Reliability - Applications

Measurement Methods

Lecture 13

Chapter 7

Today's Class

- Applications of Reliability Theory
  - Test length and reliability
  - Estimation of a true score and the error of the estimate
  - Estimation relationships between true scores on multiple tests
    - Attenuation

Upcoming Schedule

Our Upcoming Schedule

Note: THE CLASS PROJECT IS DUE ON 5/16

Date	Chapter and Topic
3/30	Chapter 7: Reliability – Applications
4/4	Computer Lab session
4/6,4/11	No Class
4/13	Chapter 8: Prediction and Multiple Regression
4/18, 4/20	Chapter 9: The Common Factor Model
4/25	Chapter 10: Validity
4/27	Chapter 11: Classical Item Analysis
5/2	Chapter 12: Item Response Models
5/4	Chapter 13: Properties of Item Response Models
5/9	Chapter 14: Multidimensional Item Response Models
5/11	Review
5/16	Final Exam (1:30pm – 4:00pm)

Test Length and Reliability

Test Development

- Typically, a large set of items are used as an initial attempt to develop a test.
  - From that large set, a smaller set of items are selected to be used in the actual instrument.
- The number of items should be kept small enough to have tests administered in a practical amount of time.
- The number of items should be kept large enough to produce an acceptable level of reliability.

How to Choose Items To Maximize Reliability

TABLE 7.1
Omitting Items—SWLS

Item	λ_j	ψ_j	$\Sigma_{(j)}\lambda_j$	$\Sigma_{(j)}\psi_j^2$	ω	λ_j^2/ψ_j^2
1	1.290	.901	4.389	6.232	.756	1.847
2	1.104	1.274	4.575	5.859	.781	.957
3	1.148	1.144	4.531	5.989	.774	1.152
4	.952	1.863	4.727	5.270	.809	.486
5	1.185	1.951	4.494	5.182	.796	.720

Note. $\Sigma_{(j)}$ denotes summation with item j omitted.

- Given the items of the SWLS, I pose to you a question:
  - Which items would you select if you wanted to make a four-item test with maximum reliability.

Item Information

- The far right column, representing the squared factor loading of an item divided by the item's uniqueness, is called the item's information.
  - This is a measure of the amount of information the item contains about the attribute.

- Higher amounts of information are desirable.
  - Tests constructed with the highest of informative items will have larger reliabilities.

- So, knowing this, how would you construct a four-item test for the SWLS?

Choosing Items in the True-Score Equivalence Model

- Recall that under the true-score equivalence model, all item factor loadings were equal.
- In this case, item information will vary only as a function of the unique variance of each item.
- Therefore, it is desirable to select items with the smallest unique variances in building a test.

Choosing Items in the Parallel-Items Model

- Recall that in the parallel items model, all items have the same factor loadings and the same unique variances.

- In this case, item information does not vary across all items.
  - Each item has the same amount of information.

- In this case, it does not matter which items we omit, only how many.

SWLS Example with Parallel Items

- To provide an example, recall the estimates of the parallel items model parameters for the SWLS:
  - $\lambda = 1.287$
  - $\Psi^2 = 1.442$
- Omega for a given number of items would then be:

1. 0.472
2. 0.641
3. 0.728
4. 0.781
5. 0.817

What About Building A Test With A Given Reliability?

- The real benefit to the parallel items model (and, really, the Spearman-Brown Prophecy Formula) is that of having the ability to estimate the number of items needed to realize a test with a specific reliability.

- For instance, what if you wanted to create a test measuring satisfaction with life that had a reliability of 0.9.
  - How many items would you need?

Spearman-Brown Prophecy Formula

- Recall that the Spearman-Brown prophecy formula states that the reliability of a single test item is:

$$\rho_1 = \frac{\lambda^2}{\lambda^2 + \psi^2}$$

- For a test with a projected *r* items, this is:

$$\rho_r = \frac{r\rho_1}{(r-1)\rho_1 + 1}$$

More Prophecy

- Re-working the formula, we can compute how many items, *r*, are needed for a parallel test to achieve a specific reliability:

$$r = \frac{\rho_r/(1-\rho_r)}{\rho_1/(1-\rho_1)}$$

- So, let's do this for the SWLS.
- Here, $\rho_1 = 1.287^2/ (1.287^2 + 1.442) = 0.472$
- What is the number of items needed to reach a reliability of 0.9?

SWLS Example

- The number of items needed is:

$$r = \frac{\rho_r/(1-\rho_r)}{\rho_1/(1-\rho_1)} = \frac{0.9/(1-0.9)}{0.472/(1-0.472)} = 10.06$$

- So, 10 items are needed for a test with a reliability of 0.9.

Additional Notes

- The assumption of parallel items can be a strict one.
  - To get around this assumption, one can compute ρ_1 based on what a person expects the average factor loadings and unique variances to be.

- Additionally, it should be noted that in item writing, it is often the first items that provide the most information.
  - Additional items cannot be expected to have similarly high factor loadings.

Scale and Reliability

Transformations of Test Scores

- Recall we talked about various transformations of test scores that could be taken.
  - For instance, imagine we wanted to have a mean of 100 and a standard deviation of 15.

- Once a transformation has occurred, the resulting estimates of variance are also changed:
  - Test score variance.
  - True score variance.
  - Error score variance.

Resulting Variances in Standardized Tests

- Imagine you standardized each examinee's test score.
  - You subtracted the mean and divided by the standard deviation.

- The resulting variances are:
  - Test score variance = 1
  - True score variance = ρ_r
  - Error score variance = 1- ρ_r

- Each of these can be found via the algebra of expectations.

Estimating a True Score
And Providing a Confidence Interval (Bounding)

Estimating A True Score

- The simple model Y = T + E has the property that the test score (Y) for an individual is an unbiased estimate of the true score for an individual (T).
- This is because we say that E has a zero mean.
- The problem is that we have a difficult time describing the E for an individual.
  - This is typically called a propensity distribution.
  - It is subject to a whole lot of processes going on in an individual's head during the time of the test.

True Score Estimates

- We call the estimate of the true score for an individual $\hat{t}$.

- The estimate of the true score is given by the test score:

$$\hat{t}_i = y_i$$

- The error variance is thus:

$$\sigma_E^2 = \sum_j \Psi_j^2$$

Bounding a True Score

- Because the estimate of error is the sum of *m* independent random components, it should approach normality for large *m*.
  - This is by the central limit theorem.

- So, just like the mean, we can use percentage points of a normal distribution.
  - The 95% confidence interval would then be:

$$y_i \pm 1.96\sigma_E$$

Bounding a True Score

- Imagine we had an omega of 0.817 and a total variance of 39.382 (from SWLS from Table 6.3).
- The error variance is (1-0.817) x 39.382 = 7.206
- The SE of measurement is then 2.684.
- So a person with a score of 20 on the test would have a true score lying within 25.261 and 14.739.

More on Estimates

- The observed score for is a satisfactory estimate of the examinee's true score for three reasons:
  1. It is unbiased.
  2. The SE of Measurement can be used to bound the score (because it is unbiased).
  3. If the items are parallel, it is the "best" estimate in the least squares sense.

Problems of Attenuation

Attenuation

- One of the earliest applications of true-score theory and the reliability coefficient was to correct the effects of errors on correlations between different tests.

- The estimated correlation between two tests will be inaccurate depending on the degree that each test is unreliable.
  - This inaccuracy shrinks the correlation.

Example Setup

- Imagine we have two tests, Y and V.
- Let T_Y be the true score for Y and T_V be the true score for V.
- Let ρ_Y be the reliability for Y and ρ_V be the reliability for V.

Error In Measurement

- Because we have Y and V and not T_Y and T_V, we are limited any correlation between Y and V will underestimate the correlation between T_Y and T_V.

- Because of this, corrections for attenuation have been developed:

$$\rho_{T_Y,T_V} = \frac{\rho_{Y,V}}{\sqrt{\rho_Y \rho_V}}$$

Correction Example

- From Spearman, imagine Y is a measure of pitch discrimination and V is a measure of intelligence.
  - $\rho_Y = 0.25$ and $\rho_V = 0.55$
- The correlation between Y and V is 0.38.

- The correlation between true scores is then:

$$\rho_{T_Y,T_V} = \frac{\rho_{Y,V}}{\sqrt{\rho_Y \rho_V}} = \frac{0.38}{\sqrt{0.25 \times 0.55}} = 1.03$$

Overcorrection

- Correcting for attenuation can lead to overestimated correlations (1.03 is not possible).
- Care must be used if correlations are to be corrected for attenuation.
- This topic is something that is cover much more in depth in other sources.
  - It is presented here to give you its relation to reliability theory.

Wrapping Up

- This chapter covers the applications of reliability theory.
  - Please read this, it is practical information.

- Reliability theory is something that has been around for many years.
  - Many of these applications are classics.

Next Time

- Computer Applications Lab
  - Meet in Room 4 – in the basement of Fraser Hall.