Ch. 5: Validity

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
- Validity as Explained 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.

Disparate Impact

- aka "Adverse Impact"...
  - disproportional...
  - adverse effect...
  - on a protected class...

Impact

- A negative Effect on
  - Employment or
  - Housing

Disproportionate

- The proportion of a protected class affected by the behavior is different from a non-protected class
  - 80% rule

- Example
  - 50% of men are hired
  - 45% of women are hired
  - 45/50 = 90%

Protected Classes

- Federal Law:
  - race, color, religion, national origin, age (40+), sex, pregnancy, citizenship, familial Status (kids), veterans, genetic status
- California Law:
  - ...disability, sexual orientation, HIV or medical status, political party, victim of domestic violence

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
- Quizzes in magazines or on the Internet -- appear "face valid" but usually have low reliability and very low validity

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 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

Norm vs. Criterion

- Norm-referenced test
  - "Hire the top 5% of applicants"
  - Pro - select the best
  - Con - can they do the job?
- Criterion-referenced test
  - "Hire those who can do _____"
  - Pro - they can do the job
  - Con - not selecting the best

Measuring Validity

- General: relationship between test and what it's supposed to be measuring

- Specific: Pearson product-moment correlation (r) between Test Score (X) and criterion (Y)

How much Validity?

- In practice, r above .60 is rare Smaller numbers like .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

- Validity depends on the cause of the correlations
- And on the meaning of the criterion
- If the reason for the relationship changes, validity may change as well.

- Examples:
  - Different *subjects*
  - Different *situation*
  - Different *criterion*

Construct Validity 1

- Construct = made-up entity. Often not directly observable or measurable.

- Problem: "what is criteria?" for many psychological concepts (example: Intelligence / IQ)

- Problem: how to measure validity of a test if the criterion can't be measured?

Construct Validity 2

- Solution -- the world is complicated. In Psychology (as in other sciences) things can exist even if they aren't easy to measure.
- Method -- collect evidence for the construct via multiple methods, multiple sources, multiple subjects

Construct Evidence

- **Convergent Evidence** — data from multiple sources all tend to point to the same conclusion.

- **Divergent Evidence** (aka **Discriminant** Evidence) — evidence that a Construct is <u>not</u> the same as another

- Example : a measure of insomnia should correlate with duration of sleep, but should not correlate with unrelated construct (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?

- Modern theory: only one type of validity -- Construct validity
- Other types of validity are just sub-types of Construct validity.

Ricci v. DeStefano (2009)

- Eighteen firefighters (17 white, 2hispanic) 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 3

- Supreme court decision:
- Found City in violation of the law
- Race-based action can be taken only if "demonstrate a <u>strong basis in evidence</u> that, had it not taken the action, it would have been liable under the disparate-impact statute"

- Tests are discriminatory:
  - if they are _not valid for the job_.
  - Not just because _protected classes_ get different results.

4 Kinds of Validity

	Description	Notes	Statistic
Face	do items "look" valid?	informal, improper, non-scientific	none
Content	do test questions cover the topic?	logic & judgement - there are no stats to calculate	none
Criterion	does the test predict a specific event?	requires a well-defined criteria	Person's R (correlation) between Test and Criteria
Construct	does the test measure the underlying thing	modern theory: all validity is Construct validity	Convergent and Divergent correlations

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. $r_{12max} = \sqrt{r_{11}r_{22}}$

- If the tests have low reliability, validity can be hidden

Reliability vs. Validity : Example

Reliability of Test	Reliability of Criterion	
CES-D	**DSM-V Depression**	**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

Reliability vs. Validity

Unreliable & Unvalid

Unreliable, But Valid

Reliable, Not Valid

Both Reliable & Valid
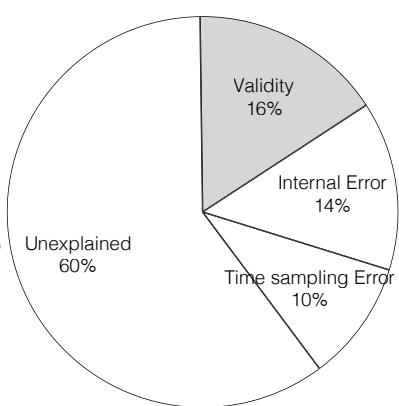
- Shotgun vs. rifle analogy
- Similar to accuracy vs. precision
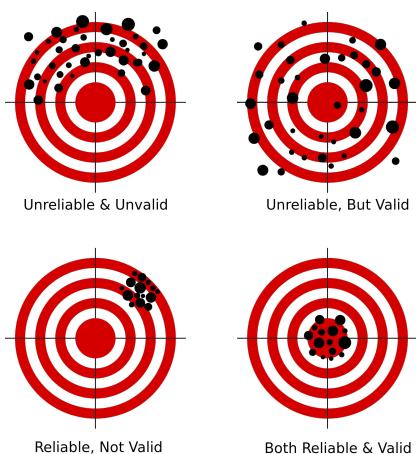
Reliability or Validity Part 1

Procedure	Reliability or Validity	What kind?
Correlation of IQ scores with ability to handle 30 tons of coal per day		
Correlation between SAT scores taken in Junior vs. Senior year of High School		
Having a committee review a high school history exam to make sure the questions cover all required topics		
Correlation between scores on two different versions of a Final Exam		

Reliability or Validity Part 2

| Procedure	Reliability or Validity	What kind?
Showing that your new measure of Anxiety correlates with an old measure of Anxiety, and does not correlate with IQ		
Showing that everyone in your research lab can rate "grumpy" expressions the same way		
Looking at a question about depression and deciding that it measures depression.		
Showing that questions on your test all correlate strongly with each other		

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*.

Classical Test-Score Theory

- T= True Score
- X = Observed
- E = Error

- X = T+E
- E = X-T

T ↔ X
  E
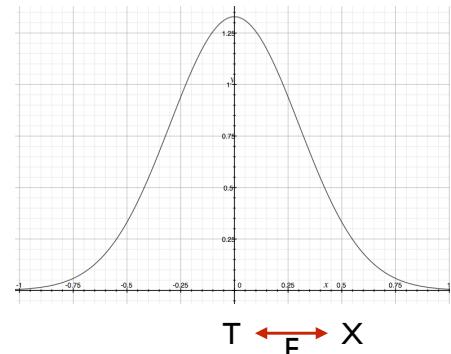
4 kinds of Reliability

	Description	Name	Statistic
Time Sampling	1 test given two times	test-retest reliability	correlation between scores at two times
Item Sampling	2 different tests given once	Alternate or Parallel forms	correlation between scores on 2 versions
Internal Consistency	One test, multiple items	Split Half or internal reliability	Cronbach's Alpha
Observer Differences	One test w/ 2+ observers	inter-observer reliability	Kappa

4 Kinds of Validity

	Description	Notes	Statistic
Face	do items "look" valid?	informal, improper, non-scientific	none
Content	do test questions cover the topic?	logic & judgement - there are no stats to calculate	none
Criterion	does the test predict a specific event?	requires a well-defined criteria	Person's R (correlation) between Test and Criteria
Construct	does the test measure the underlying thing	modern theory: all validity is Construct validity	Convergent and Divergent correlations

How much validity do you need?

- Clinical Significance
  - Effect Size of Benefits
  - vs Cost and Harm

- Examples:
  - an inexpensive test which can detect 1% of cancers (validity r^2=1%)
  - a written test for airline pilots to detect who is unsafe (validity r^2=93%)

- Answer: it depends

---

4 Kinds of Validity

	Description	Notes	Statistic
Face	do items "look" valid?	informal, improper, non-scientific	none
Content	do test questions cover the topic?	logic & judgement - there are no stats to calculate	none
Criterion	does the test predict a specific event?	requires a well-defined criteria	Person's R (correlation) between Test and Criteria
Construct	does the test measure the underlying thing	modern theory: all validity is Construct validity	Convergent and Divergent correlations

---

Constructs, Factors, Facets

- Constructs may be "high" or "low" (also called "top" or "bottom")
- Top-level constructs are made of smaller constructs
- aka Factors, Factors, Dimensions, Domains...

Construct: Anxiety

Construct: Anxiety

- Cognitive

- Emotional

- Physiological

Convergent Validity

- Multiple factors within a construct
  or
  multiple measures of a construct

- All correlate with each other

Divergent Validity

- Other factors (not part of a construct)
- Should have low to zero correlation

- What about negative correlations?

Psyc 402 Project - Overview

- Goals
- Proposal
- Data collection
- Analysis
- Paper

Psyc 402 Project Goals

- Literature Review
  - two kinds of articles
- Creating new test measures
  - writing two test questions
- Reliability of your measures
  - which kind?
- Validity of your measure
  - convergent
  - divergent (discriminant)
- APA writing

Project - Proposal

- Pick a field and sub-field
  - get 1 *Review* journal article
- Pick a construct to measure
  - get 1 *Original Research* article
- Design two 1-item tests to measure your construct
- Write Proposal: 3 pages, 2 tests, 2 articles
- **Due Date: see syllabus**
- Note: late submissions may not be included in survey

Psyc 402 - Survey

- Your 2 questions must be format that can be done on website (e.g. multiple choice, Likert, etc)
- All students will take the test answering all questions
- Anonymity
- Additional Data
  - Gender
  - Age, GPA, Midterm 1 and Midterm 2 scores
  - NEO-FFI (5 factors)

Psyc 402 - Limitations

- Construct: can not choose
  - Anxiety
  - Any of the NEO-FFI scores (Neuroticism, Extraversion, Openness, Agreeableness, Conscientiousness)
- Convergent & Divergent factors
  - Gender: no
  - May use another student's factors

Psyc 402 Project - FAQ

- APA Format
- Length
  - Proposal : about 3 pages (750 words)
  - Paper : about 8 pages (2000 words)
- How to find *Original Research* that includes a test? Use references or citation index.
- Survey Question Format:
  - Ordinal or Interval/Ratio Scale (numerical) answers : much easier to analyze
  - Avoid Nominal scales