

# 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 reliability
- ▶ 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 perfect 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

$$\sigma_D = \sqrt{\sigma_{SEM_1}^2 + \sigma_{SEM_2}^2}$$

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

- ▶ 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



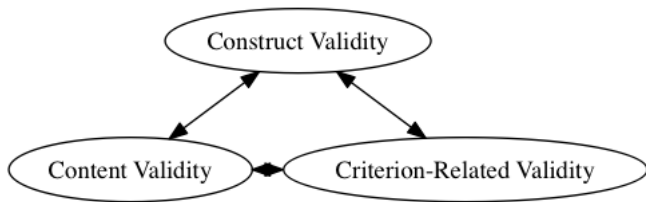
# SAT

- ▶ 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: How successful they are during their first year (often an measure of first year GPA).
- ▶ 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?

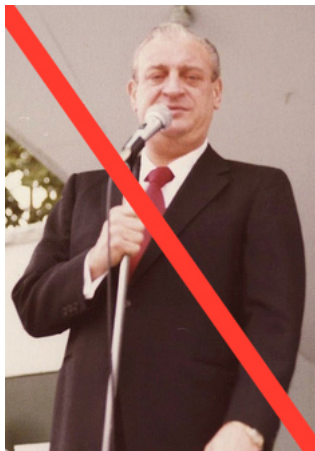
# Types of Validity



# 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 thoroughly 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, useful, or not necessary 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

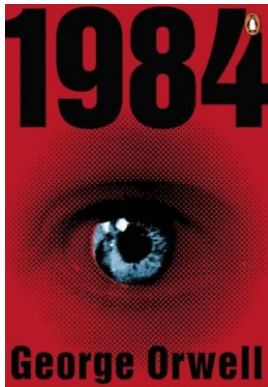
- ▶ "Does your child bite other children?"
- ▶ 20 experts, 17 say "essential"

```
CVR <- function(n, essential){  
  (essential - n/2)/(n/2)  
}  
CVR(n = 20, essential = 17)  
  
## [1] 0.7
```



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
- ▶ Scores on a new IQ instrument, *but do you really know that high scores mean high IQ?*
  - ▶ Should be **relevant**, e.g. people that are known to have high IQs (MENSA) should score highly on this instrument
  - ▶ Should be **valid** for measuring IQ, e.g. who created this instrument, does it correlate with established IQ instruments (e.g. WAIS or Stanford-Binet)?

# 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
- ▶ In addition, to self-report and parent report, you ask teachers to rate students on externalizing behaviors
  - ▶ After all the students' scores have been calculated, ask teachers to comment on them
- ▶ What is wrong with this?

# Concurrent Validity

- ▶ Concurrent Validity
  - ▶ Test scores are obtained at the *same time* as the criterion measures are obtained
  - ▶ Measures of the relationship between the test and the criterion are **concurrent validity evidence**
  - ▶ Example?
- ▶ If test scores (test new) correlate with a test (test old) that has already been validated to measure the criterion, then test old can be used as a **validating criterion**
- ▶ When might you do this?

# Predictive Validity

- ▶ Predictive Validity
  - ▶ Test scores are obtained *before* the criterion measures are obtained
  - ▶ How accurately does the test scores predict the criterion measures
    - ▶ SAT measures “college readiness”
    - ▶ What could be our future criterion?
    - ▶ What relationship would we expect between the scores and this criterion?
    - ▶ Could we use dropout (i.e. student attrition)?

# 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 your validation study and cover the scope of the test
- ▶ Read the test manual and make sure test is appropriate for your testtakers
  - ▶ Does their validity study map well to your target population and purpose?
- ▶ Coefficient should be high enough to matter

# Incremental Validity

- ▶ Refers to the degree to which an additional predictor explains the criterion measure above and beyond that already explained by those predictors already included
- ▶ Requirement: each predictor (obviously?) must have predictive validity
  - ▶ Let predict final grade in students in a statistics course
  - ▶ We have several variables to choose from:

```
## [1] "SECTION" "GENDER" "ETHDESCR" "CUM_GPA" "CUMCREDS" "ACT_TOTL"  
## [7] "ACT_ENGL" "ACT_MATH" "ACT_READ" "ACT_SCIR" "HSPR" "LTRGRADE"  
## [13] "STATGRAD" "DEVSTDNT"
```

- ▶ What should we do?



- ▶ For simplicity, let's just look at the continuous variables first.

##	SECTION	CUM_GPA	CUMCREDS	ACT_TOTL	ACT_ENGL
##	-0.009021592	0.491283854	0.250867602	0.252158752	0.188233466
##	ACT_MATH	ACT_READ	ACT_SCIR	HSPR	STATGRAD
##	0.293462052	0.167622795	0.167987726	0.239023226	1.000000000

- ▶ Which variable would you think is the strongest predictor of statistics grade?
- ▶ Which variables might have incremental validity?
- ▶ LET'S DO THIS TOGETHER!

# Expectancy tables

- ▶ Visualization tool
- ▶ Test scores (or applicant/client ratings) are obtained and placed into some interval (e.g. "excellent", "good", "ok", "bad", "miserable")
- ▶ Criterion measures obtained later (e.g. proficient in math or job performance)
- ▶ Create a chart that shows relationship between test scores and criterion measure
  - ▶ Essentially a contingency table
- ▶ A major omission from your book - we need to check and see if this is larger than chance alone!

## HR ratings and Job Performance

	Satisfactory	Unsatisfactory
Excellent	23	7
Average	12	10
Poor	12	13

```
## Error: geom_text requires the following missing aesthetics: y
```

```
##  
## Pearson's Chi-squared test  
##  
## data:  M  
## X-squared = 5.2582, df = 2, p-value = 0.07214
```

- ▶  $H_0$ : There is no association between HR rating and job performance

```
##  
## Pearson's Chi-squared test  
##  
## data:  M  
## X-squared = 5.2582, df = 2, p-value = 0.07214
```

- ▶  $H_0$ : There is no association between HR rating and job performance
- ▶ Probably need to intervene with HR!

# Construct Validity

- ▶ Evidence supporting that the test *measures* the underlying construct and that it is capable of *placing* test takers along that latent construct
- ▶ A test maker **MUST** have theories about the construct, its definition, structure, and relationship to other constructs and has theories about how their test relates to other tests
- ▶ If the test fails to discern test takers, need to know **why**
  - ▶ Recall all the various potential sources of error in testing
- ▶ All forms of validity could be considered subsets of construct validity

# Construct Validity Evidence

## ▶ Homogeneity

- ▶ 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
  - ▶ What kind of correlation is this?
  - ▶ Items that are not need to be removed or rewritten
  - ▶ What to do with items that have low correlations?
  - ▶ What does it mean to throw away items and rewrite them?
- ▶ Homogeneity implies inter item agreement ... how can we measure this?

## ▶ Change with age and pre/post

- ▶ Testtakers taking a test in reading *should* score higher on comprehension if they are older
- ▶ Students getting tutored in reading between a pre and post test should score higher on the post test
- ▶ Should we be able to predict how anxiety will change as we get older?

## Construct Validity Evidence - contd

- ▶ Groups higher on the measured construct should have higher scores (**method of contrasted groups**)
  - ▶ Administer a test measuring tendency toward violent behavior
  - ▶ Who should have higher scores: The general public or prison inmates for assault and battery?
- ▶ **Convergent**<sub>2-j</sub> - Test takers IQ scores on a new test should be correlated with their IQ score from an established and validated IQ tests (or a related construct)
- ▶ **Discriminant**<sub>3-j</sub> - Test scores should be unrelated to scores from another instrument



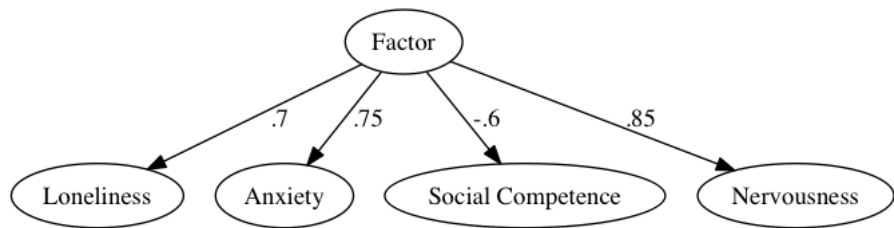
## Construct Validity Evidence - contd

- ▶ Groups higher on the measured construct should have higher scores (**method of contrasted groups**)
  - ▶ Administer a test measuring tendency toward violent behavior
  - ▶ Who should have higher scores: The general public or prison inmates for assault and battery?
- ▶ **Convergent**  $2-i$  - Test takers IQ scores on a new test should be correlated with their IQ score from an established and validated IQ tests (or a related construct)
- ▶ **Discriminant**  $3-i$  - Test scores should be unrelated to scores from another instrument

## Construct Validity Evidence - contd

- ▶ Groups higher on the measured construct should have higher scores (**method of contrasted groups**)
  - ▶ Administer a test measuring tendency toward violent behavior
  - ▶ Who should have higher scores: The general public or prison inmates for assault and battery?
- ▶ **Convergent**<sub>2-j</sub> - Test takers IQ scores on a new test should be correlated with their IQ score from an established and validated IQ tests (or a related construct)
- ▶ **Discriminant**<sub>3-j</sub> - Test scores should be unrelated to scores from another instrument
  - ▶ Ask students to score each other on leadership
  - ▶ Ask students to score each other on popularity
  - ▶ What does it mean if these two are uncorrelated?

# 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**
  - ▶ Errors by raters - too lightly, too severely, to the middle, too perfectly
- ▶ Test fairness - the degree to which a test is fair and used in an equitable way
  - ▶ What if we administer a test to a group not involved in the validation sample
  - ▶ Maybe some groups of people are just different?
- ▶ **Why do we care about bias and fairness?**