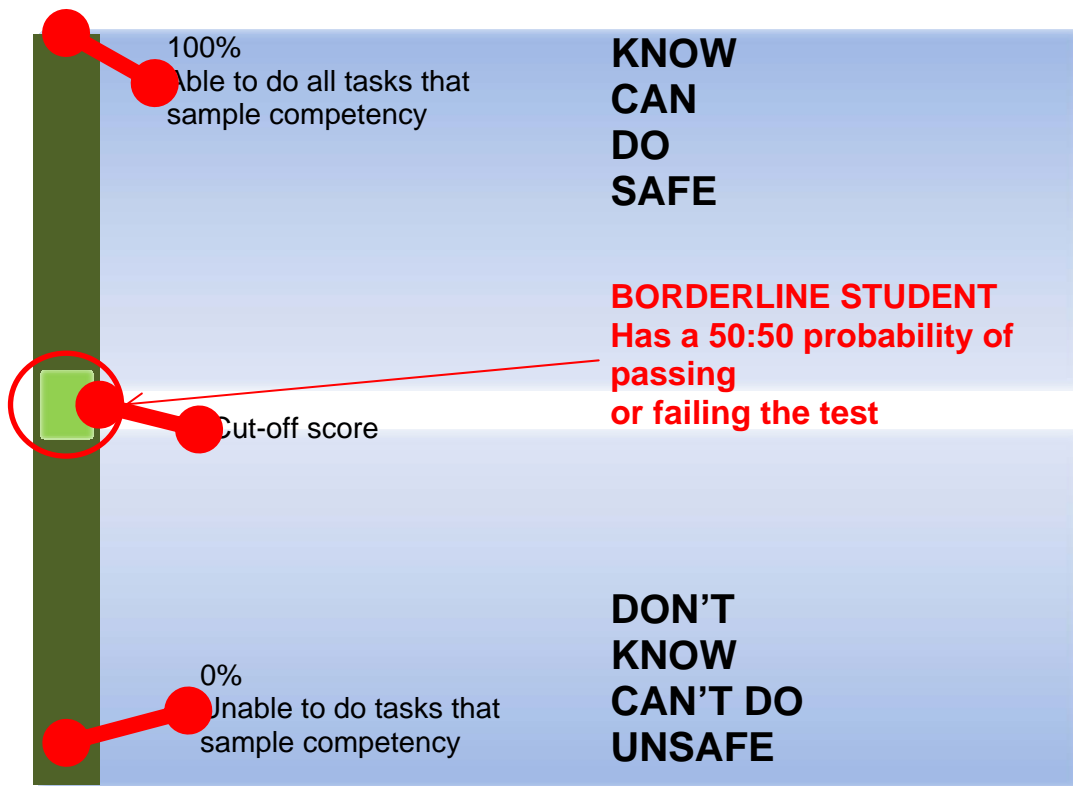


STANDARD SETTING & SCORING PROCEDURES



PERFORMANCE

1. Determining the cut-off score.

- How?
- Any objective criteria?
- The 'truth' is not 'out there'
 - "...even the most rigorous standard-setting method, followed meticulously, will be somewhat arbitrary... however, they be should be credible." – *Schnidler, Corcoran & DaRosa, 2006* –
 - Credible standards share 3 important characteristics:
 - Set by appropriate numbers and types of **judges**
 - Utilize appropriate **method**
 - Produce reasonable **outcomes**

- *Norcini & Guille, 2002* -

2. Credible standards:

- Judges
 - Content experts
 - Know the target population
 - Understand the task and assessment tool
 - Fair-minded
 - Willing to follow directions
 - Give full attention to the process
 - Demographically diverse to avoid bias
 - 5 to 6 considered minimum
- Appropriate method
 - Produces standards that are consistent with the final purpose of the test
 - Relies on informed expert judgment
 - Demonstrates due diligence (demonstrates standard of method)
 - Is supported by a body of evidence
 - Easy to explain and implement *(Norcini & Guille, 2002)*
- Reasonable outcomes
 - Compare with historical standard/external measure
 - Consider stakeholder opinion *(Norcini & Guille, 2002)*

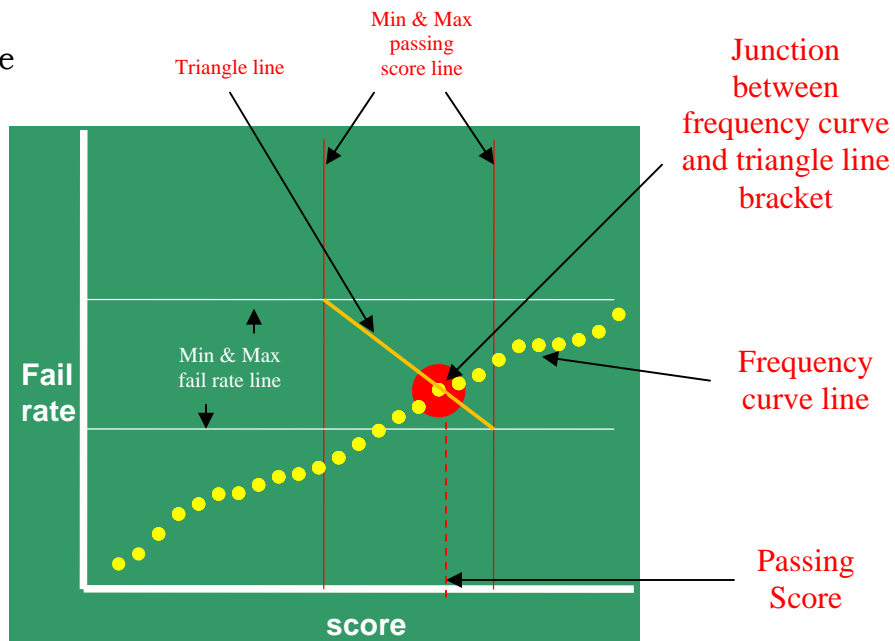
3. Methods available

- Angoff
 - Judges estimate performance of borderline student on each item
 - Judges try to answer the question for e.g. MCQ like borderline student answer the question (judges mimic the borderline student)
 - Mean estimate are combined to produce passing score

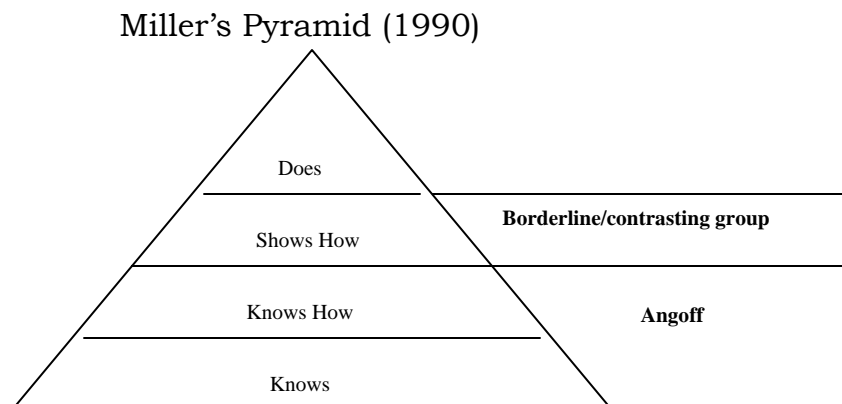
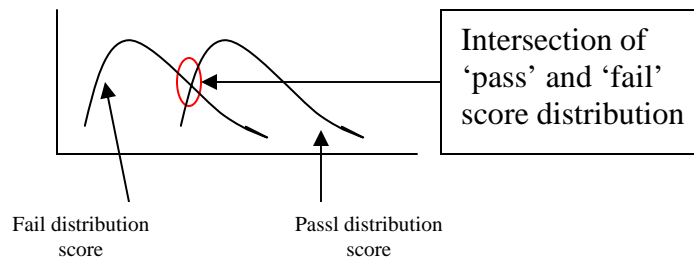
- The passing score is determined before the actual performance (predetermined passing score)
- Ebel
 - Matrix created for test items
 - Item difficulty (actual performance data)
 - Item relevance (judged)
 - Judges borderline performance for each category
 - Passing score calculated
 - Matrix created before the actual performance and passing score calculated after actual performance

	Difficult	Medium	Easy
Essential			
Important			
Acceptable			

- Hofstee



- Judges decide
 - Min and max acceptable pass score
 - Min and max acceptable fail rate
- Plotted and compare with actual performance (fail rate vs. pass score)
- Passing score set where cumulative freq dist crosses this bracketing triangle
- Borderline
 - Judges observe performance
 - Rate using checklist as well as pass/borderline/fail
 - Mean score of 'borderline' is passing mark
- Contrasting groups
 - Judges observe performance
 - Rate using checklist as well as pass/fail
 - Intersection of 'pass' and 'fail' score distribution is the passing mark.



4. Scoring Strategies

- Combination of test means test battery
- Interpretation result means
 - Scoring strategies
 - Related to validity (meaningfulness of score interpretation)
- Validity evidence
 - Content
 - Reliability
 - Consequential validity
- Types of scoring strategies
 - Compensatory
 - Sum of scores over battery of tests compared against a standard
→ pass or fail
 - Multiple measures required OR
 - Minimum performance on ANY measure → complementary
 - Can fail in 1 or more tests in battery but still pass
 - Usage of strategy
 - Content of experts decide that low performance in any single trait can be tolerated
 - Total score meaningfully reflects the construct
 - Reliability of total score high
 - Risk of false positive
 - Less diagnostic value
 - Conjunctive
 - Nonsequential
 - Each test in battery important → must pass separately
 - Less sampling of each trait → lower reliability
 - Enough for high-stakes decision?
 - Risk of false negative
 - More rigorous and demanding

- Effect on student
- Appealing from legislator/authority
- Sequential
 - A series of conjunctive decisions in the test battery for example FRCP examination must pass part I then only can proceed with part II
 - If fail any test → does not proceed
 - Often combined with opportunity to repeat test
 - Strategy useful if battery of test is
 - Time-consuming
 - Expensive
 - May minimize false-negative decisions (failing those who deserve to pass)
- Disjunctive @ complementary
 - Provides a series of equivalent testing alternatives
 - Any of parallel tests valid for making pass/fail decision
 - Emphasizes equivalence of test forms across testing sessions e.g. annual licensing body
 - Allows using initial test as basis for planning improvements for later test → disjunctive test should also provide diagnostic information
 - Can be quite expensive
 - Can be combined with compensatory/conjunctive strategies

5. Summary of strategies

Validity evidence	Compensatory	Conjunctive
Content	Implies overall performance	Emphasizes distinct traits/components
Reliability		
Score	High due to increased sampling	Low due to reduced sampling

Rater	Same	Same
Decision	- High - Risk of false positive	- Lower - Risk of false negative
Consequential	- Less demanding; may induce negative learning behaviour - Loss of diagnostic information	- Highly demanding; may induce negative stress - More diagnostic information

6. Summary by Situation

Situation	Compensatory	Conjunctive	Disjunctive/ Complementary
Measure of different construct		- Increased rigor - low Reliability - False negative	
Different measures for same construct	- Different exams/ item/ types/ times - false positive	- Validating / confirming inferences	- ////////////////
Multiple opportunities			- Minimizing false negative
Accommodation and alternate assessments			- ////////////////

7. Principles

- The manner of combining multiple measures is as important as the measures themselves
 - Should be driven by values to be promoted
- Multiple measures does not necessarily mean higher reliability
 - Especially for conjunctive strategy