E-411-PRMA

Lecture 5

Christopher David Desjardins

31 August 2015

This week

- Classical test theory
- Validity

By Hand

What is the KR-20 for this toy example?

Item 1	Item 2	Item 3
0	1	0
1	1	0
1	1	1

What is the Coefficient alpha for this toy example?

Item 1	Item 2	Item 3
4	3	4
4	3	3
5	5	5

Inter-rater reliability

- Two raters measure the same behavior
  - For example: Number of aggressive behaviors observed in a child during play time.
  - Degree to which these raters report the same incidence of aggressive behaviors is a measure of reliablity
- Correlate scores from raters (e.g. Pearson's or Spearman's rho, etc)
- Important thing to note: test scores have reliability NOT test

IRR example

Two parents are administered the CBCL (an instrument to identify problem behaviors in children) on their four children. How well do their scores for the section *Aggressive Behavior* agree (i.e. what is their inter-parent reliability)?

Child	Parent 1	Parent 2
1	5.5	6.0
2	5.2	5.2
3	4.6	4.0
4	6.6	5.6

Make sure you understand Table 5-4!

Test affects on reliability

- More homogeneous, higher reliability
- More static the characteristic, higher reliability
- Restriction range, lower reliability
- Power (difficult test with no prefect scores) vs. speed test (time limitations)
  - If speed, reliability estimates may be too high bc items are too easy
  - Everyone expected to get all of them right
  - Test-retest, alternate-forms, or split halves from two independently timed half tests
- Criterion-referenced, lower variability, lower reliability
  - If everyone has met the standard/criteria!

Calculating True Score

- Erla takes 3 tests (parallel forms) in math
- She gets an 8, 7, and 7.5
- What should we estimate as her true score/ability in math?
- Do you think that score is her true score?

Calculating True Score

- Erla takes 3 tests (parallel forms) in math
- She gets an 8, 7, and 7.5
- What should we estimate as her true score/ability in math?
- Do you think that score is her true score?
- We need a way to quantify uncertainty about Erla's score

Standard Error Measurement

$$\sigma_{SEM} = \sigma\sqrt{1 - r_{xx}}$$

- standard error of measurement = standard deviation of test scores * square root of 1 - reliability coefficient of the test

Standard Error Measurement

$$\sigma_{SEM} = \sigma\sqrt{1 - r_{xx}}$$

- standard error of measurement = standard deviation of test scores * square root of 1 - reliability coefficient of the test
- Can use this to create confidence intervals by using normality assumption of an individual's score on a large number of tests centered at the mean
- Determines the range of plausible values for a person's true score

SEM example

A math test is administered. The test scores have a reliability of 0.80 and a standard deviation of 0.5

What is the standard error of measurement?

If Anna scored a 7.5, what range of values can we be 95% confident that her true score lies between? 99% confident?

Standard Error of the difference between two scores

$$\sigma_D = \sqrt{\sigma_{SEM_1} + \sigma_{SEM_2}}$$

$$\sigma_D = \sigma\sqrt{2 - r_1 - r_2}$$

- Can be used to compare two individuals on the same test or a different test
- Can be used to compare performance of an individual on two tests

SED example

Sigrun takes the same test as Anna and scores a 6.5. Did Anna perform significantly better on the test?

If Anna took a second test and got a score of 8 and the reliability coefficient for the second test was 0.6, did Anna do significantly better on the second test?

Validity

Validity

- What is validity?
  - An indicator of how well the test measures the latent construct(s) it claims to.
  - A determination of the appropriateness of the test scores for specific uses/users
  - Validity of the test for a given purpose, at a given time, for a given population
  - You are a lawyer presenting evidence to a judge to make the case for the validity of your instrument - validation
  - Users can conduct a validation study to assess the validity of the instrument for their purposes

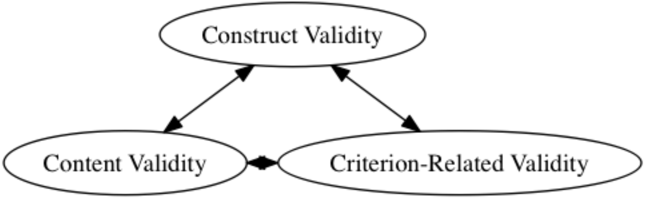
SAT

- The SAT is a standardized test measuring mathematics, reading, and writing
- Typically administered to 15, 16, and 17 year olds (sophomores, juniors, and seniors) in the USA
- Purpose to measure college readiness
  - def'n: College readiness benchmark associated with a 65% probability of earning a first-year GPA of 2.67 or higher.
- Schools within a city, within a state, and across states in the USA are quite diverse
- Would this test be valid for Iceland?
- Would this be appropriate for HÍ , HR, or UNAK?

Making the SAT valid for Iceland

- Could administer the test as it is or alter the test and conduct a local validation study
- Should translate it to Icelandic
- Update it to reflect Icelandic curriculum
- Age appropriate
- Is it for university-studies or menntaskóli?
- Anything else?

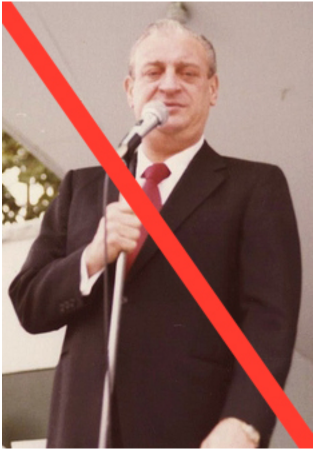
Types of Validity



Overview of Validity

- Content - Evaluation of subjects, topic, or content covered by the items in the test
- Criterion-Related - Evaluating the relationship of scores obtained on the test to scores on other tests or measures
- Construct - Evaluation relationship of scores obtained on the test to scores on other instruments measuring the same construct AND understanding how it fits within the theoretical framework of the latent construct

Face Validity is NOT Validity



source

Content Validity

- How adequately the test represents the latent construct of interest
- Do the items throughly and completely tap into the latent construct?
- Content valid test would have percentage of items on each topics to be proportional to the amount of time spent on these topics
- How can we be sure I am teaching the entire domain of psychological testing?
- Create a test blueprint
  - What could be conceivably measured and in what proportion
  - Number of questions, types of questions, areas covered, organization, etc

Assessing Content Validity

- Assume you are giving an instrument to measure aggressive behavior in children
- How can we assume this is measuring the construct of aggression quantitatively?
  - Experts assess whether each item is essential to the definition of aggression
  - $CVR = \frac{n_e - (N/2)}{N/2}$
  - Where n_e is number say "essential" and N is number of experts
  - Want this larger than chance (Table 6-1)

CVR in *R*

- ”Does your child bite other children?”
- 20 experts, 17 say “essential”

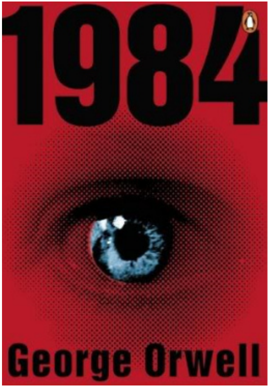
```r
CVR <- function(n, essential){
  (essential - n/2)/(n/2)
}
CVR(n = 20, essential = 17)

## [1] 0.7
```

Assessing Content Validity

BUT ... expert judgement!!!

“Who controls the past controls the future; who controls the present controls the past.”



source

Criterion-Related Validity

- What the test score tells you about where a person falls on the underlying construct being measured w.r.t a criterion
- A criterion is a benchmark or standard used for comparison
- Score high on an instrument measuring depression, *but do you really have depression*?
- Show no symptoms of depression, instrument is irrelevant and invalid

Measuring Depression

- Predict whether someone is receiving counseling services based on Beck Depression Inventory
- Find out BDI was used to determine whether someone should receive services
- What is wrong with this?

Forms of C-R Validity

- Concurrent Validity
  - Instrument provides the same “scores” as an already validated measure
  - Instruments must be administered at the same time (or nearly so)
  - Example?
- Predictive Validity
  - How well an instrument predicts some criterion in the future
  - SAT should measure “college success”
  - So it should be highly correlated with?
- validity coefficient: an “appropriate” measure of association

Validity Coefficient

- In summary, everything that affects the correlation coefficient!
- Range restriction from attrition in a study or self-selection
- Make sure testtakers are relevant in the validation study and cover the scope of the test!
- Read the test manual and make sure test is appropriate for your testtakers
- Coefficient should be high enough to matter

Incremental Validity

- Want to predict final grade in first math class in college.
- Add most important predictor first (maybe SAT math score if in the USA)
- Then add additional variables, incrementally, and see what each predictor adds
- This is akin to stepwise regression in multiple regression
- This is unwise because of inflation of type I error (the probability of incorrectly rejecting a null hypothesis when you should have retained it)

Construct Validity

- Evidence supporting that the test measures the underlying construct and that it can spread testtakers along that construct
- A test maker has theories about the construct, it's definition, structure, and relationship to other constructs and has theories about how their test relates to other tests
- All forms of validity are really subsumed within construct validity

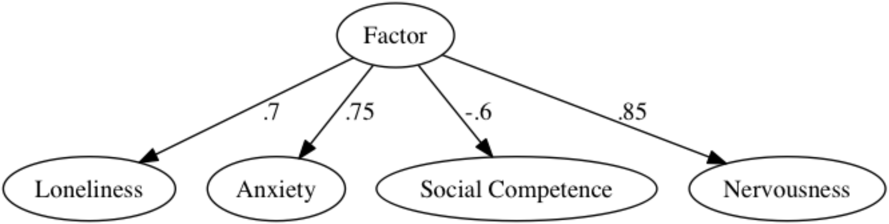
Types of Construct Validity

- Homogeniety
  - Structure of a test should be homogeneous if it is measuring a single construct
  - Responses to test items should be positively correlated with total score on the test
  - Items that are not need to be removed or rewritten
- Change with age and pre/post
  - Testtakers taking a test on algebra *should* score higher if they are older
  - Students getting tutored in algebra between a pre and post test should score higher on the post test

Types of Construct Validity

- Groups higher on the construct should have higher scores
  - Administer a test measuring tendency toward violent behavior
  - Higher scores on test: General population or prison inmates for assault and battery

Factor Analysis



- What should we call this factor?
- If Nervousness is our new instrument to measure the factor, how well does it do?
- What does it mean that social competence is negatively correlated with our factor?

Test Bias and Fairness

- Test bias - degree to which a test systematically favors one group or another
  - Can test for this statistically using logistic regression model
  - Known as differential item functioning
- Test fairness - the degree to which a test is fair and used in an equitable way
  - Administer a test to a group not involved in the validation sample
  - Maybe some groups of people are just different?