Chapter 4

Reliability and Validity
Chapter 4: Reliability and Validity

• Reliability implies consistency. Basically, if the test is repeated, we would probably get the same results.

• Validity implies accuracy. Basically, the used measurement measures exactly what it is supposed to measure.

• Reliability does not imply validity. That is, a reliable measure is measuring something consistently, but you may not be measuring what you want to be measuring.

• Validity implies reliability but, again, not vice versa.
Reliability: Conceptualization

• **Three forms:**
  
  – To what extent does a person’s (or an object’s) measured performance remain the same across repeated testings?
  
  – To what extent do the individual items that go together to make up a test consistently measure the same underlying characteristics?
  
  – How much consistency exists among the ratings provided by a group of raters?

• How to measure reliability?
  
  – Reliability coefficient (0-1).
Different Approaches to Reliability

• Test-Retest Reliability
  – Doing the test twice using the same instrument over two points of time (can be long or short).
  – The reliability coefficient is called \textit{coefficient of stability}.
  – Usually Pearson’s correlation ($p$) or something called \textit{intraclass coefficient} is used.
  – If the coefficient is high with longer times separating the two tests, the reliability is more impressive.
Different Approaches to Reliability, cont’d

• Equivalent-Forms Reliability

  – Using two forms of the same instrument (not time)

  – e.g. testing participant’s intelligence using two forms of questions that target the same answers.

  – The two sets of answers (scores) are then correlated.

  – The correlation coefficient is called coefficient of equivalence.

  – Pearson’s $r$ is widely used for correlation.
Different Approaches to Reliability, cont’d

• **Internal consistency reliability**
  – The degree to which the measuring instruments possess internal consistency.
  – No time, no different forms.
  – Consistency across parts of measurements.
  – Parts can be questions/sets of questions.
Different Forms to Measure Internal Consistency
Reliability

1. **Split-half reliability coefficient.**
   - Splitting scores by participant numbers (odd. vs. even) and then applying correlation tests.
   - The final numerical result of correlation is called: *split-half reliability coefficient*.

2. **Richardson #20**
2. Richardson #20 (aka K-R-20)

- Test one time only.
- Favorable to split-half because no odd vs. even numbers are considered.
- The formula produces the value.
- A simple variant formula is: \( KR21 = \frac{N/(N-1)}{[1 - \{M(N-M)\}]}/(N*V) \)
  - Where:
    - \( N \) = number of items in the test
    - \( M \) = arithmetic mean of the test scores
    - \( V \) = variance of the raw scores
- May be affected by difficulty of the test, the spread in scores and the length of the examination.
- Used with dichotomous measures such as gender and snoring (not continuous).
- Can be thought of as doing split-half multiple times with averaging over the resulted values.
Different Forms to Measure Internal Consistency Reliability, cont’d

• **Cronbach’s alpha (aka alpha)**
  – If the scores are dichotomous, it is the same as K-R-20.
  – More versatile: can handle three or more answers per variable.
  – Likert type scale answers is a good example.
Interrater Reliability

• The degree of consistency among the raters.
• Five known procedures:
  – Percent agreement measure
  – Pearson’s correlation
  – Kendall’s coefficient of concordance
  – Cohen’s Kappa
  – Interclass correlation
Percent Agreement Measure and Pearson’s Correlation

• Percent Agreement Measure
  – A simple percentage of agreement
  – Can be used with categorical data, ranks, or raw scores.

• Pearson’s
  – Can be used only when the raters’ ratings are raw scores.
Kendall’s and kappa

• **Kendall’s**
  – Can be any value between 0 and 1.
  – Data has to be ranked.

• **Cohen’s kappa**
  – The same as Kendall’s except that the data are nominal (i.e., categorical) with kappa.
Kappa Example

- Kappa measure
  - Agreement measure among judges
  - Corrects for chance agreement
- Kappa = \[
  \frac{P(A) - P(E)}{1 - P(E)}
\]
- \( P(A) \) – proportion of time judges agree
- \( P(E) \) – what agreement would be by chance
- Kappa = 0 for chance agreement, 1 for total agreement.
Kappa Measure: Example
Total Number of Documents = 400

<table>
<thead>
<tr>
<th>Number of docs</th>
<th>Judge 1</th>
<th>Judge 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>300</td>
<td>Relevant</td>
<td>Relevant</td>
</tr>
<tr>
<td>70</td>
<td>Nonrelevant</td>
<td>Nonrelevant</td>
</tr>
<tr>
<td>20</td>
<td>Relevant</td>
<td>Non-relevant</td>
</tr>
<tr>
<td>10</td>
<td>Nonrelevant</td>
<td>Relevant</td>
</tr>
</tbody>
</table>
Example, cont’d

• $P(A) = (300+300+70+70) / 800 = 0.925$
• $P(\text{non-relevant}) = (10+20+70+70)/800 = 0.2125$
• $P(\text{relevant}) = (10+20+300+300)/800 = 0.7878$
• $P(E) = 0.2125^2 + 0.7878^2 = 0.665$
• Kappa = $(0.925 - 0.665)/(1-0.665) = 0.776$

• Kappa > 0.8 = good agreement
• $0.67 < \text{Kappa} < 0.8$ -> “tentative conclusions” (Carletta ’96)
• Depends on purpose of study
• For >2 judges (Raters): average pairwise kappas
Interclass Correlation (ICC)

• Can be used for correlation or reliability purposes.
• ICC is associated with two numbers:
  - The first is the statistical model assumed by the researcher to underlie their data
  - The second number if:
    - 0: One rater
    - 1: Reliability of means for more than one rater
    - >1: How many scores are averaged together to generate each mean.
The standard Error of Measurement (SEM)

- Used to estimate the range within which a score would likely fall in the case of a re-measurement.

- Example:
  - IQ score of Jack is 112.
  - SEM = 4
  - Interval → 108-116, aka confidence band.
  - If retested, Jack would probably score somewhere between 108 and 116.
Warnings!

- High level of stability does not mean high consistency!!!
- Reliability coefficients apply to data not measurements.
- Any reliability coefficient is an estimate of consistency.
  - Use of similar raters may result in different reliability coefficients recorded.
- Strict time limits between tests may result in high reliability coefficients. Do not be impressed because it is the time pressure, not actually consistency.
- Reliability is not the only important criterion to assess the quality of the data.
Validity

- The results of the measurement process are accurate= the results are valid.
- A measurement instrument measures what it is supposed to measure, then it is valid.
- **Relationship between reliability and validity**
  - Data of a study can be reliable
  - Does not necessarily mean it is valid.
  - An instrument's data must be reliable if they are valid.
  - Accuracy requires consistency
Different Kinds of Validity

• **Content Validity**
  – The instrument covers the material.
  – By comparing test content to a syllabus or outline by experts.

• **Criterion-Related Validity**
  – Comparing the scores of a new test to a previous test (criterion) to measure validity.
  – Then the two test scores are correlated.
  – The resulting $r$ is called the validity coefficient.

  – **Concurrent vs. predictive validity**
    • Has to do with issue of testing time ($s$) (how far apart the criterion is from the new instrument)
Different Kinds of Validity, cont’d

• Construct validity
  – Associated with the instrument

  – *Construct validity is the extent to which a test measures the concept or construct that it is intended to measure.*
    (<http://www.psychologyandsociety.com/constructvalidity.html>)

• Three approaches:
  – Correlation evidence showing that the construct has a strong relationship with certain variables and weak relationship with other variables.
  – Show that a certain groups obtains higher mean scores than other groups on the new instrument.
  – Conduct a factor analysis.
  – Always remember to use $r$. 
Construct Validity Example

There are many possible examples of construct validity. For example, imagine that you were interested in developing a measure of perceived meaning in life. You could develop a questionnaire with questions that are intended to measure perceived meaning in life. If your measure of perceived meaning in life has high construct validity then higher scores on the measure should reflect greater perceived meaning in life. In order to demonstrate construct validity, you may collect evidence that your measure of perceived meaning in life is highly correlated with other measures of the same concept or construct. Moreover, people who perceive their lives as highly meaningful should score higher on your measure than people who perceive their lives as low in meaning.
Warnings!

• Validity coefficients (in the case of construct and criterion-related) are estimates only.

• Validity is a characteristic of the data produced by the instrument but not the instrument itself.

• When measuring content validity, experts must have:
  – The technical expertise needed.
  – The willingness to provide negative feedback to the developer.
  – Describe your experts in details when reporting on this kind of validity.