

# **E-411 PRMA**

## **LECTURE 7 - TEST DEVELOPMENT**

Christopher David Desjardins

# CONSTRUCT VALIDITY

- Evidence that the test *measures* the targeted construct and that it can *place* test takers along that construct
- Developer **MUST** have theories about the construct, it's def'n, structure, and relationship to other constructs and 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
  - Test should be homogeneous if measuring one construct
- Endorsing test items should be positively correlated with total test score
  - What kind of correlation is this?
- What should we do with unrelated items?
  - What does it mean to throw away/rewrite items?
- Homogeneity implies inter item agreement ... how can we measure this?

# CONSTRUCT VALIDITY EVIDENCE - CONTD

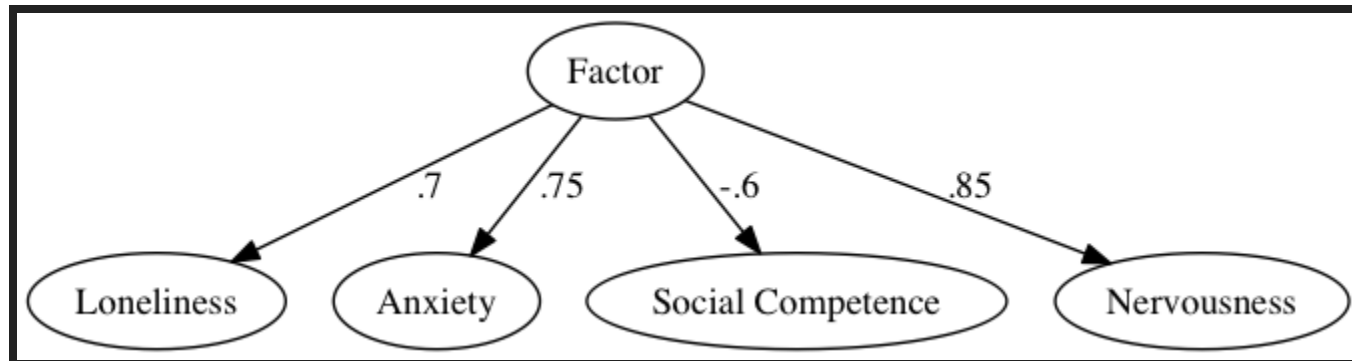
- Age and Pre/Post
  - Testtakers taking a test in reading should score higher on comprehension if they are *older*
    - Students getting tutored in reading between a pre and post test should score higher on the post test
- Can we predict how anxiety will change as we get older?

# MORE EVIDENCE

- Groups higher on the construct should have higher scores
  - **method of contrasted groups**
    - Measure tendency toward violent behavior
      - Higher scores: The general public or prison inmates for assault and battery?
- **Convergent** - Test scores on new IQ tests should correlate with scores from an established/validated IQ test (*or a related construct*)
- **Discriminant** - Test scores should be unrelated to scores from another instrument
  - Students rate each other on leadership and 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** - extent that a test systematically favors one group
  - Can test this statistically using logistic regression
- At item level: **differential item functioning**
- Errors by raters
- Test fairness - the degree 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
- **Why should we care about bias and fairness?**

# REVIEW

Classical Test Theory & Reliability

Test-Retest

Parallel Form

Internal Consistency

Quantifying uncertainty (i.e. standard error of measurement)



# REVIEW

Validity

Content

Criterion-Related

Construct

# TESTS, TEST, TESTS



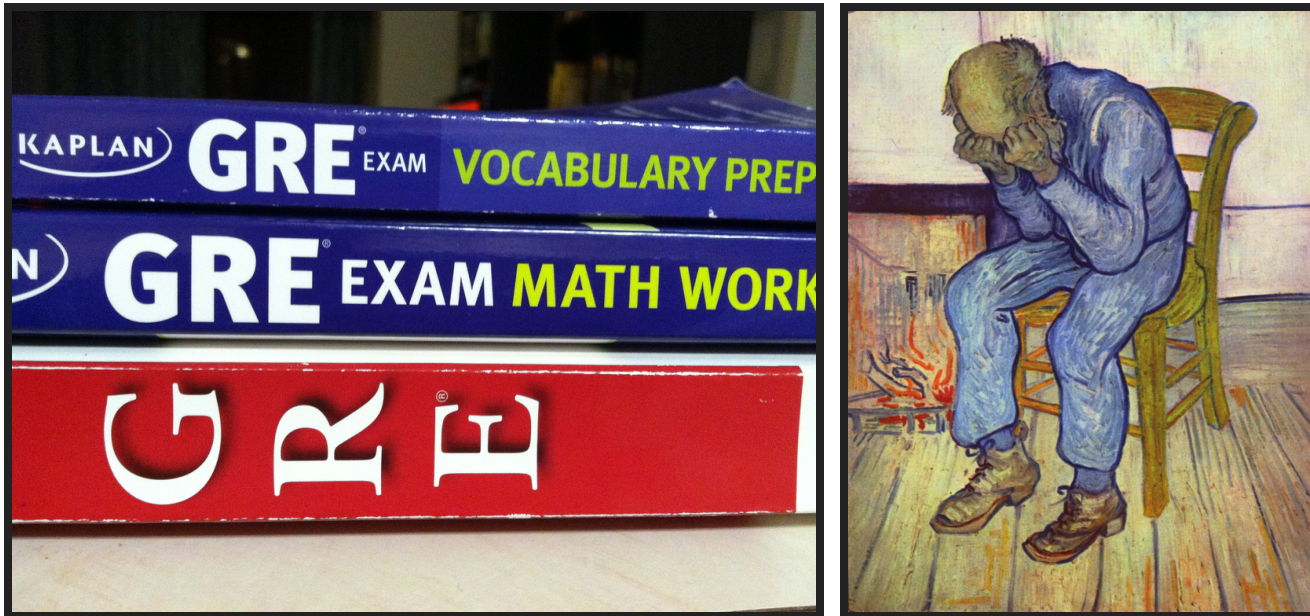
# TEST DEVELOPMENT

- Conceptualization
- Construction
- Piloting
- Item Analysis
- Revision

# TEST CONCEPTUALIZATION

- Identify a **need** for this test
- Identify a **purpose** for this test
- No test exists
- People aren't static, so tests need to be revised
- What else should we be thinking about when we're conceptualizing a test?

# NORM OR CRITERION-REFERENCED



Does it matter?

- Norm-Referenced
  - Want high scores on the test to get the item correct and low scores on the test to get the item incorrect
  - Want to spread out testtakers
- Criterion-Referenced
  - Want high scores on the test to get the item correct and low scores on the test to get the item incorrect
  - However, each item needs to measure whether a criterion is met

# TEST CONSTRUCTION

- Now that we know why, we have to know **how**
- We need a set of rules for assigning numbers in measurement - **scaling**
- In psychology, scales are instruments used to measure traits, states, or abilities

# SCALES





# TYPES OF SCALES

- Nominal, ordinal, interval, or ratio
- Examples?

# RATING SCALES

- Testtaker indicates their response to an item by selecting among strengths
- Examples: Stealing
- Likert-Type are common rating scales
- Scores from test could be summed (summative) directly; factor analysis or item response theory could be used

# SCALE ISSUES

---

*“Downloading movies is the same as stealing”*

---

Strongly Agree   Agree   Neither Agree/Nor Disagree   Disagree   Strongly Disagree

Are the distances the same between the choices?

What might affect are choices?

# MORE SCALES

- **Paired comparisons** - choose between two options scored based on some criteria
- **Comparative scaling** - items are arranged based on some criteria and **categorical scaling** - items into two or more categories
- **Guttman scale** - items written in a sequential manner such that someone higher on the trait will agree with the strongest statements through the mildest statements

# **LET'S WRITE SOME ITEMS THAT USE THESE SCALES**

Let's brain storm 10 nouns

# INTERVAL SCALES

- Could use Thurstone's equal-appearing scale (p 250 - 251)
- I am skeptical ... what do you think?

# LET'S WRITE A TEST!

$$\vec{\nabla} \cdot \vec{D} = \rho$$

$$\vec{\nabla} \times \vec{E} = - \frac{\partial \vec{B}}{\partial t}$$

$$\vec{\nabla} \cdot \vec{B} = 0$$

# WRITING ITEMS

- What content should the items **cover**?
- What should the **format** of the items be?
- How **many** items should be written and **for each content area**?
- Book recommends writing 2x the number of items for the **item bank/pool** . . . **seems a bit excessive**



# TYPES OF ITEMS

Selected-response vs constructed response



**Help improve this page** [What's this?](#)

Did you find what you were looking for?

↓

**< Great. Would you like to add anything else?** [What's this?](#)

How could this article be improved?

Please post [helpful feedback](#). By posting, you agree to transparency under [these terms](#).

# SELECTED-RESPONSE

- Types
  - Multiple-choice
  - Binary-choice
  - Matching
- Each item will have a stem, correct choice, and distractors

- A good multiple-choice item in achievement test
  - Only one correct choice
  - Grammatically parallel alternatives
  - Alternatives of similar length
  - Alternatives that fit grammatically with the stem
  - Include as much information in the stem as possible to avoid repetition
  - Avoids ridiculous distractors
  - Is not excessively long

# FINAL THOUGHTS ON SELECTED-RESPONSE

- There are more than just true/false for binary-choice items
- Matching bank should have more answers choices than items and/or be used more than once
- Guessing is a problem in an achievement setting
- Always forcing a choice in a non-achievement setting

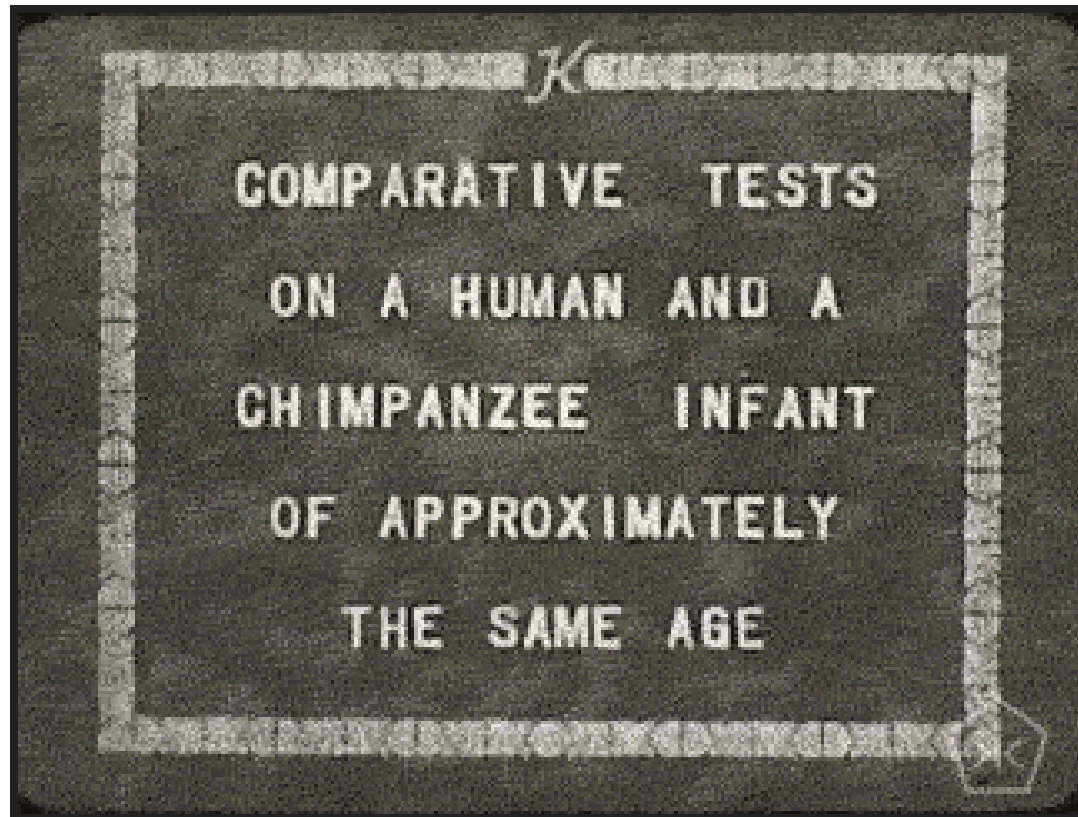
# CONSTRUCTED RESPONSE

- **Completion items** are fill-in-the blank responses
- **Short-answer** items require a response of a few sentences
- **Essay** items are long short-answer items demonstrating deeper, more thorough knowledge
- More deeply probe a specific portion of a construct, require more time
- Subjectivity in scoring essays
- **What reliability statistic would we report here?**

# SCORING THE ITEMS

- **Cumulative model** - sum up the items on the test
- **Class scoring** - based on pattern of responses placed with similar testtakers
- **Ipsative scoring** - score on a scale within a test compared to score on another scale on same test
  - Edwards Personal Preference Schedule - measures relative strength of different psychological needs
- Could look at both the cumulative scores on separate scales and the pattern of these scores, **profile analysis**

# PILOTING THE TEST



# ITEM ANALYSIS

- Many different ways to analyze items
- Can focus on
  - Difficulty of item
  - Reliability of item
  - Validity of item
  - Discrimination of item



# ITEM DIFFICULTY

- Proportion of testtakers that get the item correct
- Higher the item difficulty, the easier the item
  - item-endorsement index
- Can calculate average item difficulty for the test
- Optimal value =  $\frac{\text{Pr}(\text{Guess})+1}{2}$

# ITEM DIFFICULTY - EXAMPLE

Administer an item to 10 students and 4 students get the item correct

What is the item's difficulty?

If the item was a multiple choice with 5 distractors, what is the optimal item difficulty?

# ITEM RELIABILITY

- Internal consistency of the test
- Software often calculates changes in a reliability index (e.g. coefficient alpha) when item is deleted
- Examine factor loadings
- Calculate item-reliability index =  $s_i * r_{i,\text{ttscore}}$ 
  - $s_i$ , the standard deviation of item  $i$
  - $r_{i,\text{ttscore}}$ , correlation between item  $i$  and total test score

# ITEM RELIABILITY INDEX - EXAMPLE

Item 1
1
0
0
1
1

Assume correlation between item 1 and total test score is 0.7

# ITEM VALIDITY

- Item-validity index =  $s_i * r_{i,\text{crit}}$ 
  - $s_i$ , the standard deviation of item  $i$
  - $r_{i,\text{crit}}$ , correlation between item  $i$  and criterion measure

# ITEM DISCRIMINATION

- **Point-biserial correlations** - Are testtakers with higher abilities more likely to get the item correct?
- IRT's discrimination parameter
- **Item discrimination index**
  1. Discretize total test scores into upper and lower 27%
  2. Calculate number of "high" scores that got item correct and number of "low" scores that got item correct
  3. Calculate difference
- **Examine distractor functioning**

Example in R

# ISSUES IN TEST DEVELOPMENT

Guessing

Bias in favor of one group - differential item functioning

Test length and duration of testing session



# ALTERNATIVES TO ITEM ANALYSIS

Think Alouds

Expert Panels

Interviews

Qualitative Methods

# TEST REVISION

- On what basis should we revise our items?
- Too easy or too hard items?
- Items with similar difficulty that are measuring the same concept?
- Items with negative point-biserial correlations?
- Items that on a second/third read through seem unrelated to the construct?
- Items with low factor loadings?
- Based on IRT?

# STANDARDIZATION

- We settle on our revisions
- Administer revised version to new sample
- This becomes our comparison group, our **standardization sample**

# REVISING OLD TESTS

- Tests need to be revised when the domain has significantly changed
- Content of the items is not understood or changed
- Test norms are no longer adequate
- Theory underlying the domain has changed
- Reliability and validity of the instrument can be improved

# CROSS- AND CO-VALIDATION

- **Cross-validation** - revalidation of a test on a separate, independent sample of testtakers
- Item validities should shrink during this process (**validity shrinkage**)
- **Co-validation** - test validation conducted on two or more tests with the same sample of testtakers
- Creating norms, **co-norming**
- Cheaper, reduces sampling error by norming on the same sample