What is reliability for a test?

http://www.youtube.com/watch?v=kDW0Y5msB0A
Test Reliability Recap

• Definition

Consistency of measurements when the testing procedure is repeated on a population of individuals or groups (Standards for Educational and Psychological Testing, 1999).
Type

• Test-retest reliability
• Parallel-forms reliability
• Internal consistency reliability (Split-half)
Conceptualization of $r_{xx'} = 0.85$

Meaningful Variance 85%

Error 15%

You may use these materials only for nonprofit educational purposes. Please give credit/cite appropriately.
GROUP DISCUSSION

What can be sources of error?
Group Work Instructions
(Facilitation Technique: Typology Development Adapted)

1. Each one write three sources of error.
2. Write one source on one piece of post-it provided.
3. Post the completed post-its on the wall.
4. Read what others have written.
5. Group similar ones together.
6. Label the ideas.
CHECKLIST FOR POTENTIAL SOURCES OF ERROR

Handout
Testing Procedure

1. Identify Target Outcomes
2. Identify knowledge/skill to test
3. Test Construction
4. Test takers take the test
5. Key/Rubric Construction
6. Score or rate the answers or student production
Ranges of Reliability

0
Not reliable at all

Low-stake

0.80

0.90
High-stake

1
Totally reliable

You may use these materials only for nonprofit educational purposes. Please give credit/cite appropriately.
Calculation of reliability

- Test-retest reliability – correlation
- Parallel-forms reliability – correlation
- Internal consistency reliability (Split-half) – correlation adjusted for length

Spearman-Brown Prophecy \( r_{xx'} = \frac{n \times r}{(n - 1) \times r + 1} \)

- \( r_{xx'} = \) full-test reliability
- \( r = \) correlation
- \( n = \) the number of times the test length is to be increased

You may use these materials only for nonprofit educational purposes. Please give credit/cite appropriately.
EXCEL ACTIVITY

DEMONSTRATION + HANDS-ON

Please download the reliability calculation Excel practice file.
JUDGMENT RELIABILITY

- Interrater agreement
- Interrater reliability
- Intrarater reliability
### Rating Results of Student Essays

0 = Failing  
1 = Approaching  
2 = Meeting  
3 = Exceeding

<table>
<thead>
<tr>
<th>StudentID</th>
<th>Rater1</th>
<th>Rater2</th>
</tr>
</thead>
<tbody>
<tr>
<td>101</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>102</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>103</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Interrater Agreement = \( \frac{Pairs \ of \ same \ ratings}{All \ Pairs} \) = 0.33

You may use these materials only for nonprofit educational purposes. Please give credit/cite appropriately.
Interrater Reliability

• Consistency between raters
• Like parallel forms: The correlation between two raters
• If an aggregated score is used \( \rightarrow \) correlation adjusted for number of raters, like split-half

Spearman-Brown Prophecy

\[
r_{xx'} = \frac{n \cdot r}{(n - 1) \cdot r + 1}
\]

• \( r_{xx'} = \text{full-test reliability} \)
• \( r = \text{correlation} \)
• \( n = \text{the number of raters} \)
Intrarater Reliability

• Like test-retest reliability
• Consistency of judgment over time.
• Correlation
• Be aware of memory effect
ReCal

- Reliability Calculation for the masses
- URL: [http://dfreelon.org/utils/recalfront/](http://dfreelon.org/utils/recalfront/)

<table>
<thead>
<tr>
<th>Level of measurement</th>
<th>N of coders</th>
<th>Use</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nominal</td>
<td>2 coders only</td>
<td>ReCal2 (includes percent agreement, Scott’s pi, Cohen’s kappa, and nominal Krippendorff’s alpha)</td>
</tr>
<tr>
<td>Nominal</td>
<td>3 or more coders</td>
<td>ReCal3 (includes pairwise percent agreement, Fleiss’ kappa, pairwise Cohen’s kappa, and nominal Krippendorff’s alpha)</td>
</tr>
<tr>
<td>Ordinal, interval, or ratio</td>
<td>Any N of coders</td>
<td>ReCal OIR (includes ordinal, interval, and ratio Krippendorff’s alpha)</td>
</tr>
</tbody>
</table>

You may use these materials only for nonprofit educational purposes. Please give credit/cite appropriately.
### Sample file format

![Sample file format](image)

- **R1_SLO1**
- **R2_SLO1**
- **R1_SLO2**
- **R2_SLO2**

Save as comma delimited file

You may use these materials only for nonprofit educational purposes. Please give credit/cite appropriately.
HANDS-ON DEMONSTRATION
STRATEGIES TO IMPROVE RELIABILITY

Group Activity:
1. Install Socrative Student on your tablet or your phone. Or on your computer go to: m.socrative.com/student/#joinRoom
2. Answer the short-answer question.
Checklist for Sources of Error (Brown, 1996, p. 189)

☐ Variation due to environment
  o Location
  o Space
  o Ventilation
  o Noise
  o Lighting
  o Weather

☐ Variance due to administration procedures
  o Directions
  o Equipment
  o Timing
  o Mechanics

☐ Variance attributable to examinees
  o Health fatigue
  o Physical characteristics
  o Motivation
  o Emotion
  o Memory
  o Concentration
  o Forgetfulness
  o Impulsiveness
  o Carelessness
  o Test wiseness
  o Comprehension of directions
  o Guessing
  o Task performance speed
  o Chance knowledge of item content

☐ Variance due to scoring procedures
  o Errors in scoring
  o Subjectivity
  o Evaluator biases
  o Evaluator idiosyncrasies

☐ Variance attributable to the test and test items
  o Test booklet clarity
  o Answer sheet format
  o Particular sample of items
  o Item types number of items
  o Item quality
  o Test security

Typology Development


Purpose: to reach consensus of categories and categorical components of ideas/solutions/problems/strategies.

Time: 1 - 2 hours for 5 to 10 participants

Procedure:

1. Give each team member a large, blank Post-it(R) note pad and record as many ideas/solutions/problems/strategies as possible in 3 minutes. Each ideas/solutions/problems/strategies was recorded on a single Post-it(R) note.
2. Ask each participant to pass their note pad to one person on the left. They are directed to review the ideas/solutions/problems/strategies that are passed to them and they are given additional three minutes to add more to the list.
3. Repeat the process until the note pads had rotated through all participants and were returned to the original author.
4. The facilitator will then post each single note on a large white board. Participants are asked to eliminate duplication by placing redundant items on top of one another.
5. After consensus was achieved on these items, participants were then asked to cluster items that they perceived to be conceptually similar by physically moving Post-it(r) notes around the wall until all items have been classified.
6. Elicit consensus to label each cluster.
7. At the end of the session, the facilitator will carefully package the material for later transcription.