Validity

Measurement Methods

Lecture 17

Chapter 10

Today's Class

- Validity

Validity

- Definition:
  - "A test score is valid to the extent it that it measures the attribute of the respondents that the test is employed to measure, in the population(s) for which the test is used." (p. 197).
- To this end, we seek evidence to validate a given use of a test score.
- Evidence can come from:
  - Original design of the test
  - Intent and context of the use of the test

Considerations When Assessing Validity

- Technical methods of test theory.
  - correlation.
  - factor analysis.
  - regression.
  - path analysis.
- Substantive issues.
- Philosophical issues in social science.

History of Validity

Sam Shere, *Burning of the Hindenburg*, Lakehurst, New Jersey, May 6, 1937.

Evolution of Validity and Validation

- Earlier phase (dating to the late 1930s) was embedded in:
  - Behaviorist traditions in psychology.
  - Logical positivist philosophical framework.

- The idea that concepts embedded in test items were things being measured was considered “bad science”

- To be valid, the test score (or item scores) had to be related to some observable behavior (and predict differences between groups of individuals).

Early Objections to Content Validity

- Content validity – that item contents could serve to establish that the resulting test score was a valid measure of a specified attribute.
- Most psychologists objected to the "subjectivity" inherent in the definition of content validity.
- Also noted that a test score could be the result of many different response patterns.
  - So score is an indirect measure of the content of a test.
    - Sound like what we worry about now with multidimensionality.

Standards for Test Validation

- In 1954, the American Psychological Association (APA) issued a set of standards for validity.

- They defined four “types” of validity:
  - Predictive validity
  - Concurrent validity
  - Content validity
  - Construct validity

Construct Validity

- Construct validity – “the degree to which the individual possesses some hypothetical trait or quality [construct] presumed to be reflected in the test performance.” (p. 199).

- Chronbach and Meehl (1955) developed this concept of construct validity more thoroughly.
  - Called a construct an attribute.
  - Listed procedures for construct validity:
    - Criterion-group differences.
    - Factor analysis.
    - Item analysis.
    - Experimental studies.
    - Studies of process.

Common (current?) View of Validity

- As validity has shifted over the years, Rod reports a more common view that has evolved.

- There is but one conception of validity.
  - Rod says “by habit” this is called construct validity.

- “The validation of an application or class of applications of a test score can be taken to include every form of evidence that the score to some acceptable extent measures a specified attribute – quantifiable property or quality – of a respondent.” (p. 199).

The Labeling of Validity

- As time progressed, concurrent and predictive validity were combined into what was called criterion-related validity.

- A better term would be the predictive utility of a test.
  - There is a logical problem with saying a test is valid to the extent it correlates with another test.
  - Recall validity is the extent to which a test measures what it is supposed to measure.
  - What if the other test was not valid?

Validity in Rod's Book

- He looks at validity through several mechanisms:
  - Item content
  - Internal test evidence
  - Convergent and discriminant validity
  - Multitrait-multimethod matrices
  - External relations

Item Content and Test Score Validation

Test Content and Validity

- Rod warns about reading much into content of a test.
  - It could be possible that the content is created under a theory of the attribute that is not at all accurate (or is way to far reaching).

- But that being said, content analysis is commonly a large part of evidence for the validity of a test.

Content

- To create items, one must first begin with a concept of the attribute.
  - The concept can be multidimensional in nature.
- One can regard content design as being guided by a classification system of the test creator.
  - Differing types of items are needed to “tap” into the attribute fully.
- Content analysis can be done by having many people write items.

Content Analysis in Education

- In educational research the content of an achievement test can come from the types of objectives specified for proficiency of some type of domain.
  - Bloom's taxonomy is an organization of the level of understanding of concepts.

- Any psychometric structure shown in such tests may reflect the fact that cognitive performances are generally positively correlated.

External Review of Content

- The content of a test can be examined by having many experts rate the items already written for the type of content contained in the test.
- All types of rating systems could be employed.
  - Sorting tasks
  - Probabilistic assessments
- Additionally, test subjects can provide content assessment by employing a “think aloud” protocol.
  - Have subjects describe their thoughts about each item.
  - May work better in educational settings than in psychology, but then again, may be ok.

Internal Psychometric Analysis

Types of Analyses Needed

- Chapters 6 and 9 have provided much of the technology needed for an internal analysis of item or subtest relationships.
  - You can guess where we are going with this one…
- The main point here is that one should determine if a test (or subtest) is homogeneous.
  - That is essentially the extent of internal analysis that can be performed.
  - We can think of Omega as a coefficient of validity.

Steps to Estimating Validity Coefficients

- Rod provides the following steps to estimate the validity coefficient of a test (p. 211):
  1. A single-factor analysis gives the squared correlation of the total test score with the general factor – the domain score – by equations 6.20 or 6.21.
  2. An independent-cluster analysis gives the squared correlation of each cluster score with its cluster/group factor, again by 6.20 or 6.21.
  3. A hierarchical solution gives the squared correlation of the total test score with the hierarchical general factor, by 6.20 but not by 6.21.

Convergent and Discriminant Validity

Convergent and Discriminant Validity

- Dating to Campbell and Fiske (1959).
  - Multiple measures of a construct are said to have convergent validity if they are sufficiently highly correlated.
  - Multiple measures of a construct are said to have discriminant validity if they have sufficiently low correlations with test of other, distinct constructs.

What Does this Mean

- Recall our SWLS example where we had fit the independent-clusters model.
- We said the correlation between the two factors, while high, was sufficient to assume that the two clusters were separate attributes.
- Had the correlation been near 1.0, we would not be able to assume this point.

Quantifying Con/Dis Validity

- Rod suggests quantifying the convergent and discriminant validity by creating matrices with:
  - The correlation between the cluster sums and their own factor
  - The correlation between the cluster sums and other factors

SWLS Correlations

	F_1	F_2
Y_1	0.898	0.773
Y_2	0.672	0.780

Multitrait-Multimethod Matrices

Multitrait-Multimethod Matrices

- The section is optional, so we will skip most of it.
- I would recommend reading the section.
- Multitrait-Multimethod matrices are tools to inspect the validity of a set of measures.
- You can get some indication of convergent and discriminant validity.

Concurrent Validity and Predictive Utility

Notes About Predictive Utility

- Read the section on predictive utility – it has a couple good philosophical points.
- Mainly, I would like to point out the information at the end of the section – the effects of measurement error.
- In predictive utility studies, note that usually both the criterion (the old test) and the predictor (the new test) are not measured perfectly.
  - In this case, one can still correct for attenuation (see slides from Chapter 8).

Wrapping Up

- Validity is perhaps the most important part of the construction of a test.

- Psychometric methods can only take you so far, however.

Next Time

- Chapter 11 – Classical Item Analysis