Reliability

PSY 600K

Dr. Alyssa Gibbons

February 15, 2018

Agenda

- Items due today!
- A theoretical view and formal definition of reliability
  - Partitioning error variance.
  - Why we can do this with a correlation!
- Types of error / types of reliability:
  - Test-retest
  - Alternate forms
  - Split-half

Reliability

- What do you think of when you hear the word "reliability"?
  - In or out of a psychological context…
- Common answer: stability.
  - But this really only addresses one type of reliability.
- DeVellis:
  - "performs in consistent, predictable ways"
  - "a perfectly reliable scale would be a reflection of the true score and nothing else." (p. 31)
- McDonald (1999): "estimating the precision of measurement of a test score."

More Formally

- The reliability coefficient for a test is **defined** as the **proportion of total test variance that is due to true score.**
  - Y = T + E
  - $\sigma_Y^2 = \sigma_T^2 + \sigma_E^2$
    - Variances of uncorrelated variables add – see R & M 2.5
  - So $\rho_{XX} = \frac{\sigma_T^2}{\sigma_Y^2} = \frac{\sigma_T^2}{\sigma_T^2+\sigma_E^2}$
  - If $\rho_{XX} = .80$, we say that 80% of the variance in Y is due to variance in true scores.
    - The remaining 20% is error.
  - It makes sense to interpret this as an index of precision.

Estimating Reliability

- How can we separate $\sigma_T^2 + \sigma_E^2$ when only σ_Y^2 is known?
- Note that ρ_{XX} looks like a correlation… that's not a coincidence.
- Many methods of estimating reliability are based on correlations.

Separating True & Error Variances

- Imagine that you are giving the same test twice:
  - $Y = T + E$
  - $Y' = T + E'$
- Note that the true score stays the same, but the error changes.
  - So what kind of error is this? 💬
- We expect that the *variance* of the error terms is the same.
  - $\sigma_Y^2 = \sigma_T^2 + \sigma_E^2$
  - $\sigma_{Y'}^2 = \sigma_T^2 + \sigma_{E'}^2$
- So if $\sigma_E^2 = \sigma_{E'}^2$, then $\sigma_Y^2 = \sigma_{Y'}^2$

More…

- The correlation between Y and Y' is their covariance divided by the square root of the pooled variance.
- The covariance between Y and Y' =
  - Cov[(T + E)(T + E')]
  - $= \sigma_{TT} + \sigma_{TE} + \sigma_{TE'} + \sigma_{EE'}$
  - $= \sigma_{TT}$
- So the correlation between Y and Y' =
  - $= \sigma^2_T / \mathrm{sqrt}\{\sigma^2_Y \sigma^2_{Y'}\}$
  - $= \sigma^2_T / \sigma^2_Y$
- And that's our reliability coefficient. Correlating a test with "itself" (same T, same σ_E^2) yields an estimate of true score variance.

Correlation-Based Estimates of Reliability

- We can correlate a test with "itself" in several ways:
  - The test with the exact same test at a later point in time.
  - Two versions of the test designed to be parallel.
    - Identical structure; not identical items.
    - Measuring the same construct (T) and with the same error variance.
    - Not necessarily *strictly* parallel in the sense of the MPT.
  - Two raters using the same system (items) to rate the same targets.
  - Two halves of the same test.
- However, the "error" variance we are estimating is a little different in each case.

Test-Retest Reliability

- Measures error due to **time**.
  - Everything else – persons, items – is constant.
- Requires the assumption that true score does not change between tests…
- OR we can interpret the coefficient as the part of true score that does not change.
  - “Coefficient of stability”

Issues in Test-Retest Reliability

- What is the appropriate retest interval?
  - Usually we see that correlations between measurements decrease as they get farther apart in time – simplex model:

	Time 1	Time 2	Time 3	Time 4
Time 1	1.0			
Time 2	.90	1.0		
Time 3	.80	.90	1.0	
Time 4	.70	.80	.90	1.0

  - But then what does the correlation between any 2 measurements really mean?
    - That we picked the right (or wrong) interval?
- If we really want to study stability, a single test-retest coefficient doesn't tell us all that much.

Alternate Forms Reliability

- When is it useful to have two forms of the same test?
- Comparing two alternate forms tells us **nothing** about error or change due to time.
  - It **does** tell us about error due to differences in the forms.
  - “Coefficient of equivalence” – the part of the variance that is common across the two tests.
- Interrater reliability can be evaluated similarly – we view 2 raters as 2 “alternate forms.”
  - Of course, reliability ≠ agreement… but that’s another story.

Split Half Reliability

- Sometimes you only have one test!
  - No alternate forms, no retest.
  - How can you estimate reliability?
- Early solution: split the test in half and treat it as two alternate forms.
  - Can split randomly, theoretically, or for convenience (e.g., odd/even).
- The trouble is, this will **underestimate** the reliability of your test.
- Reliability **increases** as the number of items increases.
  - Why?
  - So by cutting your test in half, you've reduced your reliability.

The Spearman-Brown Prophecy Formula

- Estimates what the reliability of your test *would* be if you had more items.
  - Assuming your new items are equally as good as your initial items…
- Formula:
  - $\frac{n(\rho_{XX})}{1+(n-1)(\rho_{XX})}$ $\quad n = \frac{\text{\# items in final version}}{\text{\# items in the original}}$
- For split-half reliability, $n = 2$, but we can also use this to estimate:
  - What the reliability of our test would be if we added some more items.
  - How many more items we need to reach a certain level of reliability.
- What do you think of this? Is this shady?

More About Split-Half Reliability

- Split-half reliability can be considered a measure of *internal consistency reliability.*
  - Estimates error from the particular items we chose.
- We seldom use split-half methods anymore as a *primary* indicator of reliability.
  - Some reviewers, however, still ask for it.
- Split-half reliability has largely been replaced by… ??

- Cronbach’s alpha.

Questions?

For next time:

Alpha

Read: R & M 6.2 – 6.4 and 7.2 – 7.4;

Also skim Schmidt & Hunter (1996) – read a few of their scenarios.

Reading Response #5