This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike License. Your use of this material constitutes acceptance of that license and the conditions of use of materials on this site.
Measurement: Reliability and Validity Measures

Jonathan Weiner, DrPH
Johns Hopkins University
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”
Measurement Validity

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)
“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

\[
\text{kappa} = \frac{\text{Oa} - \text{Ea}}{\text{N} - \text{Ea}}, \text{ where}
\]

Ea = sum of expected counts in cells A and D:
\[
\frac{(N_1 \times N_3)}{N} + \frac{(N_2 \times N_4)}{N} \text{ (rounded to nearest whole number)}
\]

Oa = sum of observed count in cells A and D
and N is the total number of respondent pairs

<table>
<thead>
<tr>
<th></th>
<th>Rater 1 Yes</th>
<th>Rater 1 No</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Rater 2 Yes</td>
<td>A</td>
<td>B</td>
<td>(N_3 = A + B)</td>
</tr>
<tr>
<td>Rater 2 No</td>
<td>C</td>
<td>D</td>
<td>(N_4 = C + D)</td>
</tr>
<tr>
<td></td>
<td>(N_1 = A + C)</td>
<td>(N_2 = B + D)</td>
<td>(N)</td>
</tr>
</tbody>
</table>
## Criteria for Interpreting Kappa Statistics

<table>
<thead>
<tr>
<th>Level of Agreement</th>
<th>Kappa Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Almost Perfect</td>
<td>0.81-1.00</td>
</tr>
<tr>
<td>Substantial</td>
<td>0.61-0.80</td>
</tr>
<tr>
<td>Moderate</td>
<td>0.41-0.60</td>
</tr>
<tr>
<td>Fair</td>
<td>0.21-0.40</td>
</tr>
<tr>
<td>Slight</td>
<td>0.00-0.20</td>
</tr>
<tr>
<td>Poor</td>
<td>&lt;0.00</td>
</tr>
</tbody>
</table>

Section B

Measurement: Reliability 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
- \[ \frac{k}{(k-1)} \times [1 - \text{(sum of item variances/total scale variance)}] \]
  - Where \( k \) = number of items
- Range between 0 and 1
- Criteria for assessment
  - \( \geq 0.70 \) = adequate reliability for group comparisons
  - \( \geq 0.90 \) = adequate reliability for individual monitoring
They are closely inter-dependent

- There **can not** be validity without reliability
- There **can** be reliability without validity
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 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 (0.5-0.7) 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?