Reliability

PSY 395 - Oswald

Outline

- Concept of Reliability
- What are Constructs?
- Construct Contamination and Construct Deficiency
- Reliability in General
- Classical Test Theory
- Reliability Coefficients
- Internal Consistency Reliability

Reliability

- Cars (engines, brakes!)
- Friends (on time, secrets)
- Surgeons
- Tattoo Artists
- Oh, and psychological measures
Reliability Affects Decision-Making in Psychology

- Selecting airline pilots
- Placing a child into Head Start
- Determining whether a client needs treatment for depression

Constructs

Often are unseen but can be measured indirectly (often via many multiple-choice questions):
- Airline pilots $\rightarrow$ conscientiousness
- Head Start $\rightarrow$ language skills
- Clinical treatment $\rightarrow$ depression

Endorsing items related to a construct doesn’t mean you have a high standing on the construct – you must endorse many items

Example: Conscientiousness – what are some possible items?
Sample measure:
“Lack of Culture Fit”
(adapted from Lynness & Thompson, 2000, Journal of Applied Psychology)

1. I feel pressure to fit in or adapt to the situation around me.
2. I feel like I am an outsider.
3. There are few role models I can turn to.
4. I feel like I’m held to a higher standard than others.
5. I like food.
Standards of Measurement

Given that the measure of a construct is not contaminated and not deficient then we concern ourselves with:

✓ Reliability
• Validity (a measure can not be valid if it is not reliable!)

Reliability

• How consistent are the observed scores obtained by the same person tested at different time points?

• To what extent does your observed score reflect your true score on the construct measured?

Classical Test Theory (CTT)

\[ X = T + E \]

where
\[ X = \text{observed score} \]
\[ T = \text{true score} \]
\[ E = \text{random error score} \]

[the scores you have]
[construct score \(\rightarrow\) anything consistent]
[guessing, fatigue \(\rightarrow\) anything inconsistent]
CTT Assumptions

1. True scores and errors are uncorrelated (independent)
2. Errors across people average to zero
3. Across repeated measurements, a person’s average score is ≈ equal to his/her true score (i.e., #2 above also applies within people)

Classical Test Theory

\[
X_{1a} = T_{1a} + E_{1a} \\
X_{1b} = T_{1b} + E_{1b} \\
\ldots \\
X_{1N} = T_{1N} + E_{1N}
\]

Error distributions
- Across people within measures
- (and within people across measures…)

Assumption
- If different measures (\(X_1, X_2, X_3\)) measure the same construct, then for a particular person the measures should have the same true score

Reliability

If \(X = T + E\), then – since true scores and error scores are uncorrelated:
\[
\text{var}(X) = \text{var}(T) + \text{var}(E)
\]

Question: What percent of the observed variance is reliable (true-score) variance??

Classical Test Theory
Reliability Coefficients

Reliability coefficients reflect the proportion of true score variance to observed score variance

\[ r_{xx} = \frac{\text{var}(T)}{\text{var}(X)} \]

Therefore reliabilities range from 0.0 (all error variance) to 1.0 (all true-score variance)

All Indexes of Reliability are Correlation Coefficients

\[ r_{xx} = .87 \]
\[ r_{xx} = .23 \]

Errors in Measurement

What is “error”? Depends on how you define it; what variance you decide is irrelevant.
Errors in Measurement

Some questions to ask:
- Is the construct stable?
- Are there multiple dimensions to the construct? (e.g., leadership)
- Will test-takers be fatigued?
- Are there practical considerations (is test-retest possible? how are scores being used?)

Conceptual Considerations in Measurement

- Type of Test
  - Speeded
  - Power
- Type of Group Tested
  - Range-restricted?
  - Generalizable?

Different Reliability Coefficients

How you estimate reliability depends on what you consider to be error

<table>
<thead>
<tr>
<th>Source of Error</th>
<th>Type of Reliability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time</td>
<td>Test-retest</td>
</tr>
<tr>
<td>Content</td>
<td>Parallel forms</td>
</tr>
<tr>
<td></td>
<td>Split-half</td>
</tr>
<tr>
<td></td>
<td>Cronbach's Alpha</td>
</tr>
</tbody>
</table>
Different Reliability Coefficients

Hey, there’s more!

<table>
<thead>
<tr>
<th>Source of Error</th>
<th>Type of Reliability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time and Content</td>
<td>Parallel forms with time delay</td>
</tr>
<tr>
<td>Raters/Observers</td>
<td>Interrater correlation</td>
</tr>
</tbody>
</table>

Test-Retest Reliability

What is it?

The correlation between the scores of people who have taken a test twice

- Why is this useful?
- What are some possible sources of error variance?
  - practice, fatigue, health, motivation

Parallel (or Alternate) Forms Reliability

- What is it?
- Same vs. different points in time
- Practical considerations (time/$)
- Counterbalancing
Split-Half Reliability

What is it?

odd/even; first/last

What are the sources of error?
(content)

What are NOT the sources of error?
(time)

Correcting for Split-Half:
Spearman-Brown (SB) Formula

\[ r_{SB} = \frac{k \cdot r_{xx}}{1 + (k - 1) r_{xx}} \]

Correction:
\[ r_{SB} > r_{xx} \]

\( k \) = amount by which the test is multiplied
(here, it's 2 – why?)
\( r_{xx} \) = correlation between halves

EXAMPLE

• Example: 100 items, split-half \( r = .70 \), what's the estimated reliability of the full test?
• \( SB = (2 \times .7) / (1 + (2 - 1) \times .70) = 1.4/1.7 = .82 \)

\[ r_{SB} = \frac{k \cdot r_{xx}}{1 + (k - 1) r_{xx}} = \frac{2 \cdot .70}{1 + (2 - 1) \cdot .70} = .82 \]

• (note the increase)
Coefficient Alpha (\(\alpha\))

What is it?
(1) average inter-item correlation boosted by the SB formula (where \(k = \#\) of items)

Another way to say this is
(2) average of all possible split-half correlations boosted by the SB formula (where \(k = 2\))

What are the sources of error?
(content)
What are NOT the sources of error?
(time)

\[
\alpha = \frac{k \cdot \overline{r_{ij}}}{1 + (k - 1)\overline{r_{ij}}}
\]

Where you need (a) the \# of items and (b) the average inter-item correlation
Internal Consistency Reliability

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

\[ \alpha = \frac{k \cdot r_{ij}^2}{1 + (k-1)r_{ij}^2} \]

---

**Example: Correlations**

<table>
<thead>
<tr>
<th>item</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.00</td>
<td>.33</td>
<td>.33</td>
<td>.50</td>
</tr>
<tr>
<td>2</td>
<td>.33</td>
<td>1.00</td>
<td>.50</td>
<td>.33</td>
</tr>
<tr>
<td>3</td>
<td>.33</td>
<td>.50</td>
<td>1.00</td>
<td>.25</td>
</tr>
<tr>
<td>4</td>
<td>.50</td>
<td>.33</td>
<td>.25</td>
<td>1.00</td>
</tr>
</tbody>
</table>

\[ \alpha = \frac{k \cdot r_{ij}^2}{1 + (k-1)r_{ij}^2} = \frac{4 \cdot .373}{1 + (4-1)\cdot .373} = .70 \]

OR \[ \alpha = \frac{k^2 \cdot r_{ij}^2}{R} = \frac{(4^2 \cdot .373)}{8.48} = .70 \]

*Note: the average uses the “lower diagonal” not including the 1,1 values*

---

**SB and \( \alpha \) : Assumptions**

The formulas are not magical corrections that solve real measurement problems, if the test needs to be developed/revised further.

A couple of assumptions of both formulas

1. all items are equal statistically
2. a longer test will generally have higher reliability because it is assumed to be the same in terms of content as its shorter counterpart