FAIMER®

Foundation for Advancement of International Medical Education and Research

Setting Standards

John Norcini, Ph.D.

jnorcini@faimer.org

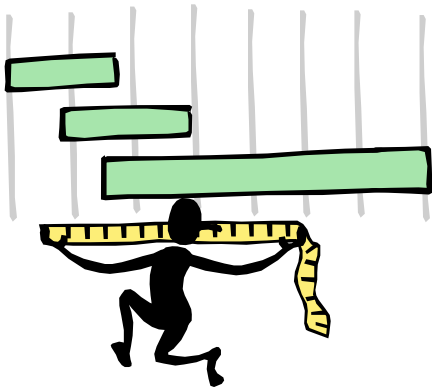
Overview

- Scores and standards
  - Definitions and types
- Characteristics of a credible standard
  - Who sets the standards, what are the characteristics of the method, and what is the outcome?
- Methods
- Steps in implementation

Scores and Standards

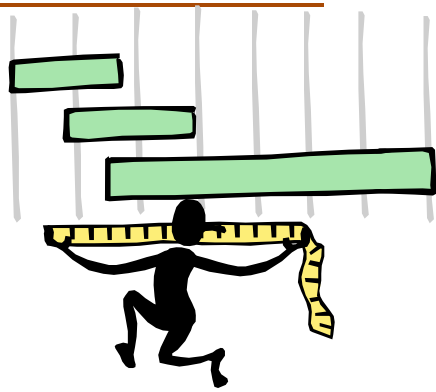
- Standard-setting is unsettled due to
  - The arbitrary nature of standards
  - Confusion over terminology
    - Norm-referenced, criterion-referenced…
- Provide a framework
  - Definition of scores and standards
  - Types of score interpretation and standards

Definition of Scores



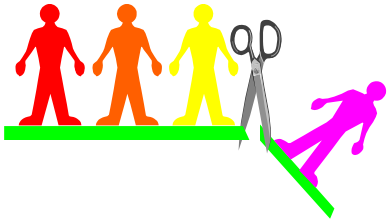
- A score is a number or letter that represents how well an examinee performs along a continuum
  - The degree of medical correctness for a response or group of responses
  - The numerical answer to the question, "how good is the examinee's performance from the perspective of the patient?"

Definition of Scores



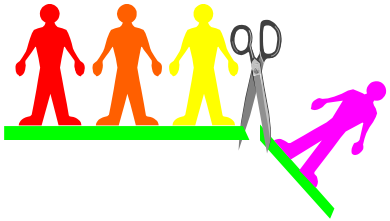
- For MCQs a score is based on the actual responses of examinees--a count
- For formats reproducing complex clinical situations with high fidelity
  - May involve weighting (degrees of correctness)
  - May involve an interpretation of the examinee's responses (e.g., oral exam)

Definition of Standards



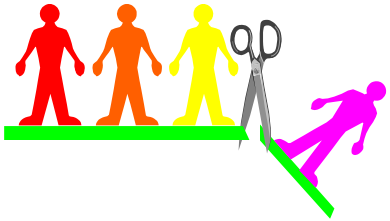
- A standard is a statement about whether an examination performance is good enough for a particular purpose
  - A special score that serves as the boundary
  - The numerical answer to the question,
    - "How much is enough?"
    - "How tall is the shortest giant?"

Definition of Standards



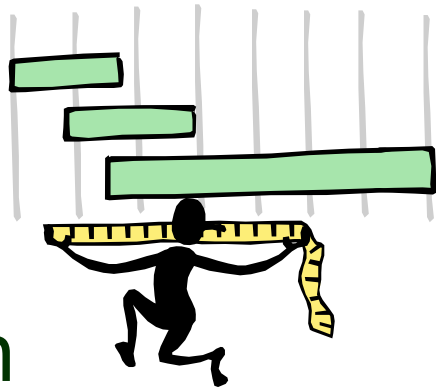
- Standards are based on judgments about examinees' performances against a social or educational construct
  - Competent practitioner or student ready for graduation
- Standards are not based on the patient outcomes that form the basis for scoring

Definition of Standards



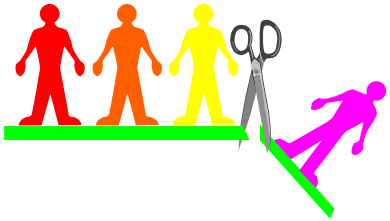
- Standards are judgmental or arbitrary
  - No ‘true’ standard
  - Not possible to collect data that definitively support a standard to the exclusion of others
  - Essential to collect data which build a case for the standard that is chosen

Types of Scores Interpretation



- Norm-referenced score interpretation
  - Based on how an examinee performs against others who took the test
  - For example, rank or percentiles
- Domain-referenced score interpretation
  - Based on how an examinee performs against the test content
  - For example, number right or percent correct

Types of Standards

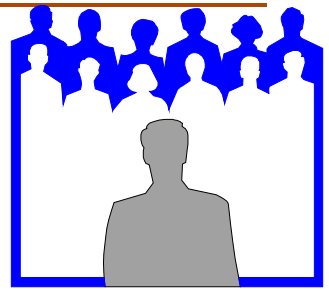


- Relative standards
  - Based on a comparison among the performances of examinees
  - For example, the top 84% pass
- Absolute standards
  - Based on how much the examinees know
  - For example, examinees must correctly answer 70% of the questions

Characteristics of a Credible Standard

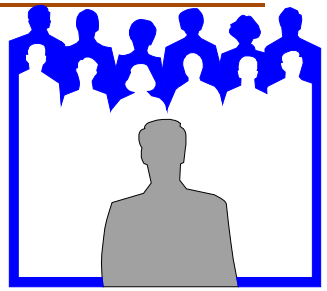
- Who sets the standards?
- What are the characteristics of the method being used?
- What is the outcome?

Who Sets the Standard?



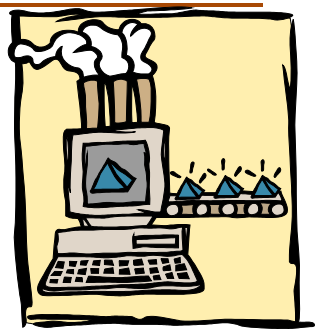
- Standard setters must
  - Understand the purpose of the test, know the content, and be familiar with the examinees
- Low stakes setting (e.g., course)
  - Single faculty member is efficient and credible but...
    - He/she has a conflict of interest
    - Standards will vary over content and time

Who Sets the Standard?



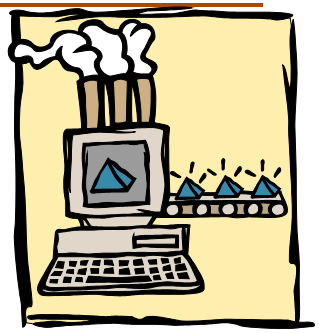
- High stakes setting (e.g., certification)
  - A significant number need to be involved
    - Increases the reproducible of standards, reduces stringency effects and differences over time
  - They need to represent a mix of attributes
    - Educators-academics
    - Practitioners
    - Balance by geography, gender, race, etc.
  - They must not have conflicts of interest

What Are the Characteristics of the Method?



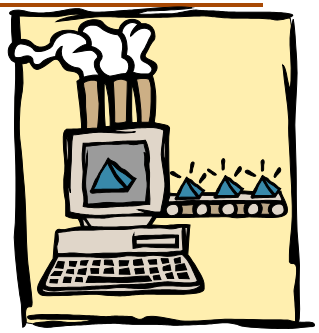
- Exact method used to set standards is less important than whether it
  - Produces standards consistent with the purpose of the test
  - Relies on informed expert judgment
  - Demonstrates due diligence
  - Is supported by a body of research
  - Is easy to explain and implement

Method: Fit for Purpose



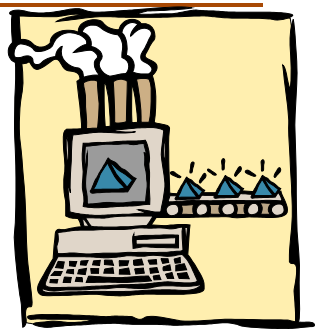
- Use the type of standards that are consistent with the purpose of the test
  - Absolute standards are preferred for most high stakes competence exams
  - Relative standards are preferred when identifying the best/worst (e.g., admissions)
    - Set without regard to how much is known
    - Vary with examinees' ability ('vintages')

Method: Based on Informed Judgment



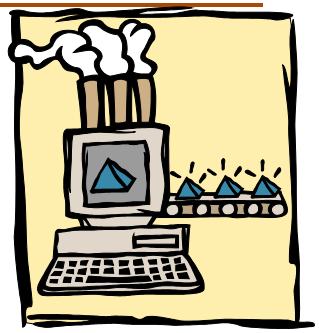
- Standard-setting methods can be based on
  - Empirical results (e.g., match with criterion)
  - Expert judgment
- Combined approaches produce better results
  - They have the most credibility with the examinees and stakeholders
  - Preference should be given to the judgment of experts in the presence of performance data

Method: Demonstrates Due Diligence



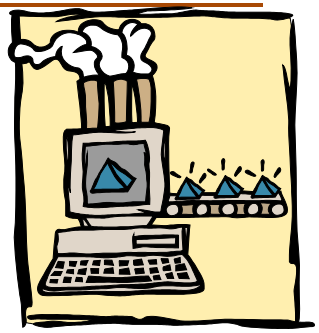
- Due diligence lends credibility
  - Method should require experts to expend considerable and thoughtful effort
- In contrast
  - Methods requiring quick, global judgments produce less credible results
  - Methods requiring several days are unnecessary and unreasonable

Method: Supported by Research



- Methods supported by a research literature produce results that are more credible
  - Ideally, studies should show that standards are
    - Reasonable compared to those produced by other methods
    - Reproducible over groups of judges
    - Insensitive to potentially biasing effect
    - Sensitive to differences in test difficulty and content
  - Research on Angoff's method is an example

Method: Easy to Explain and Implement



- Credibility is enhanced if the method is easy to explain and implement
  - Decreases the amount of training required for the judges
  - Increases the likelihood of judge compliance
  - Assures examinees everyone is treated the same way

Are the Outcomes Realistic?

- A standard that produces an unrealistic outcome will not be viewed as credible
- Building a case requires evidence that the standard
  - Is viewed as correct by stakeholders
  - Produces pass rates that have reasonable relationships with contemporaneous markers of competence
  - Is related to later performance

Summary

- Two types of standards
  - Relative and absolute
- Credible standards derive from
  - Standard-setters
    - Many with a mix of attributes but no conflicts
  - Method
    - Fit for purpose, informed judgment, diligence, researched, easy to explain and implement
  - Outcomes
    - Stakeholder support, reasonable relationships with markers of competence

Classification Scheme

- Classification system for methods of setting standards (Livingston & Zeiky, 1982)
  - Relative methods based on judgments about groups of test takers
  - Absolute methods based on judgments about the performance of individual examinees
  - Absolute methods based on judgments about test questions
  - Compromise methods

Relative Methods: Judgments About Groups of Test-takers

- Methods
  - Fixed percentage method
  - Reference group method
- Process
  - Select the judges
  - Discuss
    - Purpose of the test
    - Nature of the examinees
    - What constitutes adequate/inadequate knowledge
  - Review the test in detail

Relative Methods: Judgments About Groups of Test-takers

- Fixed percentage
  - Each judge estimates the pass rate for all examinees
- Reference group
  - Decide which group to use
  - Ask each judge to estimate the pass rate
- Discuss and permit changes
- Average the judges' pass rates

Relative Methods: Judgments About Groups of Test-takers

- Advantages
  - The methods are quick and easy
  - The process only has to be done occasionally, not every time the test is given
  - Judges usually have acceptable pass-rates in mind
  - Apply equally well to all written exam formats

Relative Methods: Judgments About Groups of Test-takers

- Disadvantages
  - Standards vary with the ability of examinees
  - Seem to manipulate size of the passing group
  - Independent of how much examinees know
  - Independent of test content

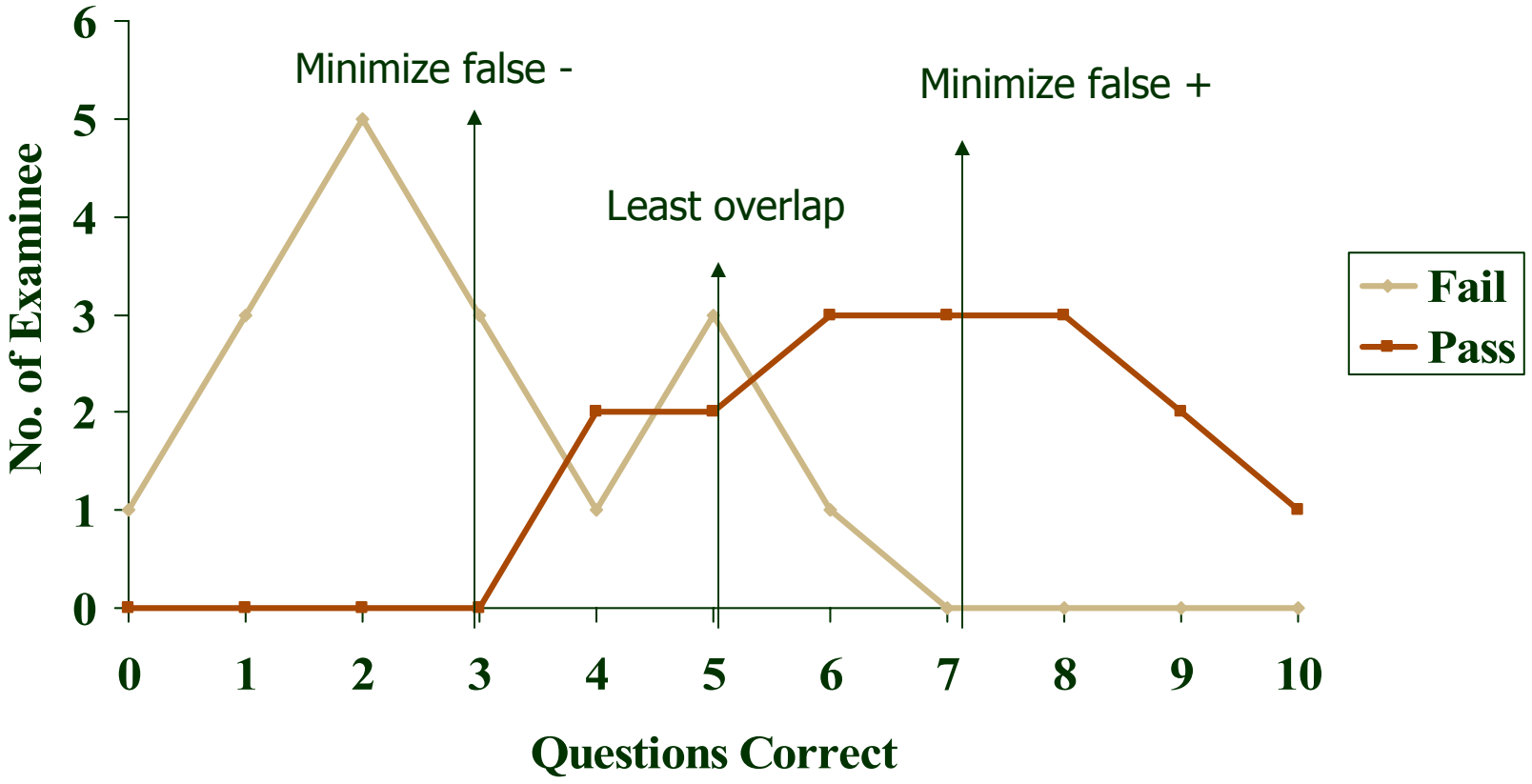
Absolute Methods: Judgments About Individual Test-takers

- Methods
  - Contrasting-groups method
  - Up-and-down method
- Process for Contrasting Groups
  - Select the judges
  - Discuss
    - Purpose of the test
    - Nature of the examinees
    - What constitutes adequate/inadequate knowledge
  - Review the test in detail

Absolute Methods: Judgments About Individual Test-takers

- Process for Contrasting Groups
  - Select a random sample of examinees
  - Give the judges their responses to the entire test
  - Ask the judges to decide (consensus, majority) whether each should pass or fail
  - Graph the scores of the passers and failers
  - Calculate the passing score
    - For example, the point of least overlap

The Contrasting Groups Method



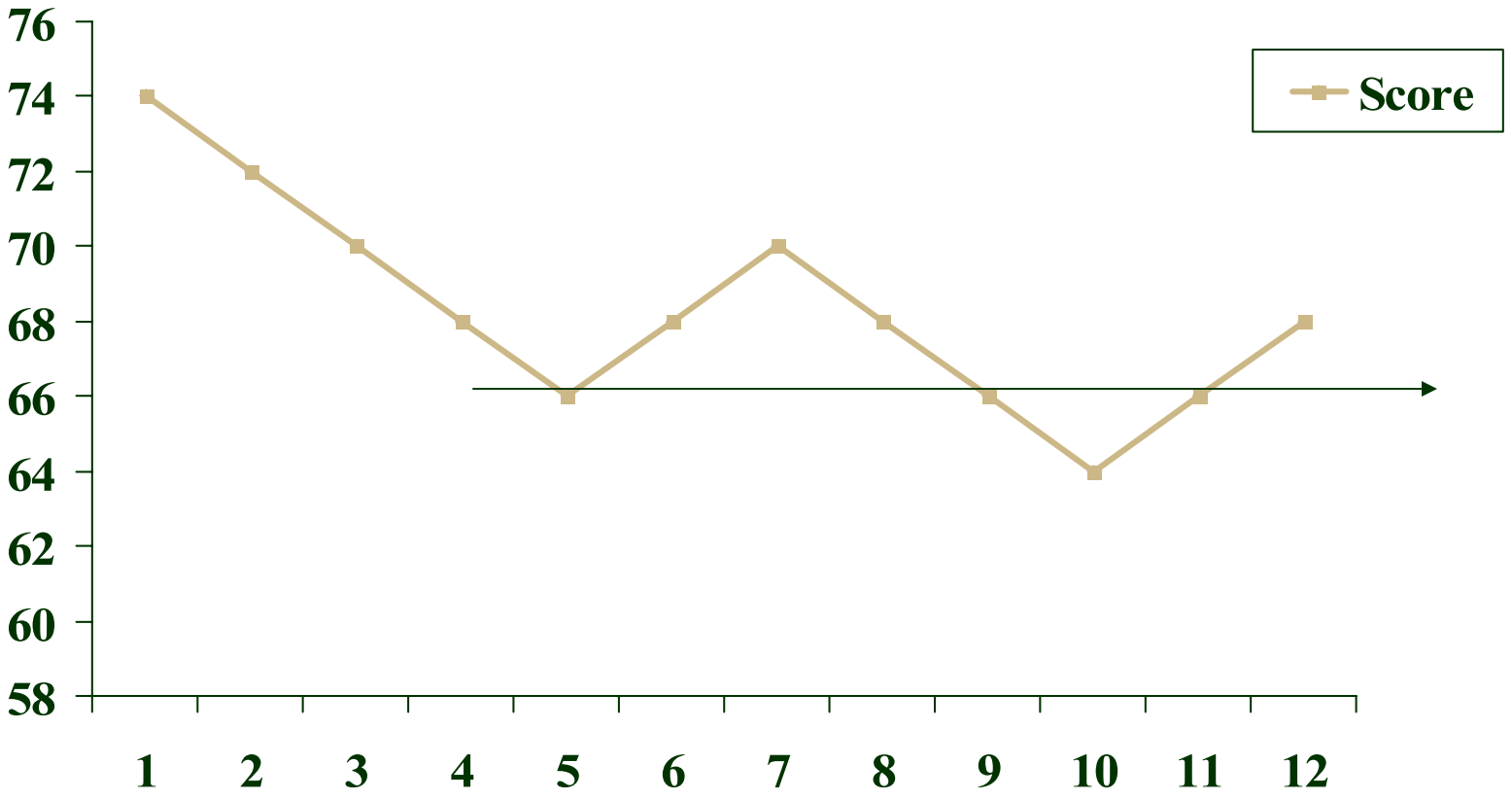
Absolute Methods: Judgments About Individual Test-takers

- Process for the up-and-down method
  - Select the judges
  - Discuss
    - Purpose of the test
    - Nature of the examinees
    - What constitutes adequate/inadequate knowledge
  - Select a sample of examinees near the cutting score
  - Give the judges the responses to the entire test of one examinee

Absolute Methods: Judgments About Individual Test-takers

- Process for the up-and-down method
  - Ask the judges to decide (consensus, majority) whether the examinee should pass or fail
  - If pass, choose an examinee with a lower score
  - If fail, choose an examinee with a higher score
  - Repeat for several examinees
  - Calculate the passing score (e.g., mean of the last 10 scores)

The Up-and-Down Method



Absolute Methods: Judgments About Individual Test-takers

- Advantages
  - Educators are comfortable making these types of judgments
  - The methods inform the judgments of experts with the actual test performance of examinees
  - Contrasting groups allow manipulation of false positive and negative rates

Absolute Methods: Judgments About Individual Test-takers

- Disadvantages
  - It is time-consuming and difficult to review entire tests and make unbiased judgments about the skills of examinees
  - Judgments must be made about a fairly large number of test-takers in order to create reliable passing scores
  - Choosing the actual passing score can be very subjective

Absolute Methods: Judgments About Individual Test Items

- Methods
  - Angoff's method
  - Ebel's method
- Process for Angoff's Method
  - Select the judges
  - Discuss
    - Purpose of the test
    - Nature of the examinees
    - What constitutes adequate/inadequate knowledge

Absolute Methods: Judgments About Individual Test Items

- Process for Angoff's Method
  - Define the "borderline" group
  - Read the first item
  - Estimate the proportion of the borderline group that would respond correctly
  - Record ratings, discuss, and change
  - Repeat for each item
  - Calculate the passing score

Angoff's Method

	Judge					
Items	1	2	3	4	5	Mean
1	.60	.70	.55	.75	.65	.65
2	.80	.90	.85	.95	.90	.88
3	.70	.75	.80	.75	.40	.68
4	.45	.55	.50	.60	.55	.53
5	.90	.95	.85	.95	.90	.91
Total						3.65

Absolute Methods: Judgments About Individual Test Items

- Process for Ebel's Method
  - Select the judges
  - Discuss
    - Purpose of the test
    - Nature of the examinees
    - What constitutes adequate/inadequate knowledge
  - Define the "borderline" group
  - Build a classification table for items based on a category scheme (like difficulty and importance)

Absolute Methods: Judgments About Individual Test Items

- Process for Ebel's Method
  - Judges read each item and assign it to one of the categories in the classification table
  - They make judgments about the percentages of items in each category that borderline test-takers would have taken or answered correctly
  - Calculate passing score

Ebel's Method

Category	% Right	# Questions	Score
Essential			
Easy	95	3	2.85
Hard	80	2	1.60
Important			
Easy	90	3	2.70
Hard	75	4	3.00
Acceptable			
Easy	80	2	1.60
Hard	50	3	1.50
		17	12.25

Absolute Methods: Judgments About Individual Test Items

- Advantages
  - They focus attention on item content
  - They are relatively easy to use
  - There is a considerable body of published work supporting their use
  - They are used frequently in high stakes testing

Absolute Methods: Judgments About Individual Test Items

- Disadvantages
  - The concept of a "borderline group" is sometimes foreign to judges
  - Judges sometimes feel they are "pulling numbers out of the air"
  - The methods can be tedious

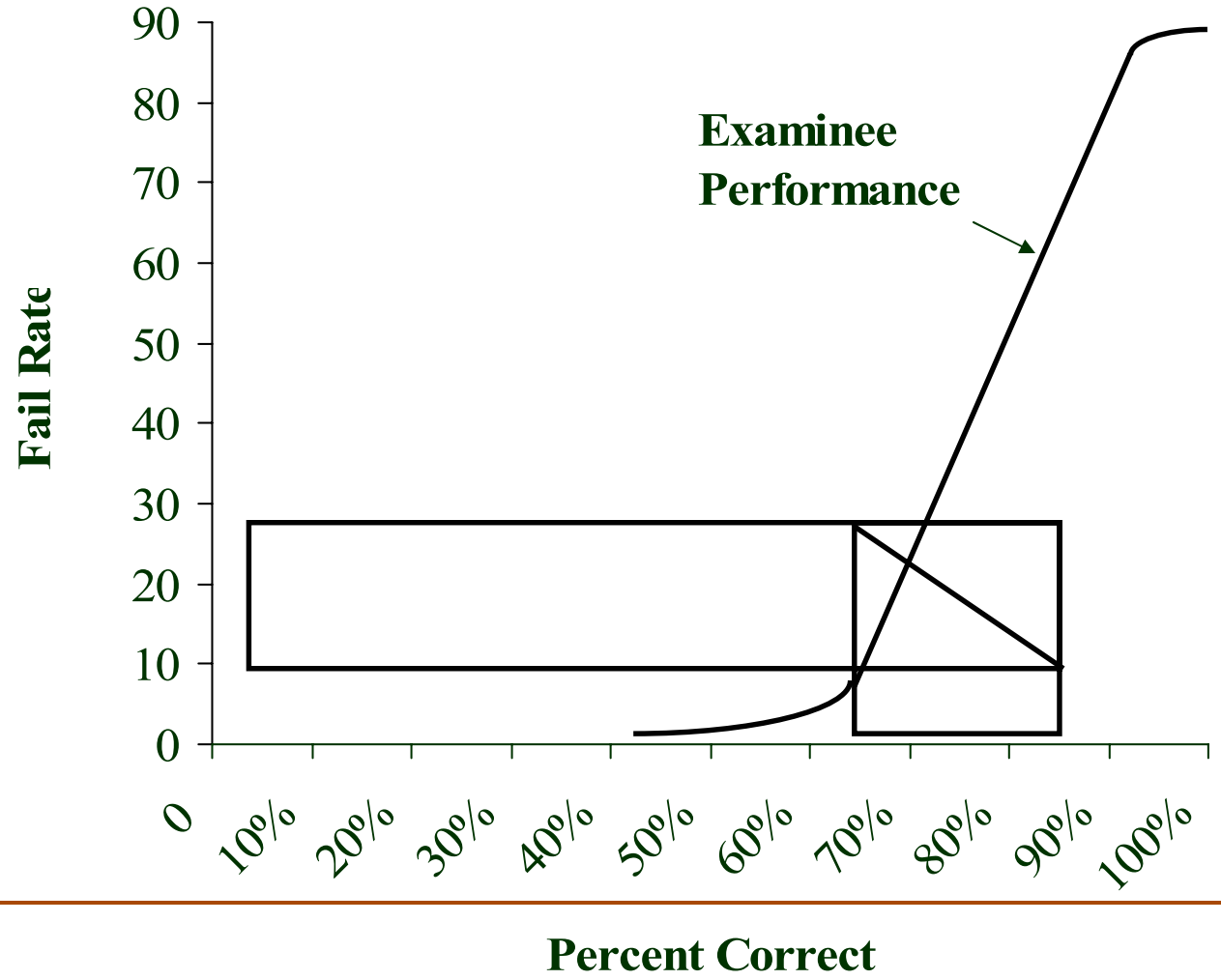
Compromise Methods

- Hofstee Method
  - Select the judges
  - Discuss
    - Purpose of the test
    - Nature of the examinees
    - What constitutes adequate/inadequate knowledge
  - Review the test in detail

Compromise Methods

- Process for Hofstee's Method
  - Ask the judges to answer four questions:
    - What is the minimum acceptable cut score?
    - What is the maximum acceptable cut score?
    - What is the minimum acceptable fail rate?
    - What is the maximum acceptable fail rate?
  - After the test is given, graph the distribution of scores and select the cut score

Hofstee Method



Compromise Methods

- Advantages
  - Easy to implement
  - Educators are comfortable with the decisions
- Disadvantages
  - The cut score may not be in the area defined by the judges' estimates
  - The method is not the first choice in a high stakes testing situation

Methods for Setting Standards on Other Written Formats

- Most methods apply directly
  - Relative methods
  - Absolute methods
    - Contrasting Groups and Up-and-Down
      - Can be done by question and then combined
    - Angoff and Ebel
      - What score would the borderline examinee get?
  - Compromise methods

Implementation Guidelines for Setting Standards

- Select the judges
  - Assign an appropriate number (at least 6-8 for high stakes testing)
  - Select the characteristics the group should possess
  - Develop an efficient design for the exercise

Implementation Guidelines for Setting Standards

- Hold the standard setting meeting
  - Make sure all judges attend throughout
  - Explain the procedure and educate the judges about the consequences of their decisions
  - Discuss
    - Purpose of the test
    - Nature of the examinees
    - What constitutes adequate/inadequate knowledge
  - Review the test in detail
  - Practice with a few items, cases, or examinees
  - Give feedback at several intervals

Implementation Guidelines for Setting Standards

- Calculate the standard
  - Decide how to handle outliers, missing data, etc.
  - Ensure that the standard is reproducible
  - Have a compromise standard available if possible

Implementation Guidelines for Setting Standards

- After the test
  - Check the results with stakeholders
  - Check to see if the pass rates have reasonable relationships with other markers of competence
  - Check to determine if the results related to future performance

Suggested Readings

- Berk, R.A. (1986). A consumer's guide to setting performance standards on criterion-referenced tests. Review of Educational Research, 56, 137-172.
- Jaeger, R.M. (1989). Certification of student competence. In R.L. Linn (Ed.), Educational Measurement. New York: American Council on Education and Macmillan Publishing Company.
- Kane, M. (1994). Validating the performance standards associated with passing scores. Review of Educational Research, 64, 425-461.
- Livingston, S.A. and Zeiky, M.J. (1982). Passing scores: A manual for setting standards of performance on educational and occupational tests. Princeton, NJ: Educational Testing Service.
- Norcini, J.J. and Guille, R.A. (2002). Combining tests and setting standards. In Norman, G., van der Vleutin, C., and Newble, D. (Eds.): International Handbook of Research in Medical Education (pp. 811-834). Dordrecht: Kluwer Press.