

Ch. 5: Validity

494

Psychology 402 - Fall 2021 - Dr. Michael Diehr

Copyright © 2020 Michael Diehr
All Rights Reserved

For use only by students enrolled
in my sections of Psyc 402
through December 2020.
May not be posted, shared or uploaded
online without permission.

495

Psychology 402 - Fall 2021 - Dr. Michael Diehr

Validity

- History
 - Griggs v. Duke Power
 - Ricci v. DeStefano
- Defining Validity
- Four Kinds of Validity
- Reliability vs. Validity
- Validity as Explained Variance

507

Psychology 402 - Fall 2021 - Dr. Michael Diehr

Griggs v. Duke Power (1971)

- Duke Power company - Coal fired power plant
- 13 people employed as Janitors
- Wanted promotions to next level (coal handler)
- Duke required certain score on IQ test for promotion
- Of 95 employees at power station, 14 were Black, 13 of 14 were janitors
- Legal case: was the IQ test requirement valid or discriminatory?
- Supreme Court decision : “invalid”

508

Psychology 402 - Fall 2021 - Dr. Michael Diehr

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.

509

Psychology 402 - Fall 2021 - Dr. Michael Diehr

Disparate Impact

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

510

Psychology 402 - Fall 2021 - Dr. Michael Diehr

Impact

- A negative Effect on
 - Employment or
 - Housing

511

Psychology 402 - Fall 2021 - Dr. Michael Dohr

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
 - $45/50 = 90\%$

512

Psychology 402 - Fall 2021 - Dr. Michael Dohr

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

513

Psychology 402 - Fall 2021 - Dr. Michael Dohr

Definitions of Validity

- Agreement between test scores and the thing (construct) it claims to measure
- Many other definitions; 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

514

Psychology 402 - Fall 2021 - Dr. Michael Dohr

Face Validity

- Common Sense / Informal Analysis
- “I like mechanics magazines” = you like mechanics magazines.
“I never tell a lie” = you never lie
- Issues:
 - 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

515

Psychology 402 - Fall 2021 - Dr. Michael Dohr

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?

516

Psychology 402 - Fall 2021 - Dr. Michael Dohr

Content Validity

- Do the test questions cover the construct?
- Commonly used in education (achievement/ aptitude testing)
- Example: does an Algebra test contain questions about a wide range of Algebra topics?
- This is a Logical, rather than statistical argument
- Fuzzy definition
- Modern theories consider Content Validity a sub-set of other types of validity

517

Psychology 402 - Fall 2021 - Dr. Michael Dohr

Content Validity 2

- If a test is supposed to test a specific *Construct*, problems may arise:
- Construct underrepresentation
 - test questions miss important topics
- Construct-irrelevant variance
 - test scores are influenced by outside factors
 - e.g. anxiety, reading comprehension, IQ, etc...

518

Psychology 402 - Fall 2021 - Dr. Michael Dohr

Content Validity

519

Psychology 402 - Fall 2021 - Dr. Michael Dohr

Improving on Content Validity

520

Psychology 402 - Fall 2021 - Dr. Michael Dohr

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

521

Psychology 402 - Fall 2021 - Dr. Michael Dohr

Criterion Validity

- Criterion -- a well defined measure of performance *in the real world*
- Criterion validity -- the correlation between a test score and the specific criterion

522

Psychology 402 - Fall 2021 - Dr. Michael Dohr

Predictive vs. Concurrent

- Predictive
High School SAT score (predictor) predicts later College GPA (criterion)
- Concurrent
Work samples from mechanics

523

Psychology 402 - Fall 2021 - Dr. Michael Dohr

Measuring Validity

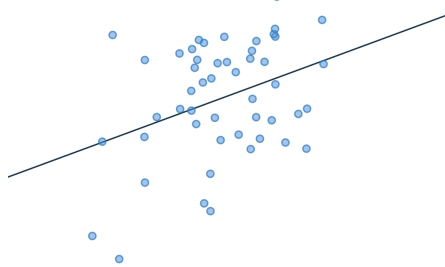
- General: relationship between test and what it's supposed to be measuring
- Specific: Pearson product-moment correlation (r) between Test Score (X) and criterion (Y)

524

Psychology 402 - Fall 2021 - Dr. Michael Dohr

Criterion Validity Example

College GPA (first year)



SAT Score

- SAT vs GPA
- Validity = $R_{X,C}$
- Predictive
- $R=0.36$
- $R^2=13.0\%$

525

Psychology 402 - Fall 2021 - Dr. Michael Dohr

How much Validity?

- Validity is often lower than you'd want
- $r > 0.60$ is rare
- $r = 0.40$ or lower is common
- $r^2 = \%$ of variance explained
- $r = .60$ means just 36% of variation in the criterion scores explained by the predictor score (means 64% is not explained)
- $r = .40$ means 16% of variance explained (84% not)

526

Psychology 402 - Fall 2021 - Dr. Michael Dohr

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*

527

Psychology 402 - Fall 2021 - Dr. Michael Dohr

Problems with Criterion Validity

- “what is the criteria?”
- Some criteria are obvious and sensible:
E.g. using SAT to predict college GPA
- For many constructs, it's not clear:
E.g. intelligence. IQ test used to predict ____?
- Without a *specific* criterion, you can't calculate criterion validity!

528

Psychology 402 - Fall 2021 - Dr. Michael Dohr

Construct Validity

- Construct = theoretical entity. In soft sciences, usually not directly observable or measurable.
- 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

529

Psychology 402 - Fall 2021 - Dr. Michael Dohr

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 not the same as another

530

Psychology 402 - Fall 2021 - Dr. Michael Dohr

Construct Validity Example

- Construct: Insomnia
- Convergent correlations:
 - # of hours slept
 - tired feeling
 - another test of insomnia
 - R should be strong (either positive or negative)
- Divergent correlations:
 - age
 - IQ
 - R should be weak (close to zero)

531

Psychology 402 - Fall 2021 - Dr. Michael Dohr

Construct Validity Example

	An old test of insomnia	# of hours slept	Feeling Tired	IQ	Age
Correlation between your test of insomnia and...	✓	✓	✓	none	none
Which kind of Construct validity?	Convergent	Convergent	Convergent	Divergent	Divergent

532

Psychology 402 - Fall 2021 - Dr. Michael Dohr

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 had 3 factors: attachment, caring, intimacy
- Convergent evidence:
 - lovers vs. friends
 - eye contact
- Divergent evidence:
 - possible to love someone w/o liking them



533

Psychology 402 - Fall 2021 - Dr. Michael Dohr

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.

534

Psychology 402 - Fall 2021 - Dr. Michael Dohr

Ricci v. DeStefano (2009)

- Written and Oral test for promotion to firefighter Captain position in Newhaven CT
- Pass rate: Whites 64%, African Americans 38%, Hispanics 38%
- Scoring rule: only top scores eligible
- None of the African Americans would get promotion
- City, worried about Disparate Impact, denied promotions to everyone
- 18 White employees filed suit against the city
- Question: was this the right action to take?

535

Psychology 402 - Fall 2021 - Dr. Michael Dohr

Ricci v. DeStefano 2

- Supreme court decision:
- Found City *did the wrong thing*
- 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”
- Tests are discriminatory:
 - if they are not valid for the job.
 - Not just because *protected classes* get different results.

536

Psychology 402 - Fall 2021 - Dr. Michael Dohr

4 Kinds of Validity

	Description	Notes	Statistic(s)
Face	do items “look” valid?	informal, improper, non-scientific	none
Content	do test questions cover the topic?	logic & expert judgement	none
Criterion	does the test predict a specific outcome?	requires a well-defined criterion	Pearson’s R (correlation) between Test and Criterion
Construct	does the test measure what it claims	modern theory: all validity is Construct validity	Convergent and Divergent correlations (Pearson’s R)

537

Psychology 402 - Fall 2021 - Dr. Michael Dohr

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.

$$r_{12max} = \sqrt{r_{11}r_{12}}$$
- If the tests have low reliability, validity can be hidden

538

Psychology 402 - Fall 2021 - Dr. Michael Dohr

Reliability vs. Validity : Example

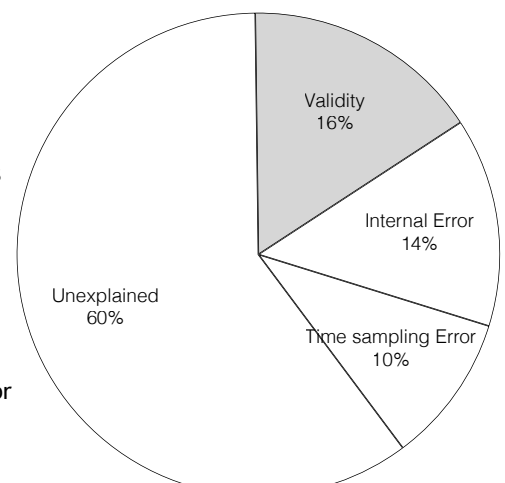
Reliability of Test	Reliability of Criterion	
CES-D	DSM-V Depression	Maximum Validity (r)
1	1	1
0.8	1	0.89
0.6	1	0.77
0.4	1	0.63
0.2	1	0.45
1	0.5	0.71
0.8	0.5	0.63
0.6	0.5	0.55
0.4	0.5	0.45
0.2	0.5	0.32

539

Psychology 402 - Fall 2021 - Dr. Michael Dohr

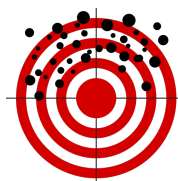
Variance: Reliability & Validity

- 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

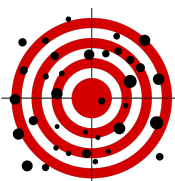


Psychology 402 - Fall 2021 - Dr. Michael Dohr

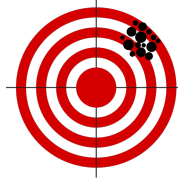
Reliability vs. Validity



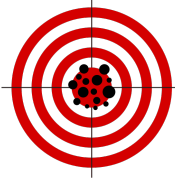
Unreliable & Unvalid



Unreliable, But Valid



Reliable, Not Valid



Both Reliable & Valid

- Target shooting 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		

Ch. 5 - Part 2

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*.

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(s)
Face	do items "look" valid?	informal, improper, non-scientific	none
Content	do test questions cover the topic?	logic & expert judgement	none
Criterion	does the test predict a specific outcome?	requires a well-defined criterion	Pearson's R (correlation) between Test and Criterion
Construct	does the test measure what it claims	modern theory: all validity is Construct validity	Convergent and Divergent correlations (Pearson's R)

555

Psychology 402 - Fall 2021 - Dr. Michael Dohr

Constructs, Factors, Facets

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

557

Psychology 402 - Fall 2021 - Dr. Michael Dohr

Construct: Anxiety

558

Psychology 402 - Fall 2021 - Dr. Michael Dohr

Convergent Validity

- Multiple factors within a construct or multiple measures of a construct
- All correlate with each other strongly

560

Psychology 402 - Fall 2021 - Dr. Michael Dohr

Divergent Validity

- Other factors (not part of a construct)
- Should have low to zero correlation
- What about negative correlations?

561

Psychology 402 - Fall 2021 - Dr. Michael Dohr