Ch. 5: Validity

- History
  - Griggs v. Duke Power
  - Ricci vs. DeStefano
- Defining Validity
- Aspects of Validity
  - Face Validity
  - Content Validity
  - Criterion Validity
  - Construct Validity
- Reliability vs. Validity
- Variance

Griggs v. Duke Power (1971)

- Group of l3 people employed as laborers -- sweeping & cleaning
- Wanted to be promoted to next higher classification (coal handler)
- Duke Power company required passing score on IQ test to be promoted
- Of 95 employees at power station, 14 were Black, 13 of 14 were assigned sweeping/cleaning duties
- Court case -- was the IQ test requirement valid or discriminatory?
- Supreme Court decision : "invalid"

Griggs v. Duke Power - 2

- Supreme court found

  If a test impacts different ethnic groups disparately, the business must demonstrate the test is a "reasonable measure of job performance"
- In scientific terms: Tests must be valid predictors of specific criteria.

Definitions of Validity

- Agreement between test scores and the quality (characteristic, feature, etc.) it is claimed to measure
- Many different definitions emerged in the 20th century, some confusing or incompatible with each other
- AREA/NCME (1985, 1999, 2012) "Standards for Educational and Psychological Testing"
- One informal definition: Face Validity
- Three formal definitions: Content, Criterion, Construct

Face Validity

- Common Sense / Informal Analysis
- "I like mechanics magazines" = you like mechanics magazines.
  "I never tell a lie" = you never lie, etc.
- Question -- what factors might influence a test-taker's response?
- Face validity is not a proper type of validity at all.
- Quizzes in magazines or on the Internet -- appear "face valid" but usually have low reliability and very low validity
- Psychometrically unsound

Does Face Validity Matter?

- Naive view = face validity
- Tests with very little face validity...
  - what does the average test taker feel about the test?
  - motivation?
  - confusion?

Content Validity

- Does the content of the test match the concept/ area in question?
- Most related to educational settings (achievement/aptitude testing)
- E.g. does an Algebra test contain questions about Algebra?

- This is a Logical, rather than statistical argument
- Somewhat fuzzy definition
- Modern theories consider Content Validity a sub-set of other types of validity

Content Validity 2

- If a test is supposed to test a specific *Construct*, problems may arise:

- Construct underrepresentation
  - test misses important information
- Construct-irrelevant variance
  - scores are influenced by outside factors
    - e.g. anxiety, reading comprehension, IQ, etc.

Criterion Validity

- Criterion -- a well defined measure of performance in the real world
- Criterion validity -- how well a test measure correlates with a specific criterion
- Predictive vs. Concurrent
- Predictive

  High School SAT score (predictor) predicts later College GPA (criterion)
- Concurrent

  Work samples from mechanics

Validity Coefficient

- Generally: relationship between test score and criterion
- Specific: often a standard Pearson product-moment correlation (r)
- In practice, r above .60 is rare! .40 is common
- Remember,
- r^2 = variance explained.
  r = .60 means just 36% of variation in the criterion scores explained by the predictor score (means 64% is not explained)
  r = .40 --> 16% of variance explained (84% not)

Evaluating Validity Coefficients

- Changes in the <u>cause</u> of relationships

  change in setting between when validity was measured (such as men vs. women in the workforce)
- What does the criterion mean?

  esp. when comparing one test with another test
- Review subject population
- Sample size? Cross-validation? (shrinkage)
- Don't confuse the Criterion with the Predictor

  e.g. requirement of certain GRE score to graduate

Evaluating Validity Coefficients 2

- **Restricted range** of predictor or criterion
  GRE is poor predictor of first-year grades in graduate school
  - Why? perhaps because in graduate school only As & Bs are given…
- How well does validity generalize?
  -- Candy Corn predictor scale given November 1st?
- Differential prediction?
  Men vs. women? English speakers vs. non-english speakers?

Construct Validity 1

- Construct = Emerging term (since the 1950s)
- Problem was "what is criteria?" for many psychological concepts (such as IQ)
- Construct = made-up entity. Often not observable or measurable.
- Big problem -- how to measure validity of a test if the criterion can't be measured
- Issue -- does inability to define or measure something mean it doesn't exist? e.g. "Love" this is the converse of the "numerical fallacy"

Construct Validity 2

- Solution -- recognize that psychology is complicated, and (just like other sciences) things can exist even if they aren't easily measured
- Method -- collect evidence for the construct via multiple methods, multiple sources, multiple subjects

Construct Evidence

- **Convergent Evidence** -- when data from multiple sources all tend to point to the same conclusion.
- **Divergent Evidence**
  - aka **Discriminant** Evidence
- Evidence that a Construct is NOT the same as another
- Example : a measure of insomnia should correlate with duration of sleep, but should not correlate with other un-related constructs (such as emotional expression)

The Love Test

- Rubin (1970)'s Love Scale
- From Literature, created 198 items on Likert scale
- Result: a "Love" scale and a "Liking" scale
- Love scale: attachment, caring, intimacy
- Convergent evidence:
  - lovers vs. friends
  - eye contact
- Divergent evidence:
  - possible to love someone w/o liking them

All Validity is Construct Validity?

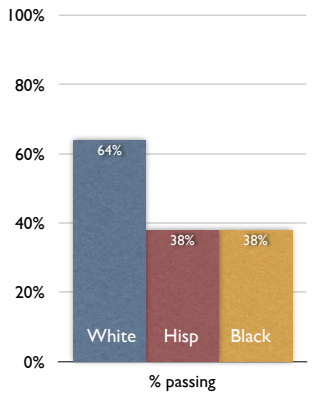
- Most modern theories consider that there is only one type of validity -- Construct validity
- All other types of validity are really sub-types of Construct validity.

Ricci v. DeStefano (2009)

- Eighteen firefighters (17 white, 2 hispanic) in New Haven, CT filed suit against the city
- Background:
  - All had passed a test (for promotion to management) scoring above a cutoff
  - None of the African Americans had scored above the cutoff (though they passed)
  - City vacated the test results, fearing lawsuit -- promotions were denied -- nobody was promoted

Ricci v. DeStefano - 2

- The Test
  - 60% written exam
  - 40% oral exam
- Passing score = 70%*
- *if weighted 30/70 2 AAs and 1 HI would have passed



Ricci v. DeStefano 3

- Supreme court decision:
- Found City in violation of the law
- Race-based action can be taken only if "demonstrate a strong basis in evidence that, had it not taken the action, it would have been liable under the disparate-impact statute"
- Summary: tests are discriminatory only if they are not related to the job. Not simply if there is evidence that different races get different results.

Review

- Reliability : easier to define and calculate. A property of the Test itself.
- Validity : harder to define, not inherent to the test, depends on the *way the test results are used*.

Reliability vs. Validity

- Validity coefficient is the correlation between a test and the criterion
- We know that *Test Measurements* and *Criterion Measurements* are unreliable
- The maximum validity is the square root of the product of their individual reliabilities. $r_{12max} = \text{sqrt}(r_{11}r_{22})$
- Thus, it's quite possible to completely miss a valid relationship if the measurements are not very reliable

Reliability vs. Validity : Example

Reliability of Test	Reliability of Criterion	Maximum Validity (r)
1	1	1
0.8	1	0.89
0.6	1	0.77
0.4	1	0.63
0.2	1	0.45
1	0.5	0.71
0.8	0.5	0.63
0.6	0.5	0.55
0.4	0.5	0.45
0.2	0.5	0.32

Variance: Reliability & Validity

- Variance in test scores can be divided into different portions
- In this example, only 16% is useful (validly predicts criterion)
- Other sources of error are known or unknown