Ch. 6: Test Development

- Goals of this chapter
- Test Items
  - Common formats
  - Alternative formats
- Item Analysis
  - Item Difficulty
  - Discriminability
- Item Response Theory

Ch. 6: Goals

- Understand several test item formats
- When to guess on multiple-choice exams, how to score exams to correct for guessing
- Understand rating scales (Likert, 10 point, etc.)
- Measure and adjust item Difficulty
- Measure and adjust item Discriminability
- Item Characteristic Curves
- Describe the "over studying" problem
- Understand limitations of item analysis / item response theory.

Test Item Formats

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

Writing test items...

- Define what you are measuring using a theoretical framework (aka "Construct")
- Write a large pool of items that cover the content area without duplication
- Avoid very long items
- Use a reading level difficulty appropriate for the test takers
- Avoid complexity -- don't mix two concepts in one question.
- Vary the "response set" with both positively and negatively worded items

Dichotomous Format

- Aka "True/False" or "Yes/No" test
- Pros: easy to write, easy to administer, easy to score, appropriate for statements of objective facts. Avoids ambivalence.
- Cons: encourages rote memorization, high scores due to guessing require increased # of items, punishes complexity or nuanced thinking, not appropriate for value judgements / shades of gray
- Summary: a somewhat unsophisticated format that should not be widely used for achievement testing, but OK for personality tests.

Poly[cho]tomous

- Aka multiple choice
- Target: correct answer
- Distractor: incorrect answers
- Pros: easy to administer (can cover a lot of material quickly as compared to essay test), easy to score, can handle shades of gray or discriminate finer nuances of meaning
- Cons: difficult to write, susceptible to guessing strategies, susceptible to "over studying"

Distractors?

- Too few distractors --> dichotomous
- Too many distractors --> slow, confusing
- Studies suggest optimal # is around 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 : Expected Score

- Probability of getting any item correct, using a random guessing strategy, is equal to 1 divided by the # of answers.
- On a dichotomous (T/F) test the probability = _____
- On a multiple choice test with M answers per question, the probability = _____
- Total score due to guessing = # of questions times average score per item or _____

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 10 item test with 4 answers = 2.5

Correcting for Guessing

- Scores can correct for guessing.
- Goal is to equalize the scores of someone who guesses randomly with 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)
- Formula - for every wrong answer, subtract (1/M) points.
- Problems?

When should you guess?

- 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
- Neuroscience/Neuropsychology suggests the two are mediated by different brain systems. Recall can be impaired but not Recognition (and vice versa)
- 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

Recall vs. Recognition

- Remember these numbers:
  - 134592618
  - 214577131

Recall vs. Recognition

- Recall both numbers now

Recall vs. Recognition

- Which of these numbers were you asked to remember?
  - 021418321
  - 134592618
  - 214577131
  - 213011764
  - 138363732

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 somewhat disagree
  3 neutral
  4 somewhat 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 : Balanced?

- Avoid un-balanced formats with even # of choices if there's a neutral answer
- 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 Format

- Similar to Likert format, but #s are used instead
- Example:

  On a 1 to 10 scale (with 1 as the lowest and 10 as the highest) how much do you like your partner?

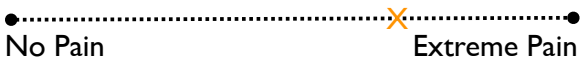
- Pros -- responses are more detailed than with Likert scales (10 vs. 5 or 6)
- Cons -- context effects stronger
  - Solution: clearly define endpoints
- Precision vs. Accuracy?

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?

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?

Checklist and Q sorts

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

- 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	% of textbooks 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%

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?
- Expressed as % who get the item correct (perhaps better called "item easiness"?)
- How hard should an item be? Ideal is halfway between chance-level performance and 100%
  - e.g. for a 4-item multiple choice, chance = 25%, so optimum would be 62.5%
  - typical range is 30% to 70%
- Test as a whole should have wide variety of item difficulty in order to work with diverse subjects.

Item Difficulty 2

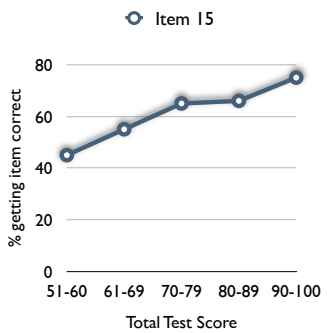
- Mathematically, 30%-70% is optimum
- What about human / emotional issues?
  - Tests or items that are too hard?
  - Tests or items that are too easy?

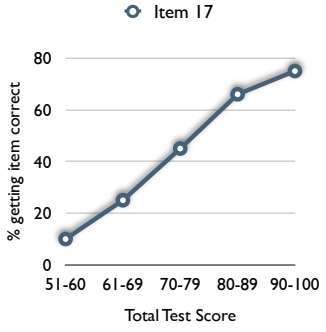
Discriminability

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

Item Characteristic Curve

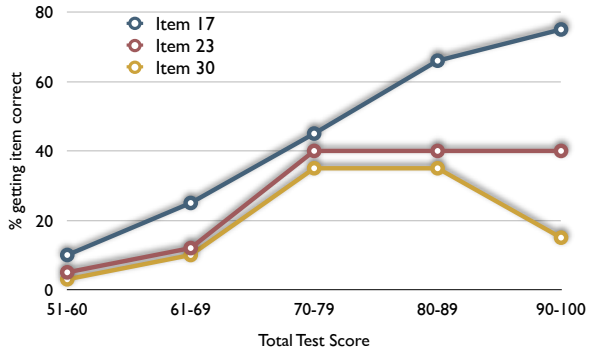
- Easier to look at this information visually
- Graph of % correct vs. total test score for one test item





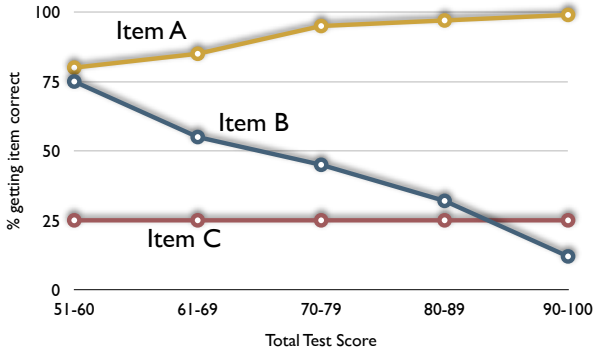
Item Characteristic Curve

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



ICC Example

- Diagnose these problems:



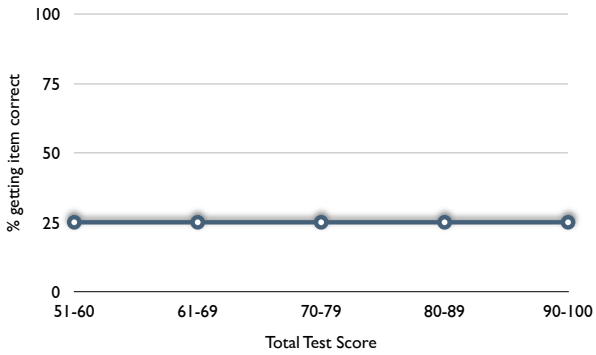
Graph the ICC

- 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



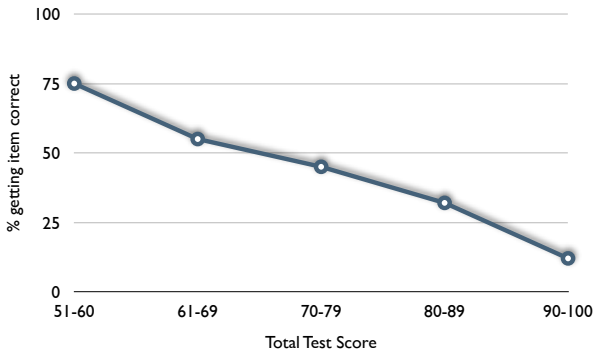
Graph the ICC

- Item 1: What is 0.34 times 0.27
  (A) 9.18
  (B) 0.61
  (C) 0.918
  (D) 91.8

- "Correct Answer" = B

ICC Example

- Test item has wrong answer



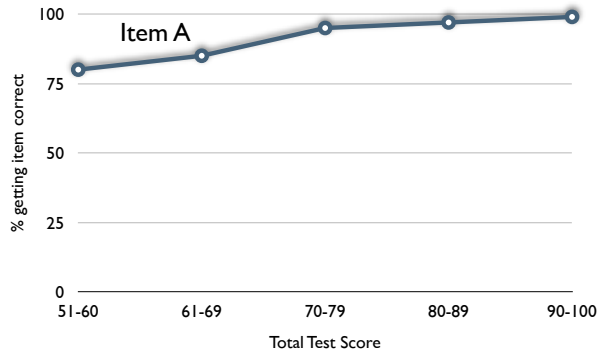
Graph the ICC

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

- Correct answer = C

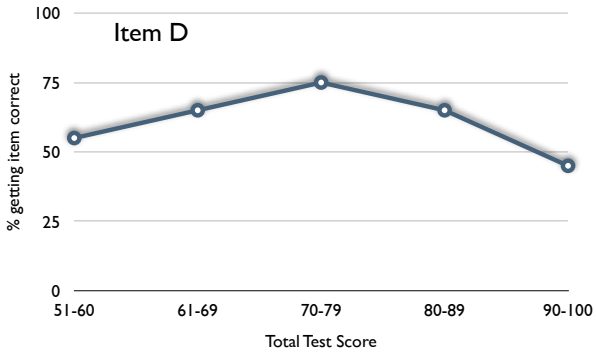
ICC Example

- Item is too easy



ICC Example

- "Overstudying" or "None of the above



Item Response Theory (IRT)

- Classical Test theory : score = # of items correct
- IRT: score = level of difficulty at which you can answer items correctly

- IRT Model : probability that item will be answered correctly is mathematically modeled using formal parameters (both Person and Test)

- IRT Procedures: using computer-based adaptive testing, test questions are given to focus in on the ability level of the test subject

IRT / Adaptive Testing

- For a test to cover a wide range of ability levels, it must have a wide range of item difficulties
- For an individual who has a particular skill level, this means many items are too easy, and many are 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")

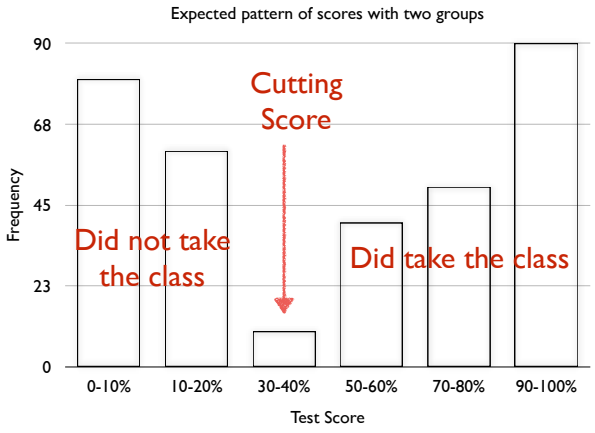
- Most Item Analysis still uses Internal criteria rather than the more correct External Criteria
- Why?

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 are designed to discriminate between different levels of performance
- Statistical tests (difficulty and discriminability) don't tell why a person missed an item
- Possible to develop items that 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)?