, Setting (ecology)
- Acknowledges, but also accounts for, the differences in measurement conditions that one is likely to encounter in applied measurement contexts

149 UCSF

GT vs. CTT

- CTT – the major concern with reliability is how well observed scores reflect corresponding true scores
- GT – attention focuses on dependability, or how well observed scores allow generalization about behavior in a defined universe of situations

150 UCSF

Generalizability Theory

- Definition & establishment of the universe of admissible observations:
  - observations that the decision maker is willing to treat as interchangeable
  - all sources of influence acting on the *measurement* of the trait under study
- Score's usefulness depends on the degree to which it allows one to generalize accurately to some wider set of situations, *the universe of generalizations*

151

GT, *continued*

- Rather than attempting to estimate a true score, as in CTT, the focus of GT is on the universe score, or the average score, that would be expected across all possible variations in the measurement procedure (e.g., different raters, forms, or items).
- Because no single obtained score is likely to perfectly match the universe score, some degree of error exists when generalizing from a particular sample of behavior to the universe score.
- The quantification of this error is a central focus of GT and is represented through the calculation of generalizability coefficients and dependability coefficients along with consideration of variances attributable to various predefined sources (e.g., persons, occasions, and raters).

152

One Facet Design
Variance Components

- 4 Sources of Variability
  - systematic differences among subjects
    - (object of measurement)
  - systematic differences among raters (occasions, items)
  - subjects*raters interaction } confounded
  - random error

153

UCSF

Two Facet Fully Crossed Design
Variance Components

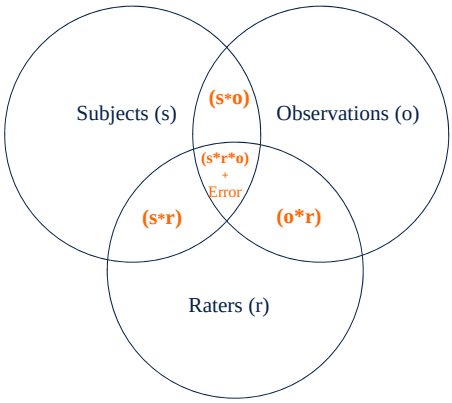
- **Seven sources of variance are estimated:**
  - subjects
  - raters
  - observations
  - s_xr
  - s_xo
  - r_xo
  - s_xr_xo,e

154

UCSF

UCSF

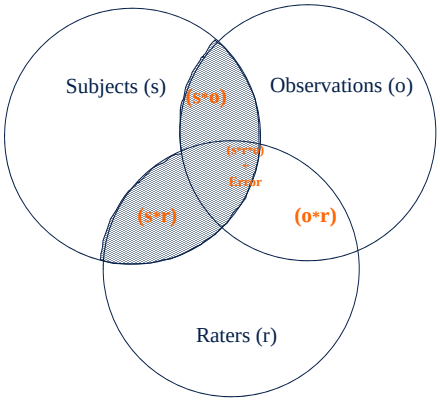
Variance Components



155

UCSF

Relative Error
Facet of Determination: Subjects

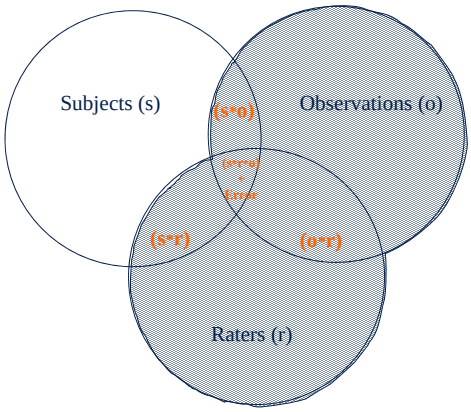


156

UCSF

UCSF

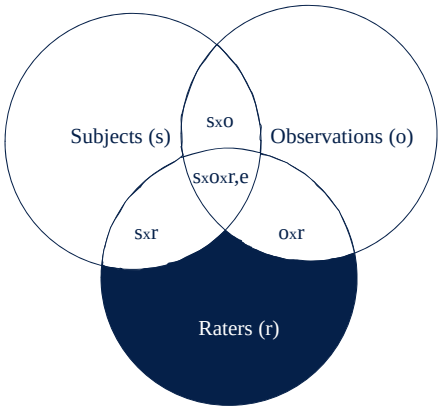
Absolute Error
Facet of Determination: Subjects



157

UCSF

D-Study: consider raters as fixed, no generalization made to other raters



Variation in **s** is no longer affected by raters:

a) average over the levels the fixed effect

b) analyze the fixed effect levels separately

158

UCSF

UCSF

Sources of variation

- **ANOVA model is the basis for sources:**
- **Between-subject sources:**
  - examples: Gender, Ethnicity, Ability, Person
- **Within-subject sources**
  - examples: Item, Time, Rater

159

UCSF

Sources of variation

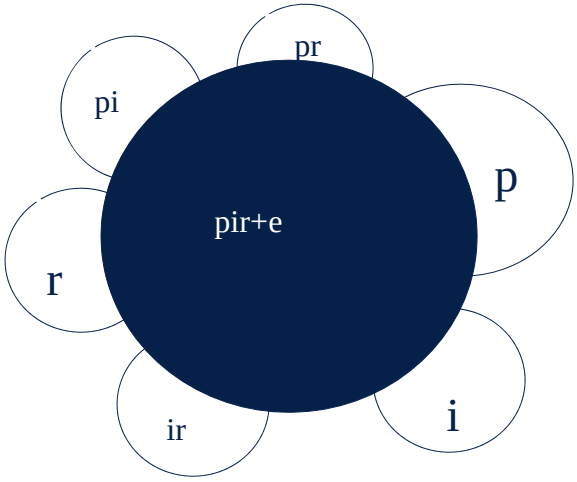
- **ASSUMPTIONS (random effects model):**
  - Person is random
  - Item is random
  - Rater is random

160

UCSF

UCSF

Sources of variation



161

UCSF

Issues for Fixed Facets

- Large variance due to fixed facet creates artificially high g-coefficient
- Preferable to do SEPARATE analyses for each level of the fixed facet
- Interpret meaning of large variance component for fixed facet

162

UCSF

UCSF

D-studies and coefficients

- Change number of facet levels to do “what if” studies
- Select specific subsets for specific levels (eg. specific raters, subsets of raters)
  - May be random or fixed factors, depending on if the raters have a particular characteristic (random) or those specific raters will be used (fixed)

163

UCSF

Generalizability Theory
Quantitative Approach

- The Intraclass Correlation (ICC) assesses rating reliability by comparing the variability of different ratings of the same subject to the total variation across all ratings and all subjects
- Partitions variance due to between subjects and within subjects
  - Has some ability to accommodate multiple sources of variance
  - Does not provide an integrated approach to estimating reliability under multiple conditions (e.g., only 1- and 2-way ANOVA)

164

UCSF

UCSF

Generalizability Theory
Quantitative Approach

ANOVA for GT

- just as ANOVA partitions a dependent variable into effects for the independent variable (main effects & interactions), G-theory uses ANOVA to partition an individual's measurement score into an effect for the universe-score and an effect for each source of error and their interactions in the design.

- **ICC partitions variance due to between subjects and within subjects**
  - assesses rating reliability by comparing the variability of different ratings of the same subject to the total variation across all ratings and all subjects
  - Has some ability to accommodate multiple sources of variance
  - Does not provide an integrated approach to estimating reliability under multiple conditions (e.g., only 1- and 2-way ANOVA)

165 UCSF

Statistical Modeling

- In ANOVA we were driven to test specific hypotheses about our independent variables and thus sought out the F statistic and p-value.
- In G-theory we use ANOVA to partition the different sources of variance and then to estimate their amount (Variance Component).

166 UCSF

Example of G-coefficients estimation (with interpretations)

Journal of Early Adolescence
XX(X) 1–28
© The Author(s) 2013
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0272431613482044
jea.sagepub.com
SAGE

Generalizability Theory as a Unifying Framework of Measurement Reliability in Adolescent Research

Xitao Fan[1] and Shaojing Sun[2]

Example:

- 10 adolescents (P) are rated by two raters (R) on two occasions (O)
- (P: object of measurement), and both *rater* (R) and *occasion* (O) may introduce measurement unreliability, thus both being the facets.
- Goal: the reliability of the average ratings across the two raters and across the two occasions
- ICC cannot handle error sources *rater* and *occasion*
- → *use 3-way ANOVA:* 3 main effects (*P*, *O*, *R*), 3 two-way interactions (*P*O, P*R, O*R*), and 1 three-way interaction (*P*O*R*) confounded with error → 7 estimable variance components.

Table 7. Ratings for 10 Adolescents (P) by Two Raters (R) on Two Occasions (O).

	Occasions			
	1		2	
	Raters		Raters	
Adolescent	A	B	A	B
1	3	5	3	5
2	5	7	7	7
3	6	6	6	5
4	6	6	6	6
5	5	6	5	4
6	5	7	7	8
7	3	5	3	5
8	4	5	5	6
9	5	5	6	9
10	7	6	7	6
$\bar{X}$	4.9	5.8	5.9	6.1

169

UCSF

Selected Output:

Variance Estimates

Component	Estimate		
Var(Person)	.617	0.617	31%
Var(Occasion)	.061	0.061	3%
Var(Rater)	.239	0.239	12%
Var(Person * Occasion)	.289	0.289	15%
Var(Person * Rater)	.311	0.311	16%
Var(Occasion * Rater)	-.028[a]	-0.028	0%
Var(Error)	.503	0.503	25%
		1.992	100%

Dependent Variable: SCORE
Method: ANOVA (Type I Sum of Squares)

a. For the ANOVA and MINQUE methods, negative variance component estimates may occur. Some possible reasons for their occurrence are: (a) the specified model is not the correct model, or (b) the true value of the variance equals zero.

Table 8. Variance Component Estimates for Table 8 Data.

Variance component	Estimate	%
σ^2_p	.61667	30.54
σ^2_o	.06111	3.02
σ^2_r	.23889	11.83
σ^2_{p*o}	.28889	14.31
σ^2_{p*r}	.31111	15.41
σ^2_{o*r}	−.02778⇒	.00
$\sigma^2_{p*r*o*e}$	.50278	24.90

170

UCSF

Table 1. Generalizability Formulas for Relative and Absolute Decisions for a Two-Facet Crossed G-study (s × i × o)

Relative Decisions

$$\sigma^2_{Rel} = \frac{\sigma^2_{si}}{n'_i} + \frac{\sigma^2_{so}}{n'_o} + \frac{\sigma^2_{sio,e}}{n'_i n'_o}$$

$$G_{Rel} = \frac{\sigma^2_s}{\sigma^2_s + \sigma^2_{Rel}}$$

Absolute Decisions

$$\sigma^2_{Abs} = \frac{\sigma^2_i}{n'_i} + \frac{\sigma^2_o}{n'_o} + \frac{\sigma^2_{si}}{n'_i} + \frac{\sigma^2_{so}}{n'_o} + \frac{\sigma^2_{io}}{n'_i n'_o} + \frac{\sigma^2_{sio,e}}{n'_i n'_o}$$

$$G_{Abs} = \frac{\sigma^2_s}{\sigma^2_s + \sigma^2_{Abs}}$$

Burns et al. 1998

Note: σ^2_{Rel} = total relative error variance; G_{Rel} = generalizability coefficient for relative decisions; σ^2_{Abs} = total absolute error variance; G_{Abs} = generalizability coefficient for absolute decisions; n'_i = number of items; n'_o = number of occasions. Adopted from Shavelson & Webb, 1991, pp. 89 & 93.

Legend: s→p; i→o; o→r

Selected Output:

Variance Estimates

Component	Estimate		
Var(Person)	.617	0.617	31%
Var(Occasion)	.061	0.061	3%
Var(Rater)	.239	0.239	12%
Var(Person * Occasion)	.289	0.289	15%
Var(Person * Rater)	.311	0.311	16%
Var(Occasion * Rater)	-.028[a]	-0.028	0%
Var(Error)	.503	0.503	25%
		1.992	100%

Dependent Variable: SCORE
Method: ANOVA (Type I Sum of Squares)

Absolute decision G-coefficient (aka index of dependability) is appropriate when we are interested in the consistency of both relative positions and actual scores, and the error term includes all the components *except* the object of measurement *P (identical to ICC)*

$$G_{abs} = \frac{.61667}{.61667 + .006111/2 + 0.23889/2 + 0.28889/2 + 0.31111/2 + 0 + .50278/4} = 0.53$$

same as Φ (dependability index) =0.53

Conclusion: Need to increase the number of items or occasions or raters to achieve better dependability.

171

UCSF

Table 1. Generalizability Formulas for Relative and Absolute Decisions for a Two-Facet Crossed G-study (s × i × o)

Relative Decisions

$$\sigma^2_{Rel} = \frac{\sigma^2_{si}}{n'_i} + \frac{\sigma^2_{so}}{n'_o} + \frac{\sigma^2_{sio,e}}{n'_i n'_o}$$

Formula for dependability coefficient

$$\Phi = \frac{\sigma^2_p}{\sigma^2_p + \sigma^2_\Delta}$$

$$G_{Rel} = \frac{\sigma^2_s}{\sigma^2_s + \sigma^2_{Rel}}$$

Absolute Decisions

$$\sigma^2_{Abs} = \frac{\sigma^2_i}{n'_i} + \frac{\sigma^2_o}{n'_o} + \frac{\sigma^2_{si}}{n'_i} + \frac{\sigma^2_{so}}{n'_o} + \frac{\sigma^2_{io}}{n'_i n'_o} + \frac{\sigma^2_{sio,e}}{n'_i n'_o}$$

$$G_{Abs} = \frac{\sigma^2_s}{\sigma^2_s + \sigma^2_{Abs}}$$

Burns et al. 1998

Note: σ^2_{Rel} = total relative error variance; G_{Rel} = generalizability coefficient for relative decisions; σ^2_{Abs} = total absolute error variance; G_{Abs} = generalizability coefficient for absolute decisions; n'_i = number of items; n'_o = number of occasions. Adopted from Shavelson & Webb, 1991, pp. 89 & 93.

Legend: s→p; i→o; o→r

$$\rho^2 = \frac{\sigma^2_T}{\sigma^2_T + \sigma^2_E} = \frac{\sigma^2_p}{\sigma^2_p + \left(\frac{\sigma^2_{po}}{n_o} + \frac{\sigma^2_{pr}}{n_r} + \frac{\sigma^2_{por,e}}{n_o n_r}\right)}$$

$$= \frac{.61667}{.61667 + \frac{.28889}{2} + \frac{.31111}{2} + \frac{.50278}{4}}$$

$$= .59$$

Relative decision G-coefficient is appropriate if we are only interested in the consistency of relative positions of the individuals, and the error term includes all interaction components involving the object of measurement *P (conceptually similar to reliability coefficients)*

Conclusion: Improve the measurement process and measurement protocol to achieve better measurement reliability.

172

UCSF

UCSF

Summary

- Generalizability is broadly understood as **representativeness of a specific sample of the traits and characteristics in the entire population**
- Generalizability Theory (GT) offers increased utility for assessment research given the ability to concurrently examine multiple sources of variance, inform both relative and absolute decision making, and determine both the consistency and generalizability of results.

173

Summary

- The G-coefficient (*relative* decision) is equivalent to coefficient alpha
- G-coefficient (*relative* decision) is *conceptually* the same as the reliability coefficients based on Pearson *r* (i.e., interrater, test-retest, and parallel form reliability coefficients), and it is *mathematically* equivalent to Pearson *r* when the two scores to be correlated in Pearson *r* have equal variances
- G-coefficient (*absolute* decision) is both conceptually and mathematically equivalent to *ICC*s

174

Additional example:

Research in Nursing & Health, 1998, 21, 83-90

Focus on Quantitative Methods

Beyond Classical Reliability: Using Generalizability Theory to Assess Dependability

Karyl J. Burns*

- Self-report physical activity questionnaire: Habitual Physical Activity Index (HPAI)
- Created in the Netherlands in 1982 using a sample of young adults
- Modified for American older adults in 1993 by Burns et al.
- Uses an 8-item index for work and an 8-item index for leisure physical activity

Step 1 Goals: The goal is to determine the optimal conditions for reliable use of the modified HPAI. Assessment of the instrument's generalizability for making relative and absolute decisions was also of interest → **G-study.**

Step 2 Sources of variance: Important facets are items and occasions → **two-facet design**.

Step 3 Sampling: To determine a number of items and occasions that a D-study would need to assess habitual physical activity with acceptable reliability, a fully-crossed deign would be necessary.

Step 4 Study design: Because the researchers anticipated that the HPAI may be used to assess a person's relative standing among peers as well as to quantify an absolute amount of habitual physical activity, information about **generalizability coefficients for relative and absolute decisions** was needed. The authors accepted the principle of exchangeability and designated both facets, items and occasions, and the object of measurement, subjects, as **random**.

175 **Step 5 Stat method and software**: ANOVA

UCSF

Additional example:

Focus on Quantitative Methods

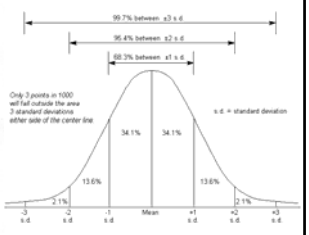
Beyond Classical Reliability: Using Generalizability Theory to Assess Dependability

Karyl J. Burns*

Table 2. Variance Components and Percent of Total Variance for the Work and Leisure Indices of the HPAI

	Work Index		Leisure Index	
Source	Component	%	Component	%
Subjects (*s*)	.317	30	.258	24
Occasions (*o*)	.000	0	.000	0
Items (*i*)	.265	25	.169	16
s × *i*	.281	27	.385	37
s × *o*	.012	1	.003	0
i × *o*	.001	0	.001	0
s × *i* × *o*, *e*[a]	.166	16	.237	22

[a] *s* × *i* × *o*, *e* = subject by item by occasion interaction plus random variance.



Step 6 G-study: A large proportion of error stemmed from the facet of items (25% and 16%) and the interaction of subjects and items (27% and 37%). Variance component for items is 0.265, i.e. SD=0.51, which means that 95% of the means lie within =/- 2 SD or 0.51*4=2.04 → **fairly wide distribution for a mean when the scale goes from 1 to 5.**

Conclusion: HPAI is useful in spreading people out on the continuum of physical activity. Agree?

176

UCSF

UCSF

Item Response Theory

- We will discuss it next week in *Measurement Theory and Practice II*
- IRT is used in Factor Analysis/ Principal Component Analysis and in Item Analysis.

177