Test Development

Strategies and Techniques

G. Donald Frey, Ph.D.
Associate Executive Director
American Board of Radiology

Disclosures

- Paid employee of the ABR
- Content of talk is my own

Different Types of Exams

High Stakes/Low Stakes

- High Stakes
  - ABR Certification
  - SAT
  - End of Term Exam
  - PhD qualification Exam
- Low Stakes
  - Weekly quiz
  - Post course surveys
  - Homework assignment

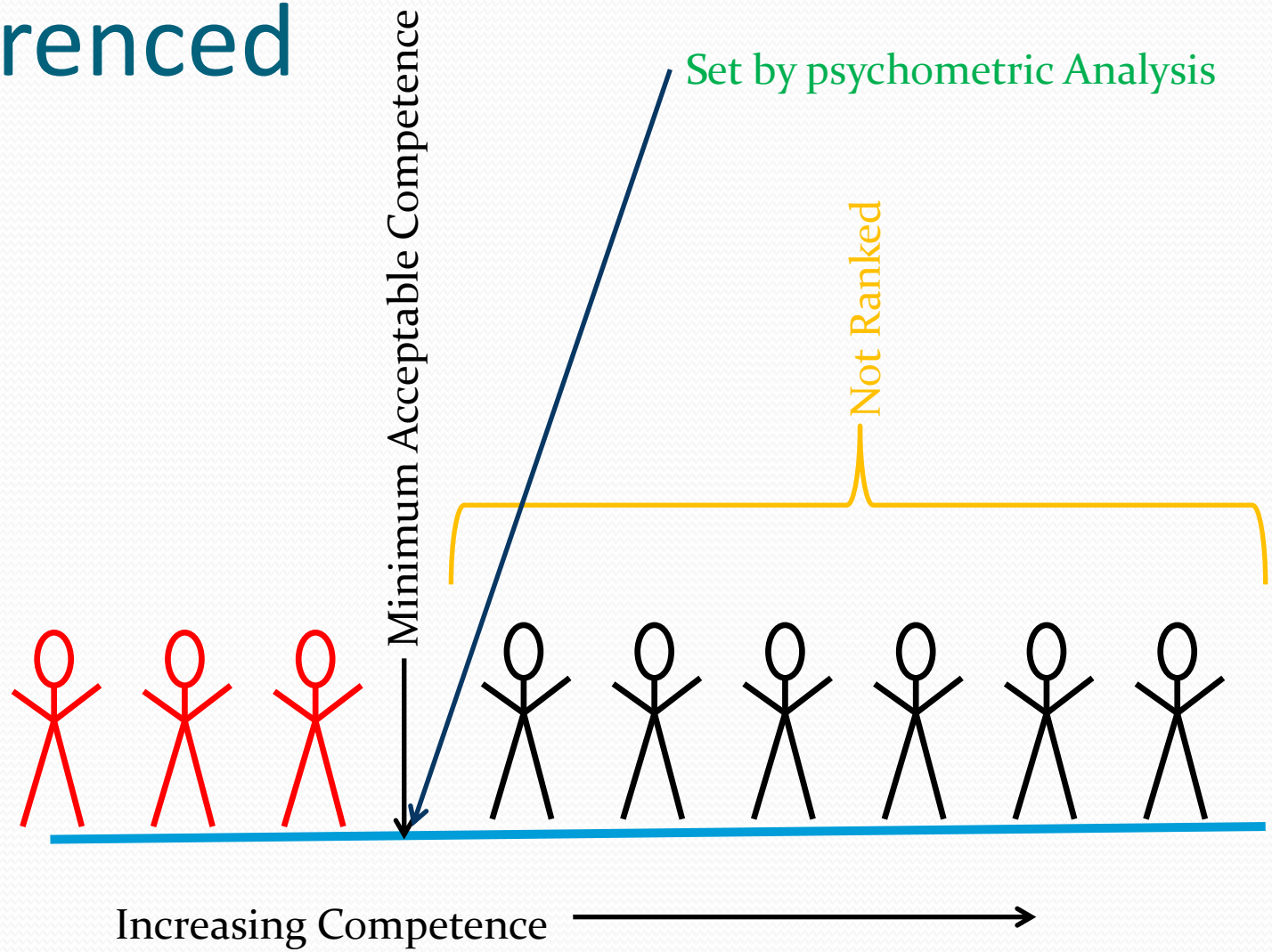
Types of Exams (Stakes)

Consequence	Low	High
Results	Easy to reverse	Hard to reverse
Reliability	Not of great importance	Very important
Development Effort	Low	High
Motivation to Cheat	Low	High

Criterion Referenced / Norm Referenced

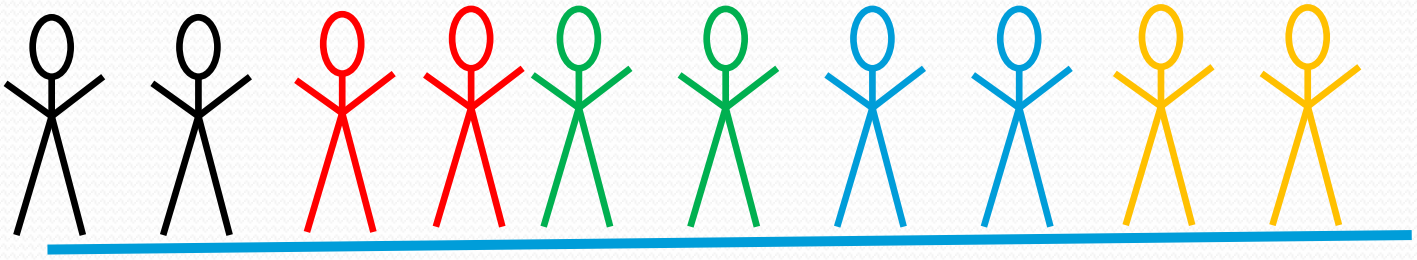
- Criterion Referenced
  - Designed to determine if an individual meets a standard
- Norm referenced
  - Designed to rank a group of individuals

Criterion Referenced



Norm Referenced

May not be a passing score



Increasing Competence

Examples

- Criterion Referenced
  - ABR Certification
  - Drivers License Test
- Norm Referenced
  - SAT
- Mixed
  - Course grade
    - Criterion referenced for passing
    - Kind of norm references for letter grades

Exam Validity

Test Score – Inflation & Reduction

$$X_{actual} = X_{true} + X_{error}$$

X_{error} Factors

Inflation	Reduction
Poorly written questions	Illness
Use of unauthorized technology	Cold room
Copying from other test takers	Poor light on monitor
Advanced knowledge of content	Uncompensated ADA

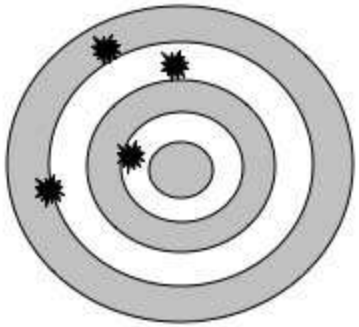
Reducing X_{error} Factors

- Improving and standardizing test environment

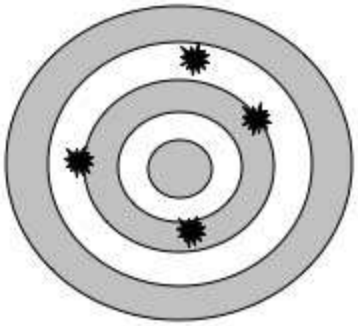
- Overall to reduce X_{error} to make X_{actual} as close to X_{true} as possible

Validity

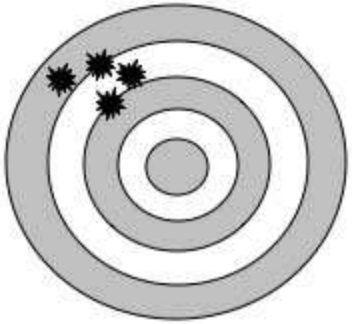
- Does a test measure what it is supposed to measure
  - Much like accuracy &precision in statistics



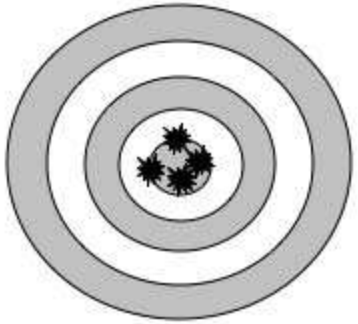
Not Reliable
Not Valid



Not Reliable
Valid



Reliable
Not Valid



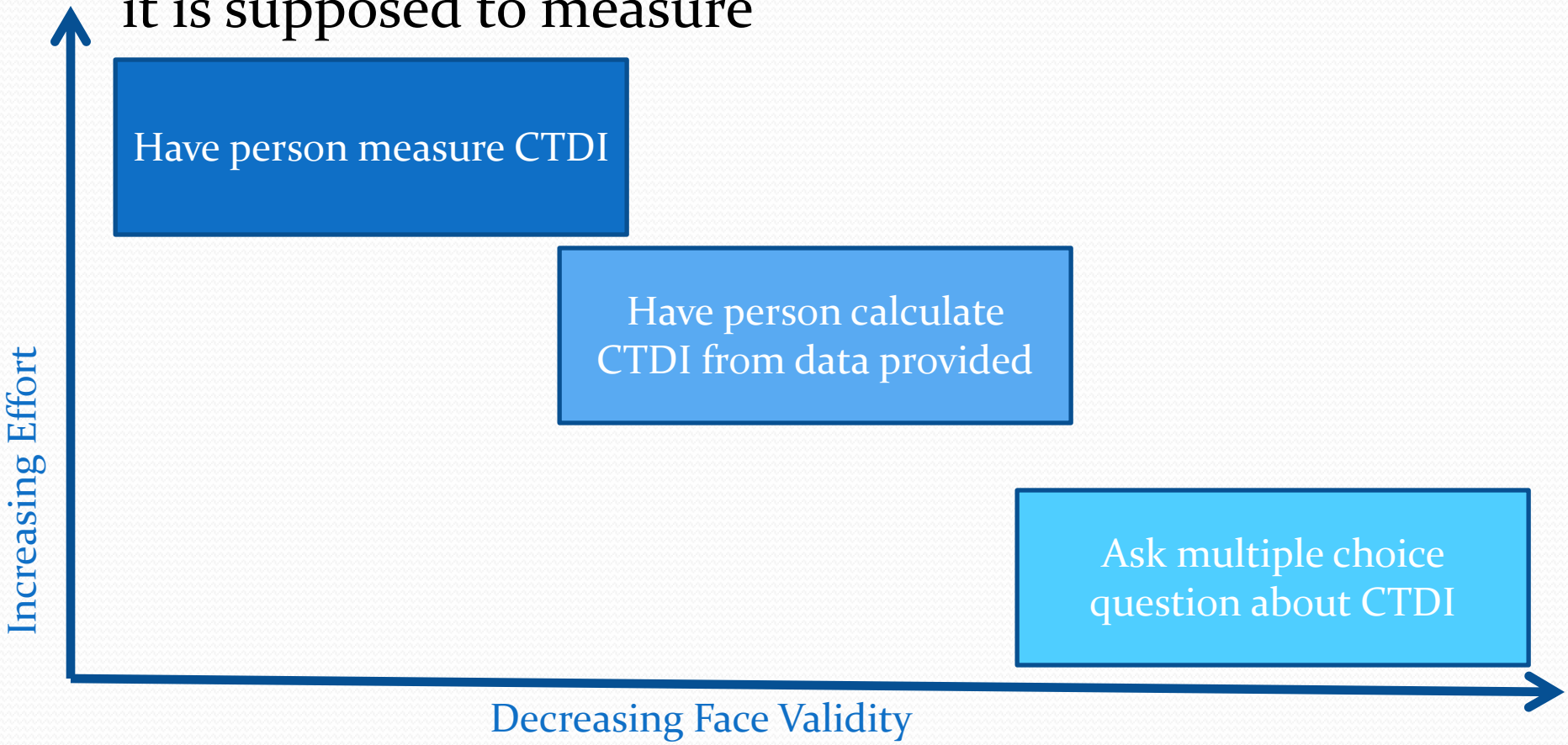
Reliable
Valid

Issues in Validity

- **Face validity**
- **Content validity**
- Concurrent validity
- Predictive validity

Face Validity

- Does the test appear to the test taker to measure what it is supposed to measure



Content Validity

- When a group of content experts verify that the test measures what it is supposed to
  - Can be one to 10 or more experts

Concurrent Validity

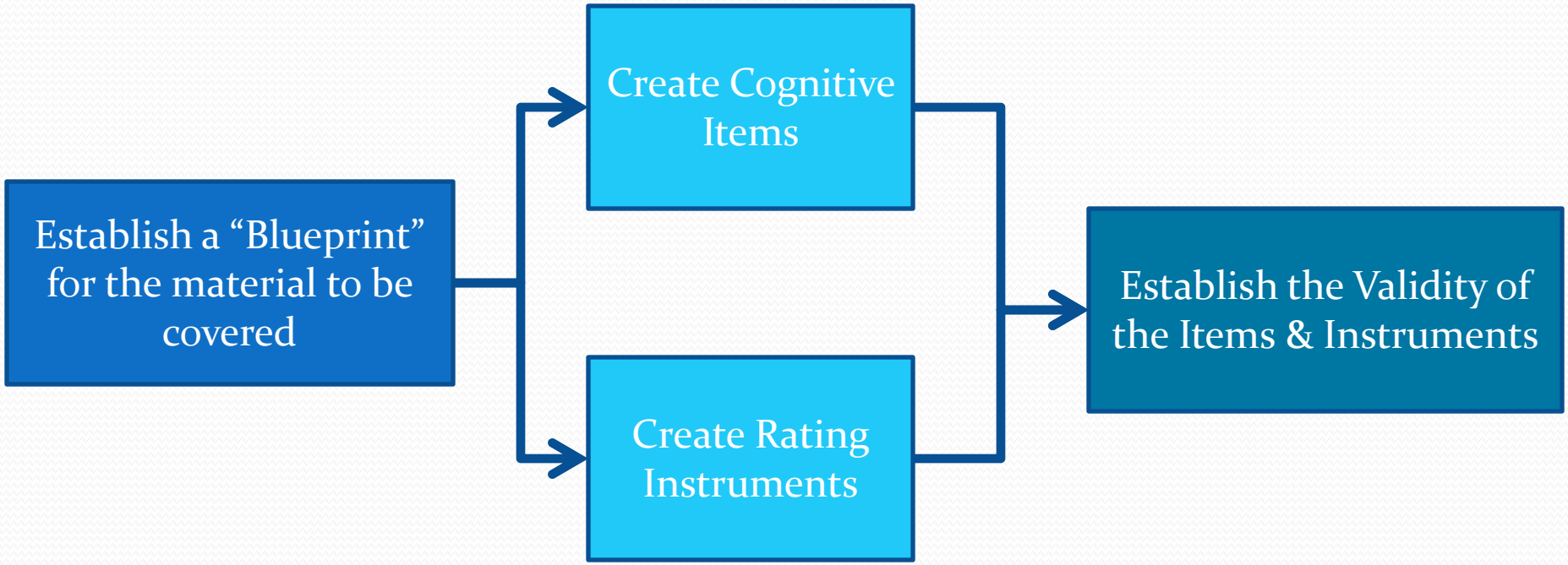
- The test actually separates the competent from the non-competent individuals
- Requires known members of each group to take the test
- Often not practical

Predicative Validity

- Does the test establish the future competence test takers
  - Do aptitude test predict future performance?
  - Do SAT scores predict college performance?

Design of Criterion Referenced Exams

Design of Criterion Referenced Tests



Blueprint

- Course syllabus
- Surveys of Practitioners
- National Recommendations
  - AAPM Task Group Reports

Blueprint should give the breakdown of topics and the percent of items on the topic

i.e. 12 % of test items are on brachytherapy

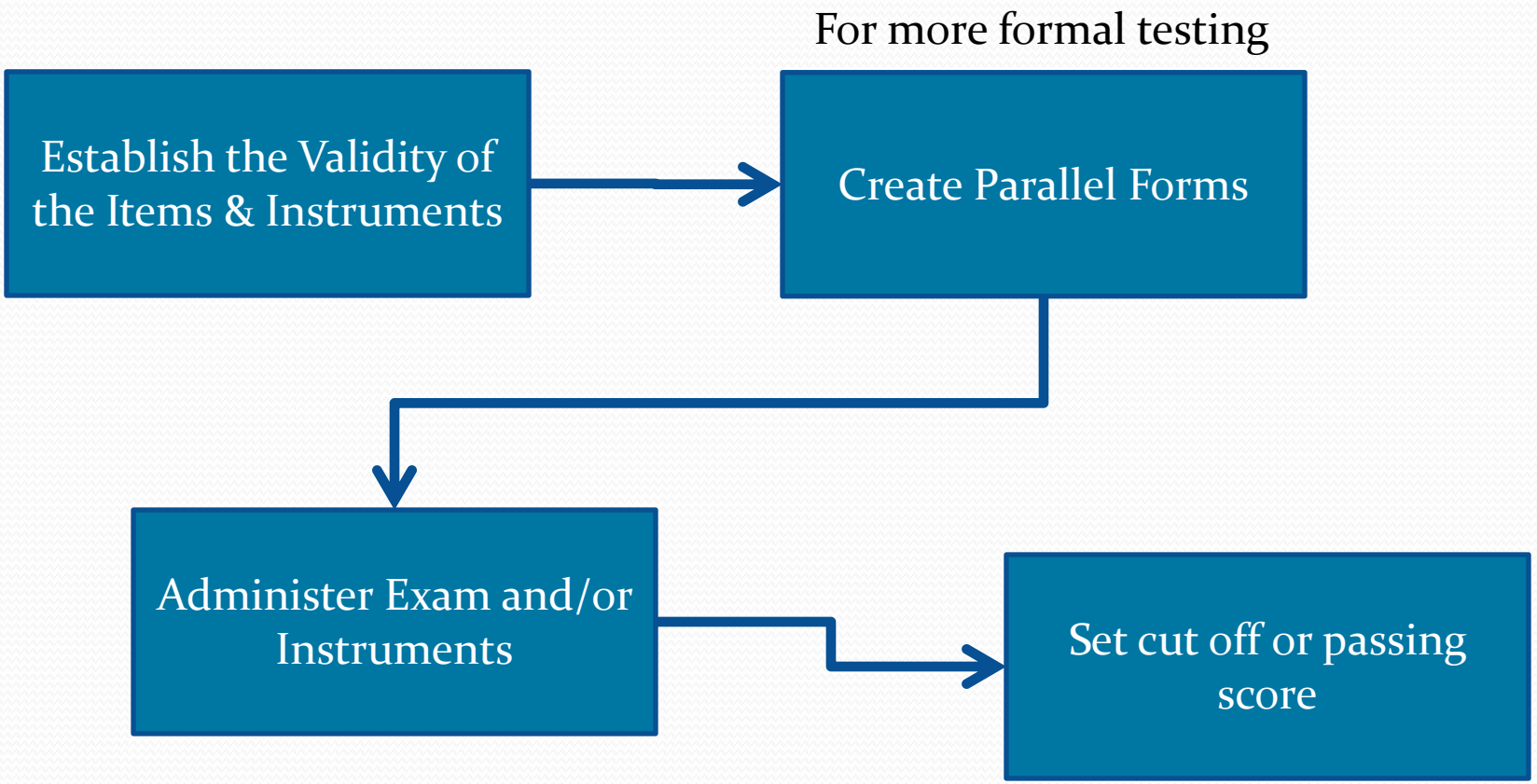
Develop Items &Rating Instruments

- Items can be multiple choice, essay, problems, matching, etc.
- Rating instruments would be things like practicums and laboratory tests
  - These need to be validated also

Establish the validity of the items

- Content experts

Design of Criterion Referenced Tests



Parallel Forms

- When test is repeated you will need parallel forms or versions
- Even in routine academic situations you need to establish validity from one semester to another.

Set Passing Score or Cutoff

- In the ideal world all forms would have the same passing score but usually there is some variation from form to form. Using (a) content expert(s) to analyze the material one can adjust for variation from form to form.
- There is a formal way to do this that is used in complex exams
  - Angoff Proceedure

Analyzing Questions

This will focus on multiple choice questions

Stem

A) Correct answer
B) Distractor 1
C) Distractor 2
D) Distractor 3

Question Criteria

- How often does the candidate get the question right?
- How do the distractors work?
- Do the high scoring candidates get the question right more often than the poor scoring candidates?

Evaluation of Multiple Choice Questions

- Difficulty
- Selection of Distractors
- Correlation

Difficulty

- Fraction of test takers that get the item correct
  - Values from 0.0 to 1.0
- Acceptable range depends on the nature of the test
  - Norm referenced would use ~0.30-0.70
  - Criterion referenced ~0.20-0.85
- High and low values do not give any useful information
  - Effectively lower the number of questions
- Reasons for high and low values
  - Wrong answer in key
  - Question too hard or too easy
  - Question “keys” the answer

Distractors

- The perfect would have the incorrect answers uniformly distributed among the incorrect distractors
  - Frequency that a distractor is selected is called “p”
    - Not usual statistics “p”
  - A distractor with a very low p-value does not provide any useful information
  - If the test has a pattern of unused distractors it suggests the questions are too easy and this will increase the odds that the correct answer can be guessed.

Correlation
(Point Biserial Correlation)

- Good performers should get an item correct more than poor performers
- Value can be from 1.0 to -1.0
- A negative correlation suggests that:
  - Key is wrong
  - Instructor has been inadvertently saying something that causes the good students to miss the item
  - Good performers see something in a distractor that poor performers do not

Calculating Correlations

- Can be done in Excel
- See for example : www.eddata.com/resources/publications/EDS_Point_Biserial.pdf

Good and Bad Questions

A	B	C	D	Point Biserial
0.65	0.05	0.20	0.10	0.58
0.25	0.40	0.25	0.10	0.35
0.20	0.25	0.40	0.15	0.05
0.00	0.75	0.10	0.15	0.10
0.90	0.10	0.00	0.00	-0.05

Reliability

Reliability

- Obtaining consistent results from one administration to the next.
- For large scale testing there are statistical measures that are used
- In the small scale setting the administrator should compare the results over time to assure that the tests are reliable
- If you have rateres (lab practicum evaluators) you should be sure they all have similar training and produce similar results.

Practical Steps

Preparation

- Be sure that design matches the course syllabus
- Insure that facilities for administration meet the requirements
- If laboratory practicals are used be sure that all the administrators are trained to administer the material in the same way

Item Preparation

- Use good item writing technique
- Determine the intellectual level of each item
- Decide on the validity of each item
  - Try to improve the face validity

Post Administration

- Determine the difficulty of each item
- Review the performance of each distractor
- If possible calculate the point biserial correlation
- Edit question if it is to be used again

Audience Response Systems

Effective Use

ARS

- Engage all students
  - Especially valuable for those reluctant to participate
- Effective use requires a dynamic presentation
  - Instructor must be ready to modify the presentation in response to the student response

Common Uses

- To develop a knowledge of the scope of knowledge at the beginning of a session
- To expose and correct student mis-conceptions
- To determine which concepts the students find difficult
- To promote interactivity and develop discussion'
- For topics that don't have a necessarily correct answer develop multiple points of view
  - What would you do next?
  - Ethics issues

Common Uses

- Creation of mock exams
  - But if presented as such the mock exam must be similar to the actual exam
- To provide immediate feed back at the end of the session

General Process

- Engage the students with a question
  - For new groups a question about them might be appropriate
  - For a recurring class a question about the previous session
- Allow sufficient time for each question
  - At least three to four minutes
  - Time to consider the question and answer
  - Time to discuss the correct and incorrect answer
  - Additional time to modify the presentation if question shows the students do not understand

General Process

- Engage students in the discussion of the item
  - Why was A) incorrect
  - How does this relate to “xx” that we discussed in the last lecture
- Because good questions take a lot of time you need to limit the number
  - The exception is exam simulation

Nature of ARS Items

- Should be as simple as possible
- Options should match the key pad
- Consider each distractor so you can discuss what is incorrect and why

Thank you