Ch. 6: Test Development

Review

- Reliability
  - kinds
  - r or r^2 - what is "good enough"
- Validity
  - kinds
  - r or r^2 - what is "good enough"

- Chapter 6
  - Writing test items w/good reliability + validity
  - Evaluating test item quality

4 kinds of Reliability

	Description	Name	Statistic
Time Sampling	1 test given two times	test-retest reliability	correlation between scores at two times
Item Sampling	2 different tests given once	Alternate or Parallel forms	correlation between scores on 2 versions
Internal Consistency	One test, multiple items	Split Half or internal reliability	Cronbach's Alpha
Observer Differences	One test w/ 2+ observers	inter-observer reliability	Kappa

4 Kinds of Validity

	Description	Notes	Statistic
Face	do items "look" valid?	informal, improper, non-scientific	none
Content	do test questions cover the topic?	logic & judgement - there are no stats to calculate	none
Criterion	does the test predict a specific event?	requires a well-defined criteria	Pearson's R (correlation) between Test and Criteria
Construct	does the test measure what it claims	modern theory: all validity is Construct validity	Convergent and Divergent correlations (Pearson's R)

Ch. 6: Test Development

- Test Items
  - question formats (T/F, Multiple Choice, Likert…)
- Correction for guessing formulas
- Cognitive Factors: Recall vs. Recognition
- Exercise: from construct to question
- Item Analysis: Difficulty, Discriminability, ICC
- Item Response Theory / Adaptive Testing
- SII (Strong Interest Inventory)

Writing test items...

- Define what you are measuring (theory of the construct)
- Write many items that cover the *content*
- Avoid very long items
- Use appropriate reading level
- Don't mix two concepts in one question.
- Vary the "response set" with both positively and negatively worded items

Test Item Formats

- Fill in the blank
- Essay
- True / False…
- Multiple Choice…
- Rating / Category scales…

Dichotomous Format

- Aka "True/False" or "Yes/No" or "Binary"
- Pros: easy to write, administer, and score, good for basic facts. Avoids ambivalence.
- Cons: rote memorization, high scores due to guessing —> increased # of items, black & white thinking: not appropriate for complexity or nuance
- Summary: unsophisticated format that should not be widely used for achievement testing

Poly[cho]tomous

- AKA "multiple choice"
- Target: correct answer
- Distractor: incorrect answers
- Pros: easy to administer (covers a lot of material quickly), easy to score, can handle shades of gray / nuance
- Cons: difficult to write, susceptible to guessing strategies, susceptible to "over studying"

Distractors?

- Too few distractors --> dichotomous
- Too many distractors --> slow, confusing
- Optimal is 3-5 distractors. Thus, most multiple-choice tests should have between 4 and 6 possible answers per question.
- Distractors should cover a wide range of abilities w/o being cute or trite

Guessing : Probability

- M = # of answer choices per question
- $P_{correct}$ with random guessing = 1/M
- On a dichotomous (T/F), P = ____
- On a multiple choice test with M answers per question, the probability = _____
- Total score from guessing:
  - $N_{questions}$ x $P_{correct}$

Guessing : Expected Score

- Probability of getting any item correct, using a random guessing strategy, p is equal to 1 divided by the # of answers.
- On a dichotomous (T/F) test the probability P = 1/2 = 50% = 0.5
- On a multiple choice test with M answers per question, the probability = 1 / M. For a 4 item test P= 1/4 = .25 = 25%
- Total score due to guessing = # of questions times average score per item or N * P.
- Example: an 100 item test with 4 answers = 25

Correcting for Guessing

- Scores can correct for guessing.
- Goal: person randomly answering should get same score as someone who doesn't answer.
- Expected score of someone who answers no question = zero
- Expected score of someone who guesses randomly is N* (1/M)
- For every wrong answer, subtract 1/(M-1) points.

Correcting for Guessing : Example

- Example:
  - a 100 item test (N=100)
  - each question has 5 choices (M=5)
  - probability of right answer by guess? (P = 1/M = 1/5 = 20%)
- A student takes the test, guesses on each item, and gets 20 correct (P*N = 0.2 * 100 = 20)
- Correction for guessing subtracts (1/M-1) points for each wrong answer = 1/(5-1) = 1/4 = 0.25 points.
- Adjusted score?

Correcting for Guessing - Real World

- Formula is simplistic
- College Board removed guessing penalty for AP exams in 2010
- SAT revisions in March 2016
  - Removes penalty for Guessing
  - other changes:
    - Essay is optional
    - Vocabulary test changed

When should you guess?

- Almost always
- Worst case: if a correction formula is in use, and you truly have zero information for a given item, guessing gains you nothing
- However, chances are that you actually have some knowledge. This increases your chances slightly above chance, giving you a positive expected score.

[di|poly]chotomous Issues

- Pros:
  - neutral, fair scoring
- Types of knowledge:
  - Recall vs. Recognition
  - Receptive vs. Expressive
- Skill =? test taking ability
- Solution: Essay test format

Accessing Knowledge

- Recalling information is different than Recognizing it
- Neuropsychology suggests different brain systems. Recall can be stronger or weaker than Recognition
- Issues for testing:
  - What type of access is involved in polychotomous testing?
  - Is it fair to test using a method which prefers one type over the other?

Recall vs. Recognition

Other question formats

- Likert Scale
- Category Rating Scale
- Visual Analogue Scale
- Q-Sorts
- Checklists

Likert Format

- Asked to rate statements on a scale with a small fixed number of answers
- Example:
  I am afraid of heights:
  1 strongly disagree
  2 disagree
  3 undecided
  4 agree
  5 strongly agree
- Numbers : sometimes shown, sometimes not shown.

Likert : Neutral?

- Sometimes, want to avoid the middle (neutral, undecided) answer
- Example:
- I am afraid of heights:
  1 strongly disagree
  2 somewhat disagree
  3 somewhat agree
  4 strongly agree
- Like T/F, forces subject to take a position

Likert : Balance & Symmetry

- Answers should be balanced & symmetrical
- Example:
- I am afraid of heights:
  1 strongly disagree
  2 somewhat disagree
  3 neutral
  4 somewhat agree
- Poor design
  - Answers will be biased towards 3 or 4

Category (Rating Scale) Format

- Similar to Likert format, but #s are used instead
- Pros -- responses are more precise than with Likert scales (10 vs. 5 or 6)
- Cons -- context effects stronger
  - Solution: clearly define endpoints
- Precision vs. Accuracy?

Category Example

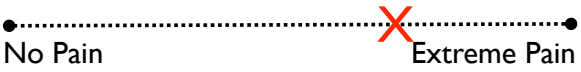
- On a 1 to 10 scale how much do you like your partner?
  1 Planning to break up
  2
  3
  4
  5
  6
  7
  8
  9
  10 Planning to get Married soon
- Issues:
  - Unbalanced (is 5 or 6 the middle?)
  - Hard to interpret : what does a "2" or "3" really mean?

How many choices?

- Research suggests optimal # of choices is between 4 and 7
- Using up to 10 choices is OK if
  - raters are motivated
  - good anchors & examples are giving
  - Otherwise, 10 choices leads to random responding

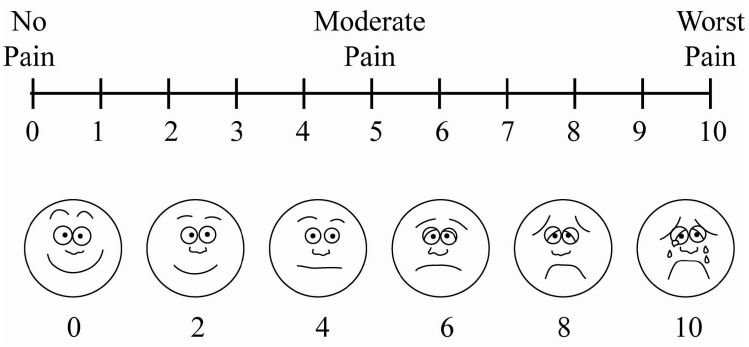
Visual Analogue Scale

- Similar to Category format, except use of a visual stimulus & graphical measurement
- Example:
  How much pain are you in right now?



- Pros: allows a precise, finely detailed response
  Cons: hard to score, precision vs. accuracy?

Visual Analogue Scale



Checklists

- Checklists:
  - Agree/disagree with large # of statements
- Example

- "I am currently having trouble with…"
  - ☐ Money
  - ☐ Relationships
  - ☐ Appetite
  - ☐ Sleep
  - ☐ …

Q sorts

- Q sort:
  - sort large # of statements into piles depending on how much you agree/ disagree (like Likert format)
  - Responses follow bell-shaped curve, extreme responses are most interesting

Advice from Textbooks

Advice	% endorsing
Don't use "All of the above"	80%
Don't use "None of the Above"	75%
All choices should be plausible	70%
Negative wording shouldn't not be un-used	55%

Review

- Reliability and Validity of entire Test
- Individual Test Items
  - dichotomous / polychotomous
  - recall vs. recognition
  - Likert
    - neutral, balanced
  - Category
    - anchors, context effects
  - Ideal # of answers per question?

Item Analysis

- In Ch 5 we discussed the reliability and validity *of the entire test*.
- Now we look at psychometrics of *individual test items*.
- Item Difficulty
- Item Discriminability

Item Difficulty

- How hard is this item?
- % who get the item correct (item easiness)

Too hard / Too easy

- Floor effect: many scores near the bottom range of possible scores
- Ceiling effect: many scores near the top range of possible scores

Item Difficulty

- How hard is this item?
- % who get the item correct (item easiness)
- Ideal= halfway between chance and perfect
  - for a 4-item multiple choice, chance = 25%, so optimum would be 62.5%
  - typical range is 30% to 70%
- Overall test should have wide variety of item difficulty because people are different

Item Difficulty 2

- Mathematically, 30%-70% is optimum
- What about human / emotional issues?
  - Tests or items that are too hard?
  - Tests or items that are too easy?

Discriminability

- Difficulty = how many people answer correctly?
- Discriminability = who answers correctly?
- Does performance on one item correlate with overall test performance?

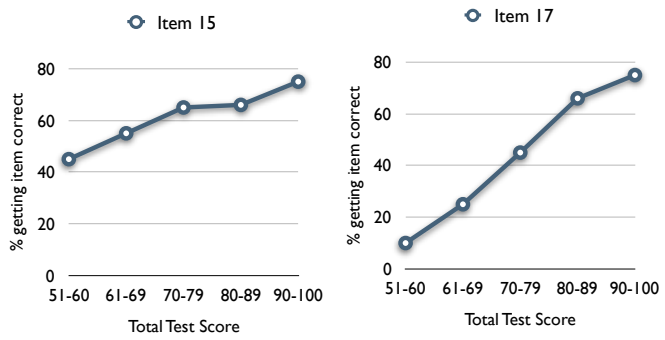
- Two ways
  - statistical
  - graphical

Discriminability - Statistical

- Extreme Group:
  - divide test takers into thirds
  - % correct : top third vs. bottom third
- Point Biserial
  - p.b. correlation between item and test score
  - low or negative values represent "bad" items

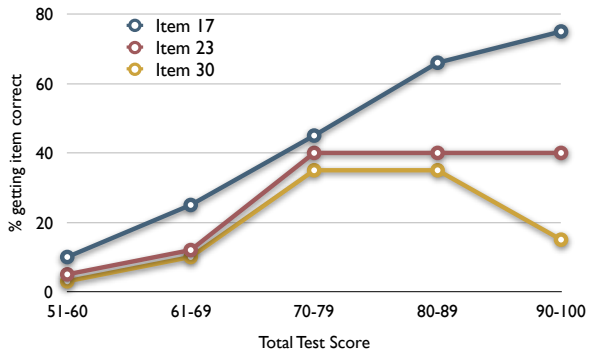
Discriminability - Graphical

- Item Characteristic Curve
- Graph % correct vs. total test score for one test item



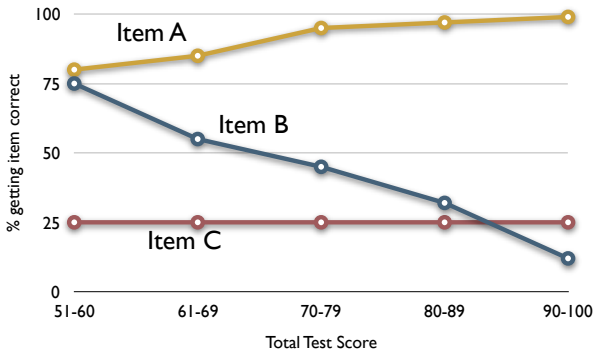
Item Characteristic Curve

- Good items show steady increase
- Bad items show decreases or flat spots



ICC Example

- Diagnose these problems:



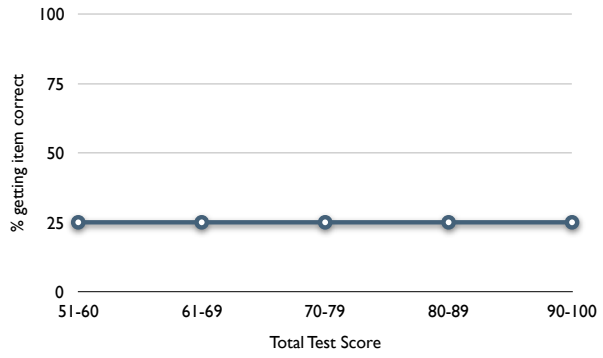
Graph the ICC

- Item 1: What was the exact population of the town Bodie, California, in 1879?
  (A) 6142
  (B) 6143
  (C) 6144
  (D) 6145

- Correct answer = A

ICC Example

- Random guessing



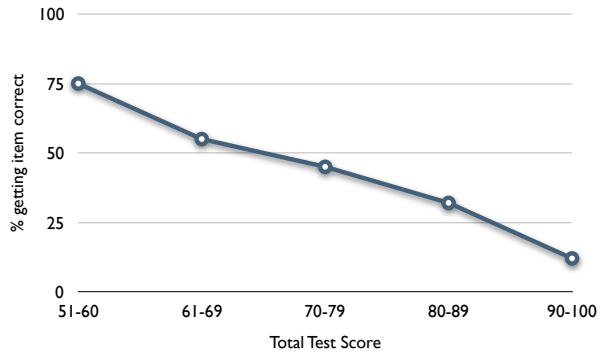
Graph the ICC

- Item 1: What is 0.34 times 0.27
  (A) 9.18
  (B) 0.61
  (C) 0.0918
  (D) 91.8

- "Correct Answer" = B

ICC Example

- Test item has wrong answer



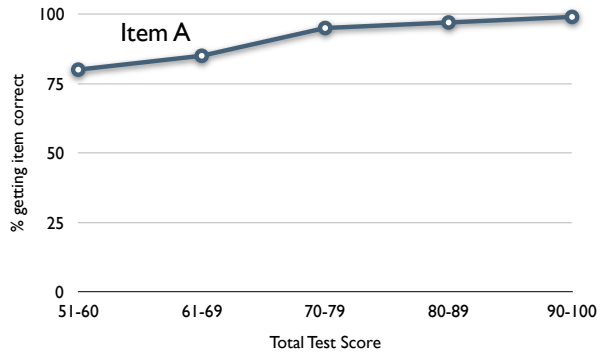
Graph the ICC

- Item 1: What is 1 + 2
  (A) 11
  (B) 21
  (C) 3
  (D) 0.3

- Correct answer = C

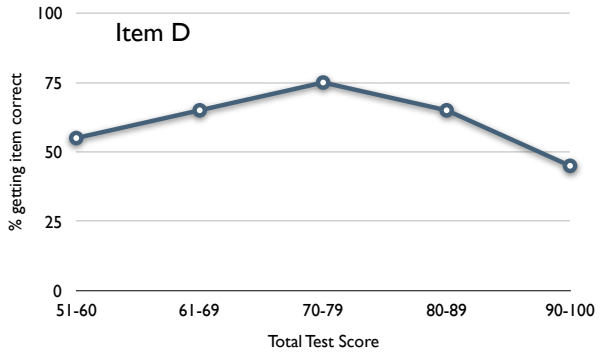
ICC Example

- Item is too easy



ICC Example

- "Overstudying" or "None of the above



Item Response Theory (IRT)

- Classical Test theory
  - your ability = *number of items correct*
- IRT
  - your ability = *level of difficulty* at which you can perform
- IRT Model : probability of correct answer is modeled using several variables (for the test and the test-taker)
- IRT Procedures: using computer-based *adaptive testing*

IRT / Adaptive Testing

- To cover a range of ability levels, tests must have a range of item difficulties
- For one individual, therefore many items are much too easy and much too hard
- "old fashioned" solution = have many tests, choose right one based on pre-existing knowledge of person.
- IRT solution = one test that automatically detects person's level and gives questions mainly in that difficulty level.

IRT in the real world

- IRT is theoretically better
- Adoption in curriculum is slow
- some tests use it but vast majority do not

- Continuing research

External Criteria

- Internal Criteria = total test score
- External Criteria = thing that actually matters (e.g. "do you crash the plane")

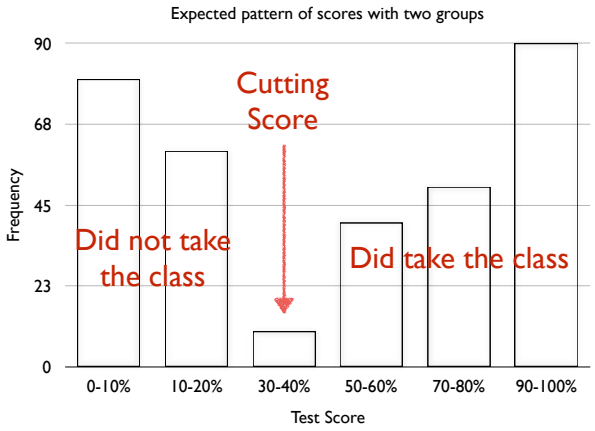
- Most Item Analysis still uses Internal criteria rather than the more correct External Criteria
- Why?

Criterion-referenced Test

- Instead of arbitrary criteria such as "70% = pass" use one with more validity.

- Criteria = the learning outcome(s) desired
- Method:
  - create a good test
  - give it to two groups of students
    - those who have had the material
    - those who have not
  - Determine cut-point score from histogram

Criterion-referenced Test



Limitations of Item Analysis

- Tests discriminate between levels of performance
- Statistics (difficulty and discriminability) don't tell why a person missed an item
- Items might discriminate well (statistically) but for the wrong reasons (educationally)
- Tests don't directly help people learn
- Tests can harm, if they dramatically change learning behavior (e.g. study for the test rather than the subject)

Example of a poor test item?

- What is 0.4 plus 0.3
  (A) 0.3
  (B) 0.4
  (C) 0.7
  (D) .07

- Is answering (A) better or worse than answering (D)?