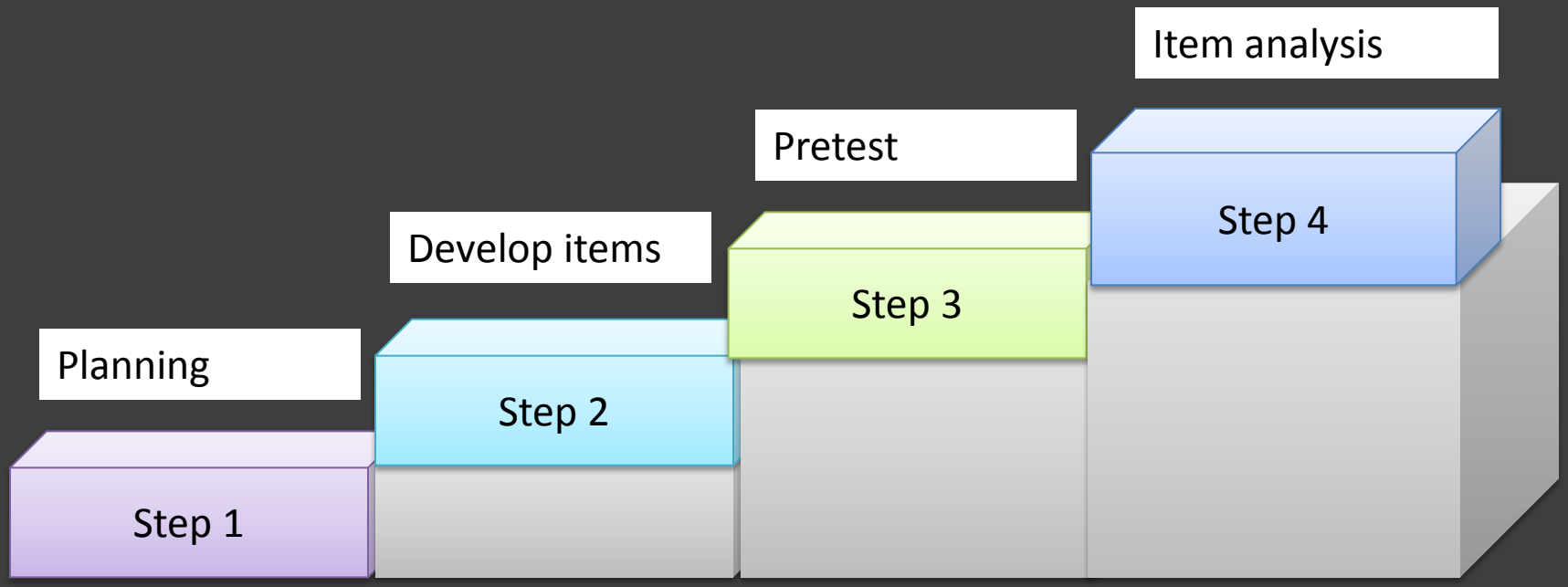
Putting Tests to the Test

Using the Results of Item Analysis to Evaluate the Multiple-Choice Exams

Developing a test

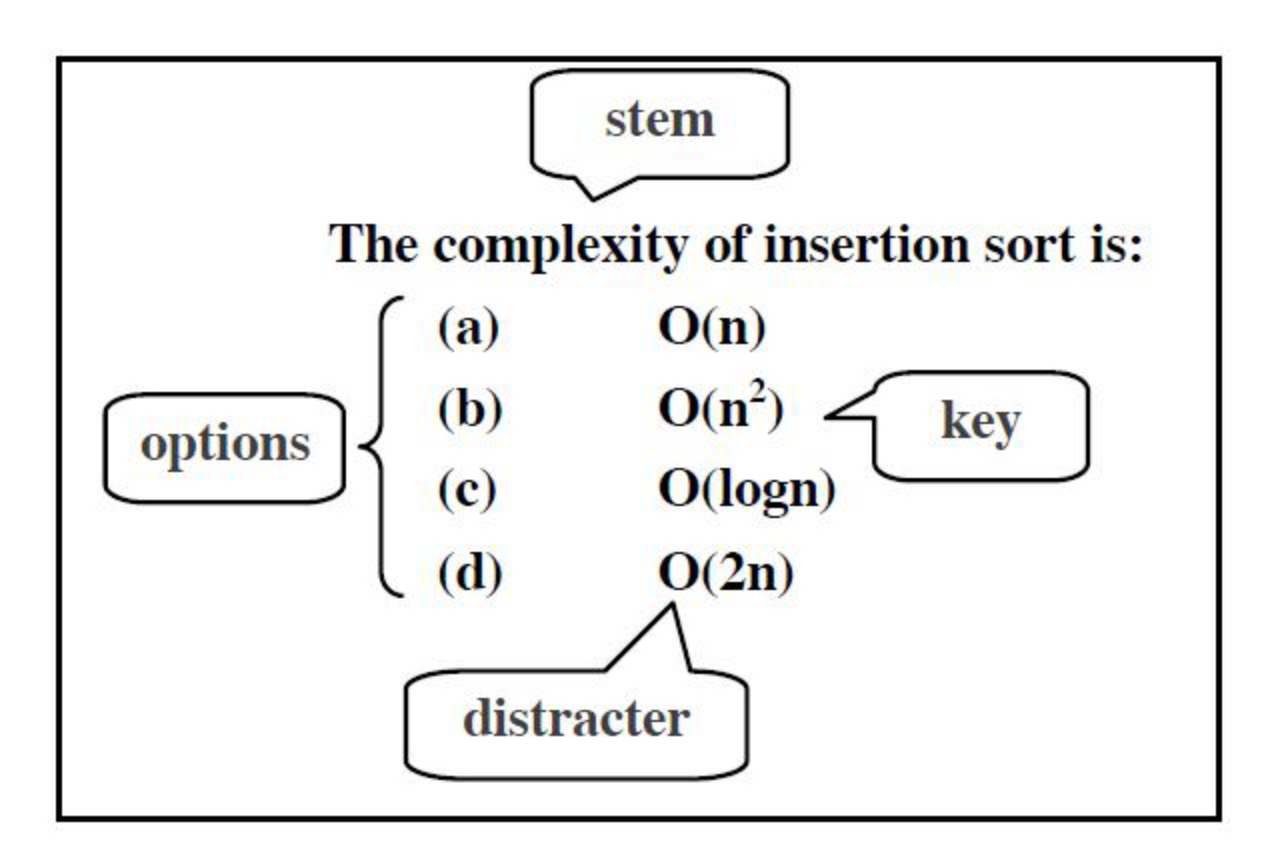


Test construction

Initial considerations . . .

- Classical Test Theory or Item Response Theory?
- Norm-referenced or criterion-referenced?
- Purpose and resources?
- Format?
- Length?
- Content & cognitive levels?

Anatomy of the MCQ



Writing items

- One best/correct answer
  - Avoid “none of the above”, “all of the above”
- Stem clearly stated
- Avoid negative wording
- Verbal/linguistic consistency
- Multiple cognitive levels
- Sufficient number of items

Item analysis

- Item difficulty
- Item discrimination
- Reliability
  - Coefficient alpha
  - Split-half correlation
- Distracter evaluation

Item difficulty

- Proportion with correct response (P)
- Poor items have . . .
  - P > 90% (too easy)
  - P < 20% (too hard)
- Ideal difficulty .
  - 5 options: 60%
  - 4 options: 62%
  - 3 options: 66%
  - 2 options: 75%

Item discrimination

- Point-Biserial (item-total) correlation
- Range: 0 to 1
- Higher score means more discriminating
- Guideline
  - 0.40 or higher — very good item
  - 0.30 to 0.39 — good item
  - 0.20 to 0.29 — fairly good item
  - 0.19 or less — poor item

Reliability – Coefficient alpha

- Ranges from 0 to 1
- Higher means more reliable
- Guideline
  - 0.90 or higher — excellent (standardized tests)
  - 0.80 to 0.90 — very good (for classroom tests)
  - 0.70 to 0.80 — good (typical classroom test)
  - 0.60 to 0.70 — low
  - 0.50 to 0.60 — need to revise
  - 0.50 or below — questionable
- Larger number of items improves reliability

Distracters

- Should be clearly incorrect
- Appeal to low scorers; infrequently chosen by high scorers
- Examine frequency table
  - Chosen by few = implausible, replace
  - Chosen by many = confusing, misleading
  - Should be reasonably uniform distribution among those getting question wrong
- Crosstab by ability can also give insight