Reliability Theory for Total Test Scores

Measurement Methods

Lecture 9

Today's Class

- Some applications of the true score model.
- Before we begin…do you have any questions about homework?

Preliminaries

- Our upcoming schedule:
  - 2/28 (today) – The end of Chapter 5.
  - 3/2,7,9 – Chapter 6.
  - 3/14 – Midterm review session.
  - 3/16 – Midterm.
  - 3/21, 23 – SPRING BREAK!!!!

Great Moments in Measurement

2006 Rose Bowl Game

- January 1, 2006.
- Vince Young leads Texas to a come-from-behind victory over USC to claim College Football's National Championship
- He subsequently declares for the NFL draft.

Detroit Free Press www.freep.com

More Pro football

- ROUNDUP: Lowdown on Young's Wonderlic results

Today's top stories

- DETROIT 84, CLEVELAND 72: The last laugh
- FEEDING FRENZY: Demand for Polish pastries gets too big for Fat Tuesday alone
- State busts nearly 9,000 who dodged cigarette tax
- SPRING TV: A month you can't refuse
- Lose 10 pounds in the next 10 months

PRO FOOTBALL

ROUNDUP: Lowdown on Young's Wonderlic results

February 28, 2006

FREE PRESS NEWS SERVICES

Email this Print this

INDIANAPOLIS -- Even though Vince Young did not work out at the NFL scouting combine, the former Texas quarterback was by far the biggest story of the event.

zoom

Bevilling Carl Eller

According to various reports, Young either scored poorly on the Wonderlic test, or the test scores were improperly calculated. As of Monday, three people involved in personnel decisions maintain Young scored a 6 on the aptitude test, while Houston general manager Charley Casserly said that figure was inaccurate.

A player taking the Wonderlic test is given 12 minutes to answer 50 multiple-choice questions, which have nothing to do with football. Teams use the test to see how a player learns. Most teams want their quarterbacks to score in the 30s or higher.

"We don't treat the Wonderlic score for a quarterback any different than we do any other position," said Casserly, whose team owns the top pick in the draft. "We have a process that we go through that I don't think is any different than any other team. You have a Wonderlic score, and then you evaluate how the player plays on film."

The Wonderlic scores for all of the prospects in Indianapolis have not been released to the teams.

But if the reports of the score are accurate, it could hurt Young's standing as one of the top picks, although how much will only be known by draft time in April. One former NFC personnel chief said a low score raises some red flags, but a team would be apt to retest the player if that team had questions about him.

THE LAST ROW

WONDERING ABOUT WONDERLIC?

February 28, 2006

With all the hubbub is about Texas' QB Vince Young's reported poor score on the 50-question Wonderlic exam, the folks at ESPN.com have posted some sample questions from the 12-minute exam. Here are a few:

Which number in the following group of numbers represents the smallest amount?

7

.8

31

.33

2

The hours of daylight and darkness in SEPTEMBER are nearest equal to the hours of daylight and darkness in:

June

March

May

Nov.

A boy is 17 years old and his sister is twice as old. When the boy is 23 years old, what will be the age of his sister?

When rope is selling at $.10 a foot, how many feet can you buy for sixty cents?

Assume the first two statements are true. Is the final one: True, False, not certain?

Tom greeted Beth. Beth greeted Dawn. Tom did not greet Dawn.

In printing an article of 48,000 words, a printer decides to use two sizes of type. Using the larger type, a printed page contains 1,800 words. Using smaller type, a page contains 2,400 words. The article is allotted 21 full pages in a magazine. How many pages must be in smaller type?

Wonderlic Test

- From Wikipedia:
  - The **Wonderlic Personnel Test** (often referred to as ***Wunderlich***) is an intelligence test primarily known for being administered to prospective players in the National Football League since the 1970s. The Wonderlic is a twelve minute, fifty question exam to assess aptitude for learning a job and adapting to solve problems for employees in a wide range of occupations. The score is calculated as the number of correct answers given in the alloted time. A score of 20 is intended to indicate average intelligence (corresponding to an intelligence quotient of 100). It is rumored that at least one player has scored a 1 on the test.

What About Reliability?

- From http://cps.nova.edu/~cpphelp/WPT.html:
  - **Description:** The Wonderlic Personnel Test (WPT), so named to reduce the possibility that job applicants will think they are taking an intelligence test, was originally a revision of the Otis Self-Administering Tests of Mental Ability. The WPT is a 50-item, 12-minute omnibus test of intelligence. The items and the order in which they are presented provide a broad range of problem types (e.g., analogies, analysis of geometric figures, disarranged sentences, definitions) intermingled and arranged to become increasingly difficult. The WPT exists in 16 forms, and was designed for testing adult job applicants in business and industrial situations.
  - **Scoring:** The WPT yields one final score which is the sum of correct answers.
  - **Reliability:** The manual reports odd-even reliabilities, which are not appropriate for speeded tests; however, it also reports test-retest reliabilities of .82 to .94, and interform reliabilities of .73 to .95.

Applications of the True-Score Model

Applications of the Model

- To apply the true-score model, we wish to interpret T as the true of the examinee and E as the error of measurement.
  - We assume that a person's true score does not change between measurements.

- To the extent that factors outside of the test enter into the equation, the test-retest methods yield a valid coefficient of precision.

Methods Used to Estimate Reliability

- There are three main methods used to estimate the reliability coefficient:

1. Test-retest methods.
2. Parallel or alternate-form methods.
3. Internal analysis.

- We will revisit #3 in the next Chapter.

Test-Retest Methods

- A test of *m* items is administered to a large sample of examinees at two points in time
  - Yielding Y and Y'.

- We estimate $\rho_{YY'}$ from the test scores.
  - We then call this entity ρ_r
  - We can then estimate σ_E^2 and SEM(Y).

Assumptions

- In doing this whole process, we are making two very strong assumptions:
  1. The true scores of the examinees do not change between administrations of the test.
     - If this is the case, the errors are independent.
     - $\rho_{YY'}$ can be taken to be an approximation of the coefficient of precision.

- One big problem with this assumption is that there is no way to tell if it is violated.

Assumptions, Continued

- Our second assumption under the retest methods:
  2. We must say that the retest true score is defined as the component of the observed score that does not change between administrations.

- This is the reason that the reliability coefficient is sometimes called the coefficient of stability.

Time Intervals

- The main problem with #2 is that what we are observing is a small fraction of behavior over time.
  - Generally, the longer the time interval, the lower the coefficient.

- By having multiple measures, we do not have a single retest true score.
  - Our method has only been defined for a pair of tests.

Longitudinal Measurement

- There are very good reasons for collecting test data longitudinally.
- As of yet, we have not introduced a mechanism for relating longitudinal data to the ideal coefficient of precision.
- Lord and Novick (1968) mentioned that any coefficient of stability underestimates the coefficient of precision because the "error" variances includes unstable "true" variance.

Alternate Form Methods

- In **alternate form** methods, two tests – the alternate forms – contain disjoint sets of items.

- The two tests are administered to a set of examinees yielding Y and Y'.
  - Such administrations are usually very short in time.

Alternate Forms and the True Score Model

- To treat each test under the true score model, we must assume that the variances of Y and Y' are the same in the population of interest.

$$\sigma_Y = \sigma_{Y'}$$

- By same, we mean not significantly different as tested in the sample.

- Once we have verified this is the case, we then get an estimate of $\rho_{YY'}$.
  - This estimate is then used to estimate σ^2_E and the SE of measurement.

More on Alternate Forms

- In making the transition from retest reliability to alternate-form reliability, we note that the retest reliability from the total score has no relation to the precision with which we measure the attribute itself.

- The *m* items are indicators of the attribute, but may not be closely related to the attribute.
  - A set of items may have high stability and low-alternate-form reliability.
    - Or vice versa.

Even More on Alternate Forms

- We might say that it is an implicit assumption that the two forms equally measure the examinee's true scores.
  - They only differ by independent errors of measurement.
  - This is an untestable assumption.

- Rather, we might define the alternate-form true score to be a component in the two total test scores that is common to the forms we have constructed.
  - Then their errors are components of the total test scores that are unique to each form.
  - The resulting reliability coefficient is then called a coefficient of equivalence.
  - We will learn more about this when we introduce the common factor model in the next chapter.

How Many Alternate Forms?

- Technically speaking, a test can have as many alternate forms as there are tests of the same variance to correlate with.
  - Each correlation would yield another coefficient of equivalence.
- In application, content restrictions are placed on the items of the alternate forms.
  - We want these items to reflect the attribute being measured.

Coefficient of Equivalence

- When using alternate forms, we hope that the coefficient of equivalence (the estimate of reliability) will become a coefficient of precision.

- Lord and Novick (1968) state that generally the coefficient of equivalence will be less than the coefficient of precision.
  - This is because the error component will include true-score variability due to lack of parallelism of the tests.

- We will find more stringent requirements for parallel forms as we progress through the book.

Alternate Items

- We would like for items of alternate forms to be equivalent to the items of the original form.
  - Similar yet not identical.

- For example, consider a word-fluency test consisting of one frequency-count item:
  - Write down as many words you can think of beginning with the letter E

- How would you construct similar items?

More Terms

- We make a distinction between **content-parallel test forms** and **content-equivalent test forms.**

- Content-parallel test forms – two forms containing the same number of items, in which the items are paired to be similar in content, while distinct items within each form may be less similar.

- Content-equivalent test forms – two test whose items can be recognized as content-homogeneous when they are combined to make a test.

Content-Parallel Test Form Example

- To illustrate a content-parallel test, consider the following sets of items:

Form L	Form M
1. What day of the week is it?	1. What month is it?
2. If I buy 4 cents worth of candy and pay 10 cents, what change do I get?	2. If I buy 12 cents worth of candy and pay 15 cents, what change do I get?
3. Repeat in reverse order 6-5-2-8.	3. Repeat in reverse order 3-6-2-5.

Content Parallel Example #2

- Here is a test online…can you create some content-parallel items?
- http://www.blogthings.com/couldyoupasseighthgrademathquiz/

Content-Equivalent Example

- To demonstrate a content-equivalent test, consider the list of items shown to the right ->
- There are a number of ways to form similar-content subsets.
- Judgment of content-equivalent tests is largely subjective.
  - Statistical evidence of equivalence can be attained by methods from Ch. 6.

TABLE 5.2
Satisfaction With Life Items

You should agree or disagree with each item using the 1–7 scale below. Place a number from 1 to 7 next to each item on the answer sheet to indicate your degree of agreement with that item.

7. Strongly agree
6. Agree
5. Slightly agree
4. Neither agree nor disagree
3. Slightly disagree
2. Disagree
1. Strongly disagree

A	1. In most ways my life is close to my ideal.
	2. I frequently think about unhappy times or events of my past.
	3. I am a person who can feel happy very easily.
E	4. The conditions of my life are excellent.
	5. I am satisfied with the current state of affairs in my life.
	6. I like the life I have led.
C	7. I am satisfied with my life.
	8. I frequently experience intense negative emotions that make me unhappy.
D	9. So far I have gotten the important things I want in life.
	10. When something makes me happy, this emotion usually lasts a long time.
B	11. If I could live my life over, I would change almost nothing.
	12. My life does not live up to the standards I have for a good life.
	13. I am satisfied with my present life.
	14. If I imagine the most desirable life for myself (the ideal), my life is very close to that point.

Winding Down – Retest Methods

- A test-retest correlation is called a coefficient of stability
  - It generally bears no clear relation to anything we would regard as the precision of measurement of the test.
- Even if it approximates a coefficient of precision, such a value is commonly not of interest.
  - Approximation of the coefficient of precision requires some relatively difficult to attain conditions – no effects of previous responses, fatigue, etc…
- In most cases, the quantity of interest is the precision with which the attribute itself is measured.

Winding Down – Alternate Form Methods

- An alternate form test correlation is called a coefficient of equivalence.
  - It can be a possible measure of the precision of the measurement of the attribute itself.
    - This requires that all the items in the two forms are indicators of just that attribute.
      - It might be sufficient to have content-homogeneous items.

Winding Down – Alternate Form Methods

- One might think about what we have with alternate forms.
  - 2*m* items measuring the same attribute.

- There are numerous ways to divide such items into alternate forms.
  - In each of which, the coefficient of equivalence would vary.

Content Parallel Test Forms

- The set of content parallel items presents a measurement challenge.
  - The pairs of items measure an attribute at a higher level of abstraction than the attribute measured by the whole test.

- As a consequence, the coefficient of equivalence will be artificially inflated.

Wrapping Up

- The classical true-score model deals with test scores – the sums of items.
- To continue further, we need theory at the item level itself.
- The true-score model was presented for foundational reasons.
  - We will keep the model but improve the method.

Next Time

- Test homogeneity, reliability, and generalizability.
  - Chapter 6 – Part 1.