E-411 PRMA

WEEK 4 - TEST DEVELOPMENT

Christopher David Desjardins

REVIEW

Classical Test Theory & Reliability

Test-Retest

Parallel Form

Internal Consistency

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

REVIEW

Validity

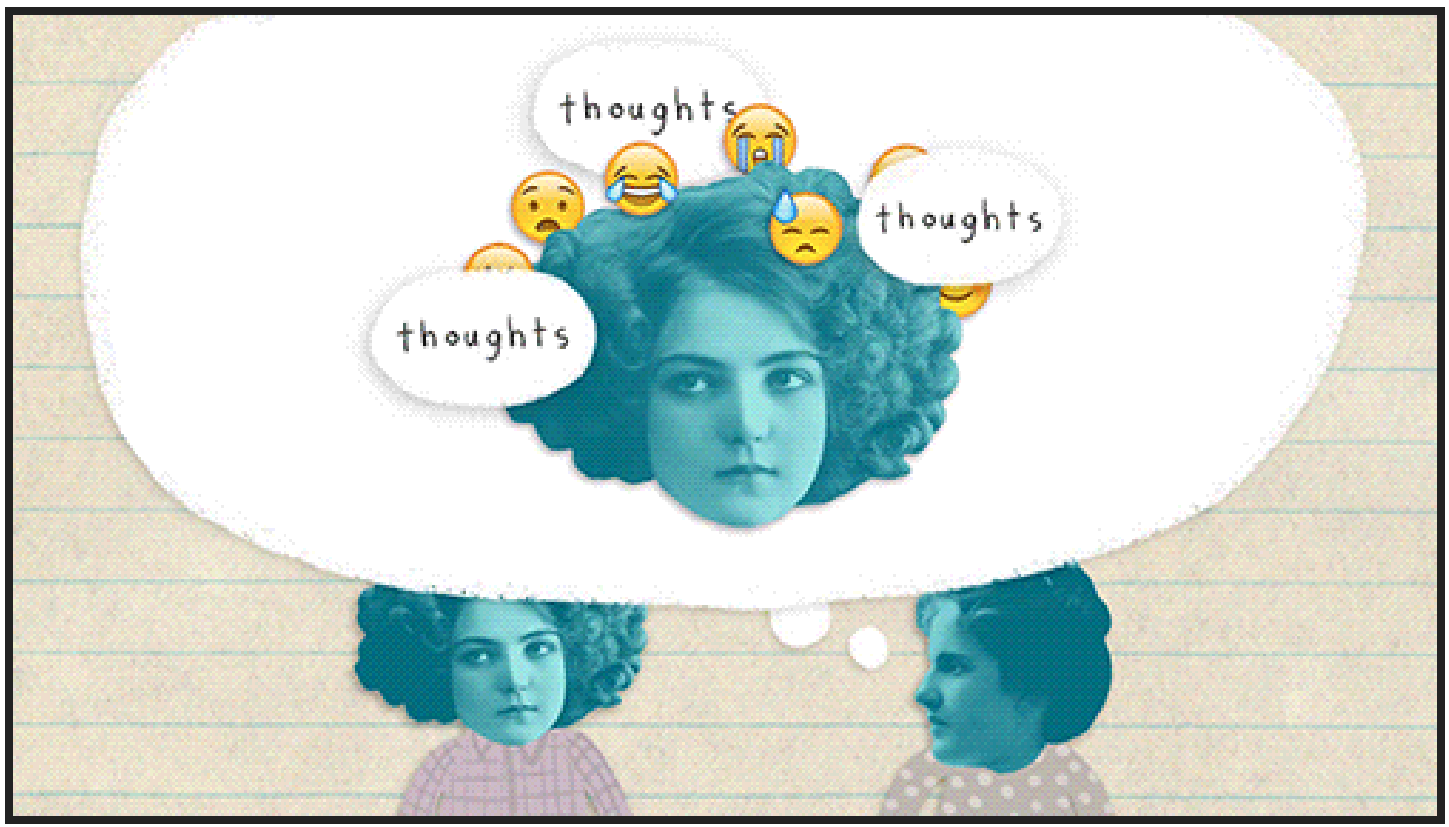
Content

Criterion-Related

Construct

Quiz next time!

TESTS, TEST, TESTS



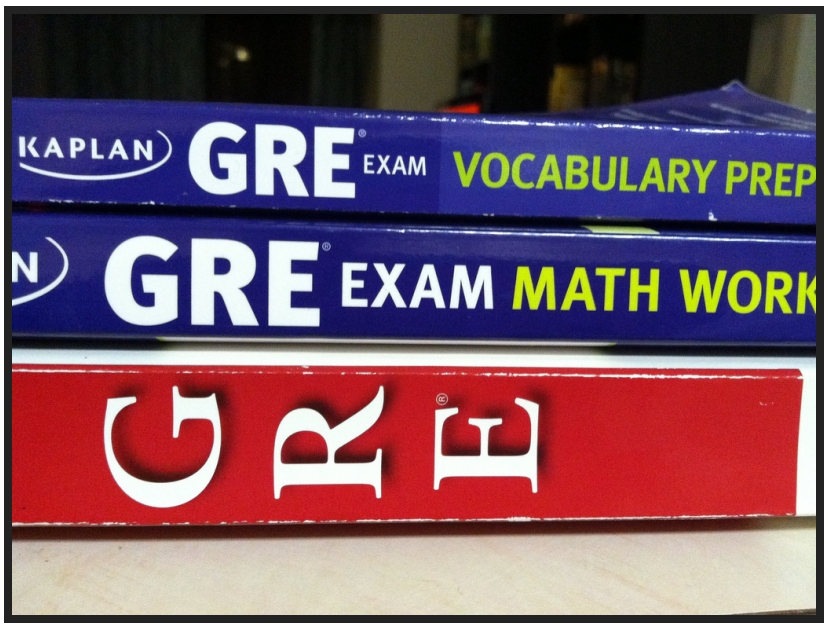
TEST DEVELOPMENT

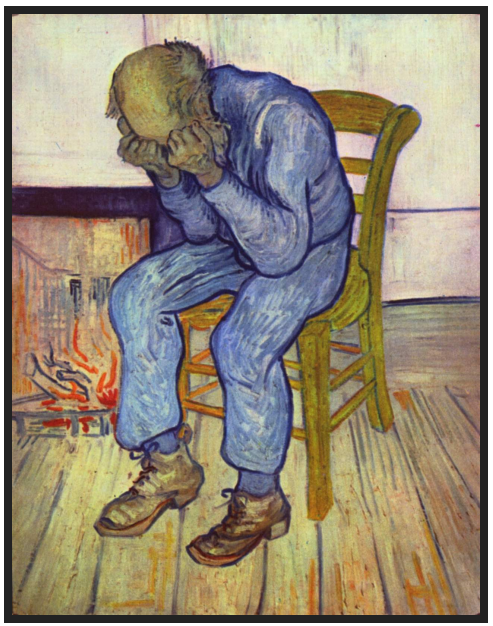
- Conceptualization
- Construction
- Piloting
- Item Analysis
- Revision

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 should we be thinking about when we're conceptualizing a test?

NORM OR CRITERION-REFERENCED





Does it matter?

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 directly or 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 betwen the choices?

What might affect this?

MORE SCALES

- Paired comparsions - 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 statments

INTERVAL SCALES

- Could use Thurstone's equal-appearing scale (p 250 - 251)
- I am skeptical ...

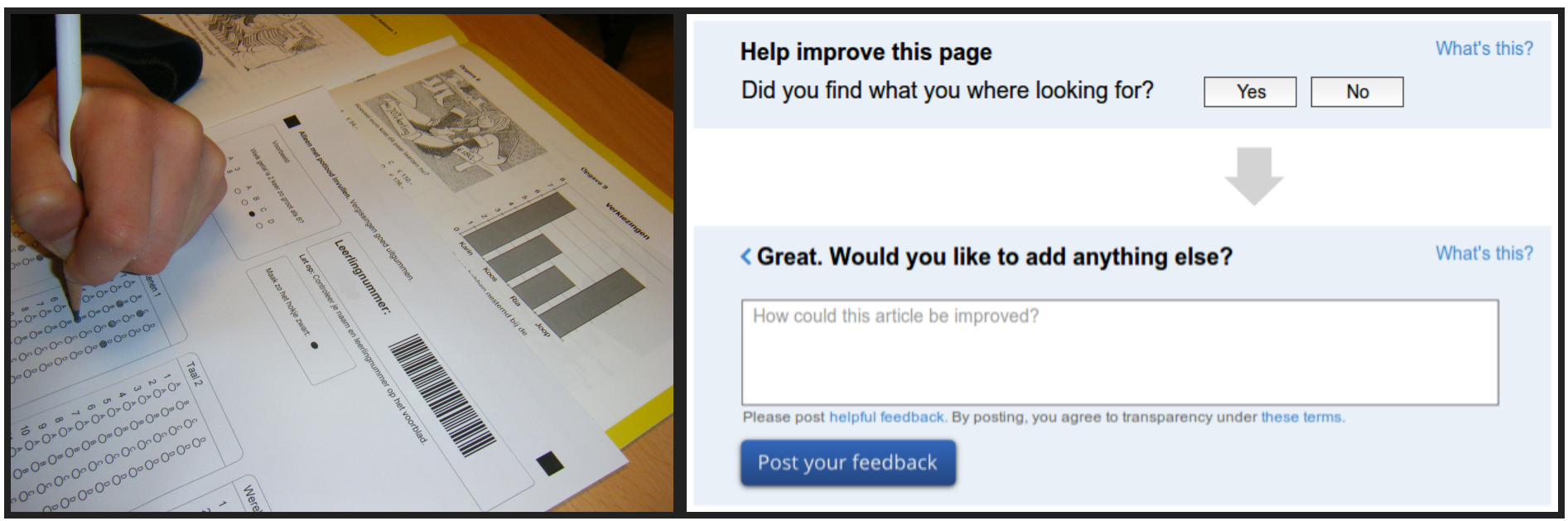
LET'S WRITE A MATH TEST!

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



SELECTED-RESPONSE

- Types
  - Multiple-choice
  - Binary-choice
  - Matching
- Each item will have a stem, correct chioce, 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 repetiiton
  - 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

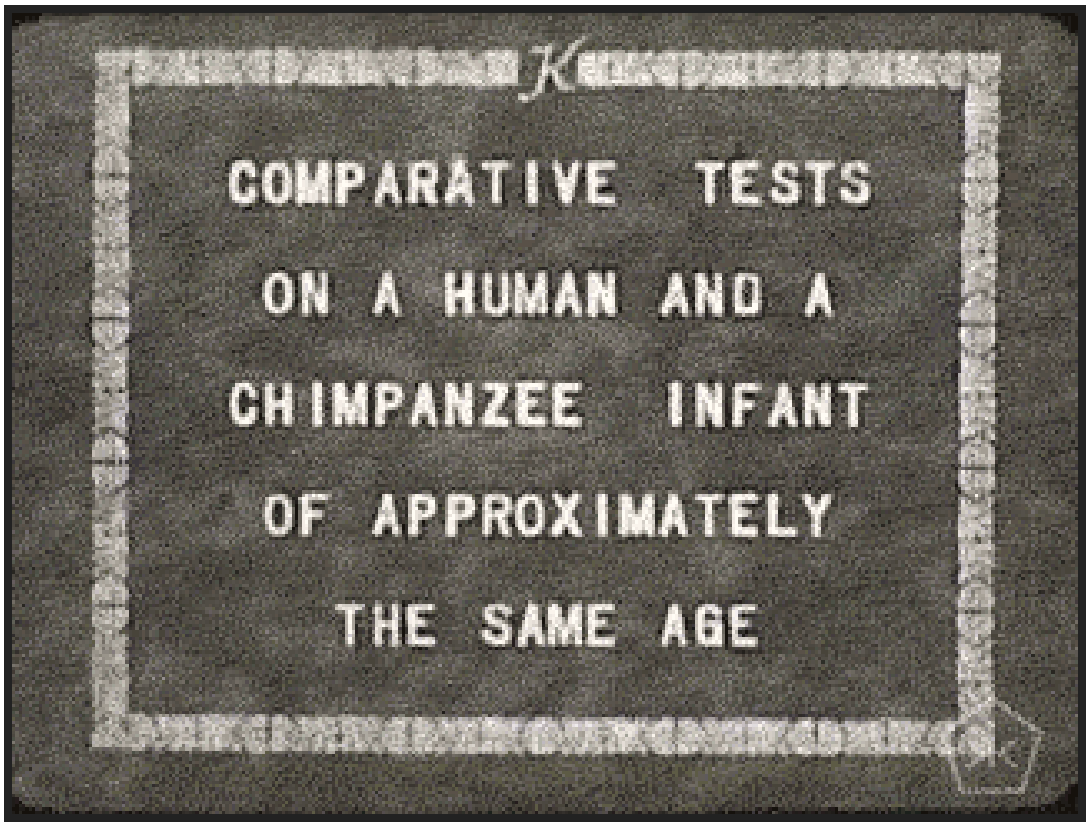
CONSTRUCTED RESPONSE

- 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 based on comparsion to score on another scale on the test
- Edwards Personal Preference Schedule - measures relative strength of different psychological needs
- Could look at both the cumlative scores on seperate 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 is 0.5 or $\frac{\Pr(\text{Guess})+1}{2}$

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

ITEM VALIDITY

- Item-validity index - product of standard deviation of the test item and the correlation of the test item with the total score

ITEM DISCRIMINATION

- Point-biserial correlations - Are testtakers with higher abilities on the construct more likely to get the item correct?
- Item response theory's discrimination parameter
- Item discrimination index in the book (why discretize?)
- Examine distractor functioning

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 comparsion 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 seperate, independent sample of testtakers
- Item validities should shrinks 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