JOHNS HOPKINS
BLOOMBERG
SCHOOL of PUBLIC HEALTH

Measurement: Reliability and Validity Measures

Jonathan Weiner, DrPH
Johns Hopkins University

JOHNS HOPKINS
BLOOMBERG
SCHOOL of PUBLIC HEALTH

Section A

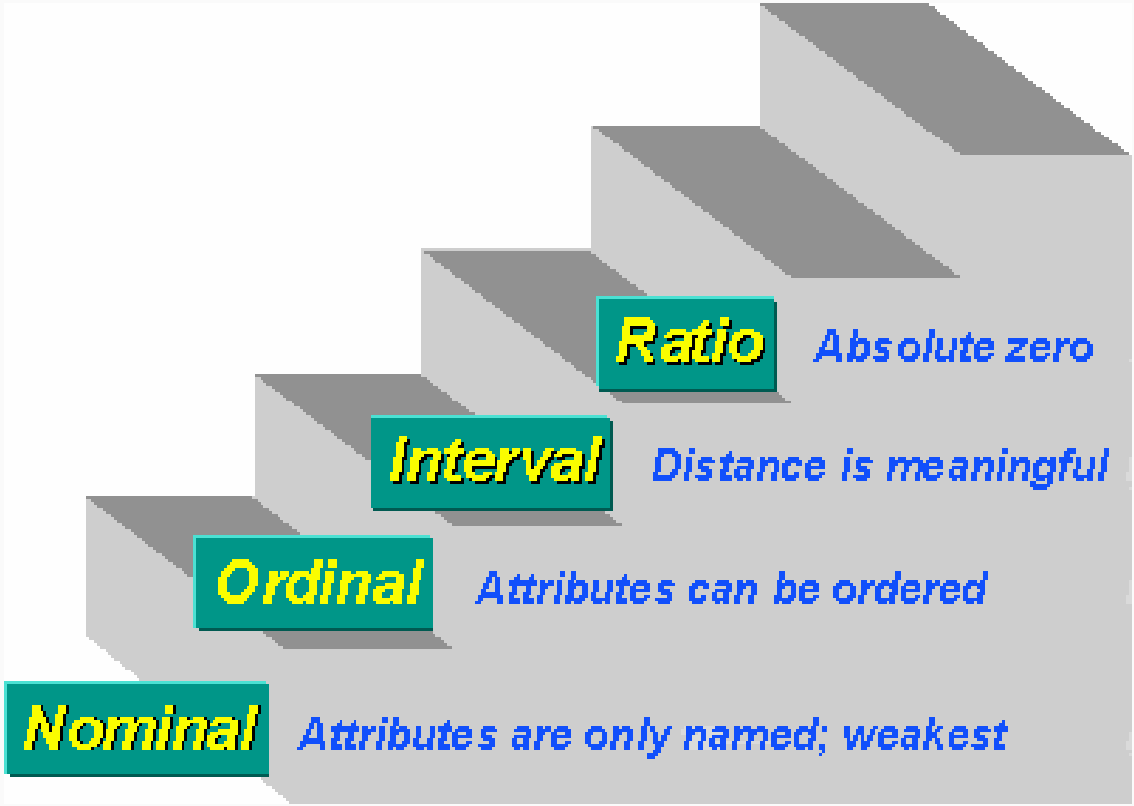
Definitions and Reliability

Measurement

- Measurement is a systematic, replicable process by which objects or events are **quantified** and/or **classified** with respect to a particular dimension
- This is usually achieved by the assignment of numerical values

Levels of Measurement

1. Nominal measure
2. Ordinal measure
3. Interval measure
4. Ratio measure



Reliability of a Measure

- The degree to which a measurement technique can be depended upon to secure consistent results upon repeated application
  - "The rubber ruler issue"

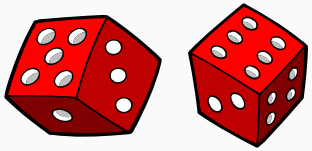
Validity of a Measure

- The degree to which any measurement approach or instrument succeeds in describing or quantifying what it is designed to measure
  - "The 35-inch yardstick issue"

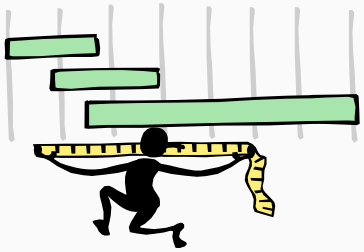
So, Variation in a Repeated Measure

- Can be due to the following reasons:

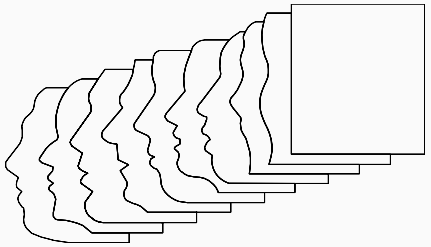
1. Chance or unsystematic events



2. Systematic inconsistency



3. Actual change in the underlying event being measured



Measurement Reliability

- Not just the property of an instrument
- Rather, a measure or instrument has a certain degree of reliability when applied to certain populations under certain conditions

Sources of "Unsystematic" Threats to Reliability

1. Subject reliability—factors due to research subject (e.g., patient fatigue, mood)
2. Observer reliability—factors due to observer/ rater/interviewer (e.g., abilities of interviewer, different opinions)
3. Situational reliability—conditions under which measurements are made (e.g., busy day at the clinic, new management)

Source: Suchman

"Unsystematic" Threats to Reliability

4. Instrument reliability—the research instrument or measurement approach itself (e.g., poorly worded questions, quirk in mechanical device)
5. Data processing reliability—manner in which data are handled (e.g., miscoding)

How Do We Evaluate Observer Measurement Reliability?

- Inter-rater agreement
  - Compare two or more of the observers/raters at a point in time
  - Percentage of overall agreement, Kappa
- Test-retest
  - Compare measurements made by the same observer/rater at two points in time
  - Timeframe should be short enough that the construct itself hasn't changed

Calculating Kappa

Kappa = Observed Agreement – Expected Agreement due to Chance

kappa = (Oa – Ea) / (N - Ea), where

Ea = sum of expected counts in cells A and D:

$[(N_1 * N_3)/N] + [(N_2 * N_4)/N]$ (rounded to nearest whole number)

Oa = sum of observed count in cells A and D

and N is the total number of respondent pairs

	Rater 1 Yes	Rater 1 No	
Rater 2 Yes	A	B	N_3= A+B
Rater 2 No	C	D	N_4= C+D
	N_1= A+C	N_2= B+D	N

Criteria for Interpreting Kappa Statistics

Level of Agreement	Kappa Value
Almost Perfect	0.81-1.00
Substantial	0.61-0.80
Moderate	0.41-0.60
Fair	0.21-0.40
Slight	0.00-0.20
Poor	<0.00

Source: Landis, J.R., Koch, G.G. (1977). The measurement of observer agreement for categorical data. *Biometrics*. 33:159.

JOHNS HOPKINS
BLOOMBERG
SCHOOL of PUBLIC HEALTH

Section B

Measurement: Reliablilty and Validity of Measures

How Do We Evaluate Instrument Reliability?

- General "congruence" of instrument/ questionnaire (at same point in time)
  - Item-total correlation
  - Internal consistency—the extent to which the items in a scale “hang together” (Cronbach’s coefficient or “alpha” statistic)

How Do We Evaluate Instrument Reliability?

- General "congruence" of instrument/ questionnaire (at same point in time)
  - Item-total correlation
  - Internal consistency—the extent to which the items in a scale “hang together” (Cronbach’s coefficient or “alpha” statistic)
- Approaches developers use to assess / improve congruence and efficiency
  - Split half (split questionnaire into two parts)
  - Short form-long form (two different lengths)
- Test-Retest
  - Similar to observer re-test

Calculating Cronbach's Alpha

Reliability Measure for Multi-Item Scales

- Total scale variance = sum of item variances and all item covariances
- **[k/(k-1)]* [1- (sum of item variances/total scale variance)**
  - Where k = number of items
- Range between 0 and 1
- Criteria for assessment
  - ≥ 0.70 = adequate reliability for group comparisons
  - ≥ 0.90 = adequate reliability for individual monitoring

Relationship between Reliability and Validity

- They are closely inter-dependent
- There **can not** be validity without reliability
- There **can** be reliability without validity

Measurement Validity (Recap)

- Validity of a measure
  - The degree to which any measurement approach or instrument succeeds in describing or quantifying what it is designed to measure
  - Validity reflects those errors in measurement that are systematic or constant

The Term "Validity" Has More than One Connotation

- The general concept of "validity" is broader than just "validity of approaches to measurement"
- In general, measurement reliability and validity issues fall into Campbell and Stanley's "instrumentation" category

How to Evaluate Measurement Validity

- **Face Validity**
  - Measurement is accepted by those concerned as being logical on the "face of it" (also expert validity)
- **Content Validity**
  - Do the items included in the measure adequately represent the universe of questions that could have been asked?

How to Evaluate Measurement Validity

Criterion-Related Validity

- Does the new measure agree with an external criterion, e.g., an accepted measure?
- Predictive evidence
  - Predictive of future event or outcome of interest
- Concurrent evidence
  - Correlation with "gold standard" at the same point in time
  - Shortened scale with full scale

How to Evaluate Measurement Validity

- Construct validity—is the measure consistent with the theoretical concept being measured?
  - All tests of validity ultimately designed to support/refute the instrument's construct validity
  - Construct validity never fully established

Assessing Construct Validity

- Convergent evidence
  - Demonstrate that your measure correlates highly (.5-.7) with measures of the same construct
  - Groups known to differ along construct have significantly different scores on measure
- Discriminant evidence
  - Low correlation with instruments measuring a different construct; or differences between known groups
- Factorial evidence
  - Clustering of items supports the theory-based grouping of items

Some Advanced Measurement Terms/Topics

- Field of "psychometrics" is advancing
- Well developed "scales" and "composite indices" are available for most measures
  - For example, Short-Form (SF) 36, Euro-QoL
  - Many advantages to well standardized robust multi-item measures
- Supported by computer generated questions in the field of "Item Response Theory"
- IRT is gaining popularity (also known as "adaptive" testing)
  - If a patient can walk a mile, why ask if they can walk 100 yards?

Practical Application of Measurement Reliability and Validity—Factors to Consider

- How much time and money do you have to carry out your own tests?
- How small a difference in the measurement do you expect?
- Can you use a previously validated measure?
- Does the previous measure work within the context of your setting?