Test Validity: What it is, and why we care.

Validity

- What is validity?
- What is a construct?
- Types of validity
  - Content validity
  - Criterion-related validity
  - Construct Validity
  - Incremental Validity

What is validity?

- The validity of a test is the extent to which it measures the construct that it is designed to measure
  - As we shall see, there are many ways for a test to fail or succeed = validity is not a single measure

Paul Meehl: What is a construct?

- Meehl's definition of a construct has 6 main elements, as follows:

1.) To say *what a construct is* means to say *what laws it is subject to.*

- This is *a definition* = you can refuse to work with it or say why you think it is bad, but you can't disprove it
- The sum of all laws is called a construct's nomological network.

What does 'nomological' mean?

I had always believed it came from:

ad. L. *nomin* meaning 'name'

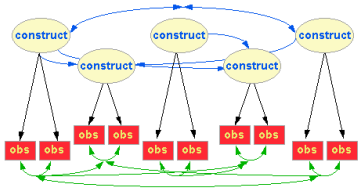
I was wrong. In fact it comes from:

ad. Gr. *nom* combining form of a word meaning 'law'

So 'psychonomics' is the study of the laws of the psyche, and 'nomological network' refers to a network whose components can be described by laws or rules

The Nomological Network

a representation of the concepts (constructs) of interest in a study,



...their observable manifestations, *and the interrelationships among and between these*

Image from: http://trochim.human.cornell.edu/kb/nomonet.htm

Paul Meehl: What is a construct?

2.) Laws may relate observable and theoretical elements

- The relations must be 'lawful', but they may be either causal or statistical (what's the relation?)
- What are the 'theoretical elements'? Constructs!

Paul Meehl: What is a construct?

- What are the 'theoretical elements'? Constructs!
- To escape from circularity and pure speculation about the properties of constructs, we need to anchor the nomological net, hence:

3.) A construct is only admissable if at least some of the laws to which it is subject involve observables

- If not, we could define a self-consistent network of ideas that had no relevance to the real world (and many such networks have been defined!)
- You should be able to relate this idea of observables to our earlier discussion of information: what counts as observable is what counts as information (detectable differences)

Paul Meehl: What is a construct?

4.) Elaboration of a construct's nomological net = learning more about that construct

- We elaborate a construct by drawing new relations, either *between elements already in the network*, or *between those elements and new elements outside of the network*
- This elaboration is precisely the work of psychometrics, as well as the work of science in general

Paul Meehl: What is a construct?

5.) Ockham's razor + Einstein's addendum

- That is: make things as simple as possible, but no simpler

6.) Identity means 'playing the same role in the same network'

- If it looks like a duck, walks like a duck, and quacks like a duck: then it is a duck!*
- Or (in the spirit of Gregory Bateson): If it makes no difference, then it makes no difference

* at least pending further investigation

How to measure validity

- Analyze the content of the test
- Relate test scores to specific criteria
- Examine the psychological constructs measured by the test

Content validity

- Content validity = the extent to which the test elicits a range of responses over the range of of skills, understanding, or behavior the test measures
- Most important with achievement tests, because there are usually no external criteria
- How can we determine content validity? (or: How will you know if you get given a good exam in this class?)
  - Compare the questions on the test to the subject matter
  - If it looks like a measure of the skill or knowledge it is supposed to measure, we say it has *face validity*

Criterion-related validity

- Criterion-related validity depends upon relating test scores to performance on some relevant criterion or set of criteria
  - i.e. Validate tests against school marks, supervisor ratings, or dollar value of productive work
- There are two kinds of criterion-related validity: concurrent and predictive

Criterion-related validity II

- Concurrent validity = the criterion are available at the time of testing
  - i.e. give the test to subjects selected for their economic background or diagnostic group
  - the validity of the MMPI was determined in this manner
- Predictive validity = the criterion are not available at the time of testing
  - concerned with how well test scores predict future performance
  - For example, IQ tests should correlate with academic ratings, grades, problem-solving skills etc.
  - A good *r*-value for most psychological questions would be .60

What affects criterion-related validity?

i.) **Moderator variables**: Those characteristics that define groups, such as sex, age, personality type etc.

- a test that is well-validated on one group may be less good with another

- validity is usually better with more heterogeneous groups, because the range of behaviors and test scores is larger

And therefore:

ii.) **Base rates**: Tests are less effective when base rates are very high or very low (that is, whenever they are skewed from 50/50)

What affects criterion-related validity?

iii.) **Test length**

- For similar reasons of the size of the domain sampled (think of the binomial rabbits or trying to decide how biased a coin is), longer tests tend to be more reliably related to the criterion than shorter tests

  - What are those reasons?

- Note that *this depends on the questions being independent* (= every question increasing information)

- when it is not, longer tests are not more reliable

  - eg. short forms of WAIS

- However, note that independence need only be partial ($r < 1$, but not necessarily $r = 0$)

What affects criterion-related validity?

iv.) **The nature of the validity criterion**

- Criterion can be contaminated, especially if the interpretation of test responses is not well-specified, allowing for results to 'feed back' to criterion

  - In such cases, there is confusion between the validation criteria and the test results = self-fulfilling prophecies

- In essence we are then stuck at the theoretical level of the nomological net, with no way for empirical study (= no information) to tell us we are wrong

Construct validity

- Construct validity = the extent to which a test measures the construct it claims to measure
  - Does an intelligence test measure intelligence? Does a neuroticism test measure neuroticism? What is latent hostility since it is latent?
- It is of particular importance when the thing measured by a test is not operationally-defined (as when it is obtained by factor analysis)
- As Meehl notes in the paper we just read, construct validity is very general and often very difficult to determine in a definitive manner

How to measure construct validity

i.) Get expert judgments of the content

ii.) Analyze the internal consistency of the test (Tune in next class for how to do this.)

iii.) Study the relationships between test scores and other non-test variables which are known/presumed to relate the same construct (sometimes called 'empirical validity')

- eg. Meehl mentions Binet's vindication by teachers

iv.) Question your subjects about their responses in order to elicit underlying reasons for their responses.

v.) Demonstrate expected changes over time

How to measure construct validity

vi.) Study the relationships between test scores and other test scores which are known/presumed to relate to (or depart from) the construct

- Convergent versus discriminant validity
- Multitrait-multimethod approach: Correlations of the same trait measured by the same and different measures > correlations of a different trait measured by the same and different measures

What if correlations of measures of different traits using the same method > correlations of measures of the same trait using different methods?

Incremental validity

- *Incremental validity* refers to the amount of gain in predictive value obtained by using a particular test (or test subset)
- If we give N tests and are 90% sure of the diagnosis after that, and the N+1th test will make us 91% sure, is it worth 'buying' that gain in validity?
  - Cost/benefit analysis is required.

Validity coefficient

- Validity coefficient = correlation (r) between test score and a criterion
- There is no general answer to the questions: how high should a validity coefficient be? Or: What shall we use for a criterion?

Measuring validation error

- Coefficient of determination = r^2
  = the percent of variation explained
- Coefficient of alienation = $k = (1 - r^2)^{0.5}$
- k is the inverse to correlation: a measure of nonassociation between two variables
  - If k = 1.0, you have 100% of the error you'd have had if you just guessed (since this means your r was 0)
  - If k = 0, you have achieved perfection = your r was 1, and there was no error at all*
  - If k = 0.6, you have 60% of the error you'd have had if you guessed

* N.B. This never happens.

Example

- The correlation between SAT scores and college performance is 0.40. How much of the variation in college performance is explained by SAT Scores?
- $r^2 = 0.16$, so 16% of the variance is explained (and so 84% is not explained).
- What is the coefficient of alienation?
  - Sqrt(1- 0.16) = Sqrt(0.84) = 0.92

Why should we care?

- k is useful in reporting accuracy of a test in a way which is unit free BUT notice that it tells you nothing you didn't already know from being told r