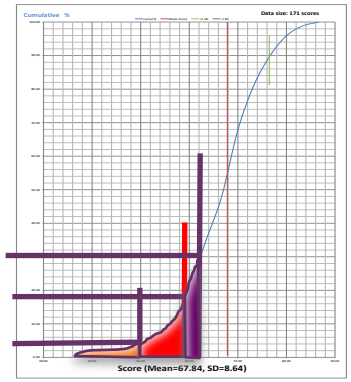
Setting Performance Standards

TLRC Workshop Series 2013



Shekhar Kumta

Yan Jin • Alex Yung • Joseph Leung

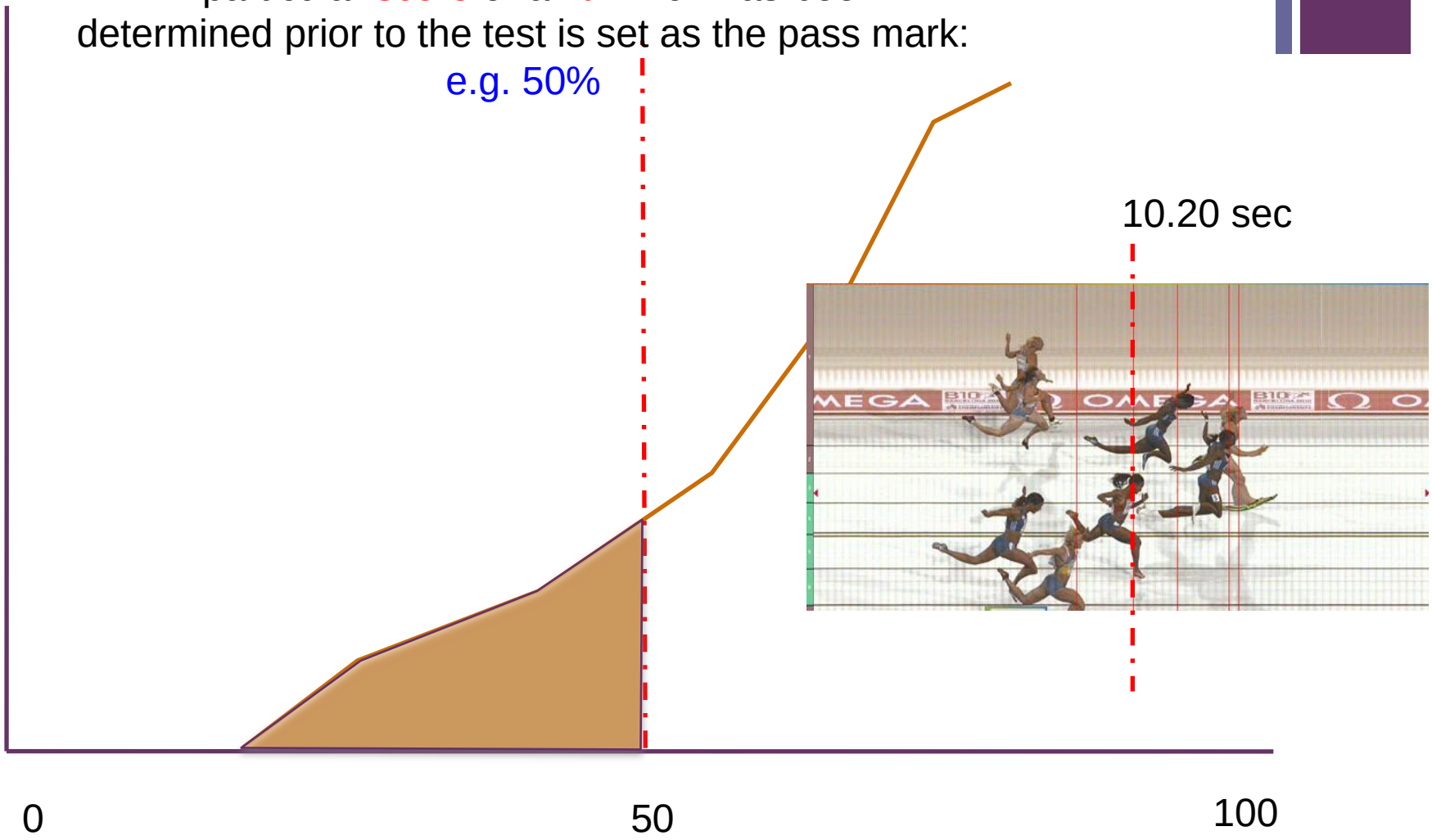
Setting A Performance Standard

- Is a Matter of Policy
- Those who set the standard *must be empowered to do so* under the relevant authority • credentialing body • Institution
- Determining Cut-off Scores on tests is merely the *operational* aspect of this policy

Cut-off Scores → Pass- Fail
Distinction – Vs Good Pass
Competence Certification

ABSOLUTE & FIXED STANDARD

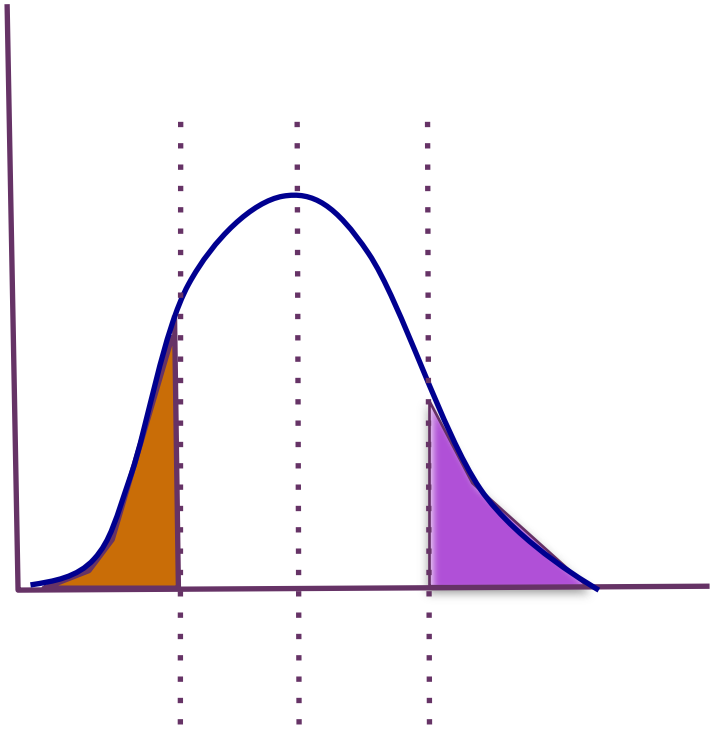
A particular score or a % which has been determined prior to the test is set as the pass mark:
e.g. 50%



Relative & Fixed (example: Rank or Order..)

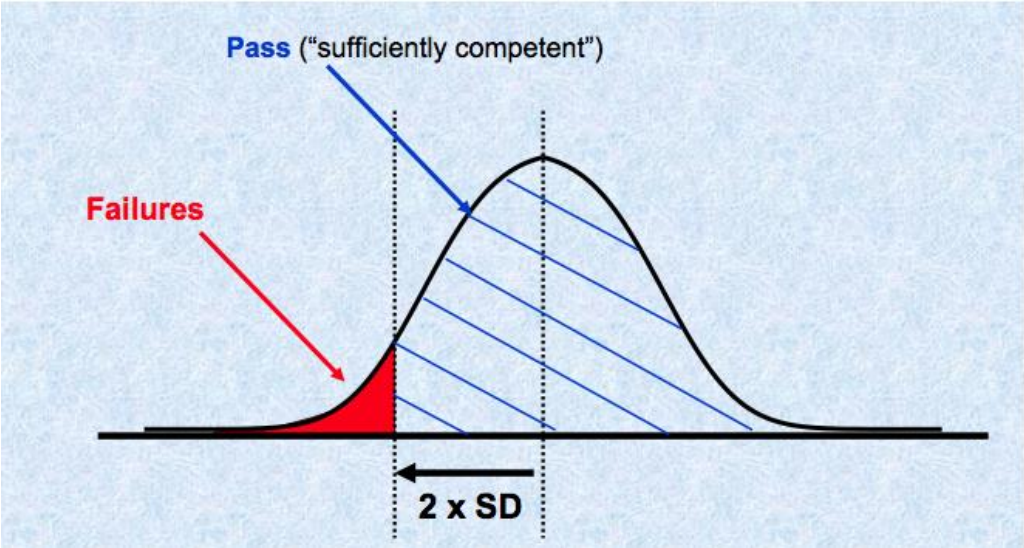
- 1st 3 – Gold Silver Bronze

- Top 25% • Bottom 25%

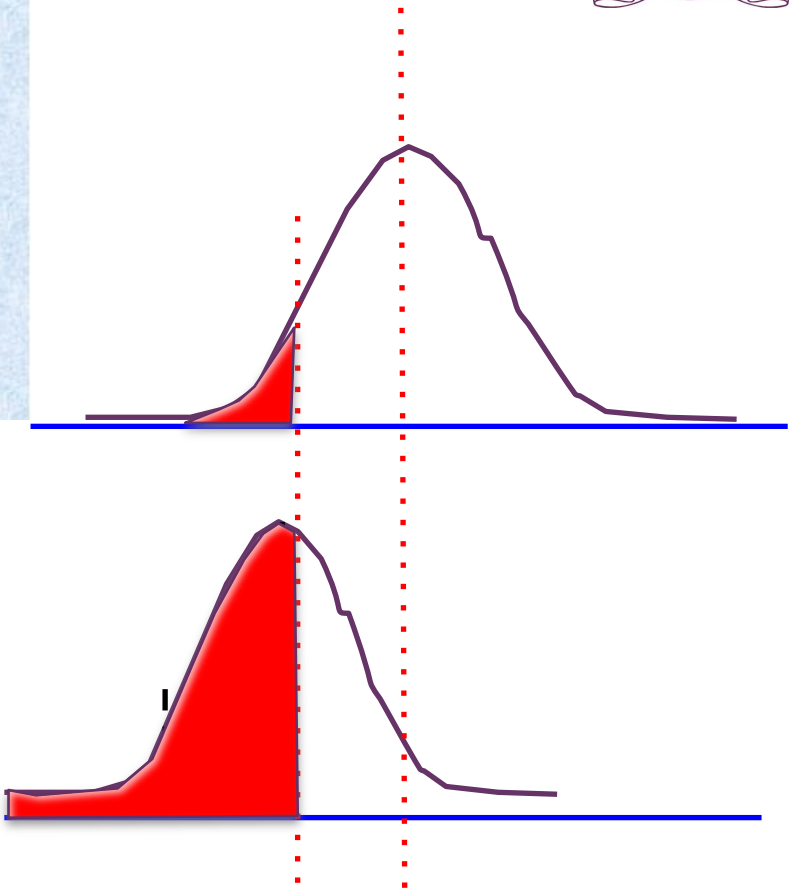


Relative & Not-Fixed

- Cut-off at -2SD from the mean



Mean

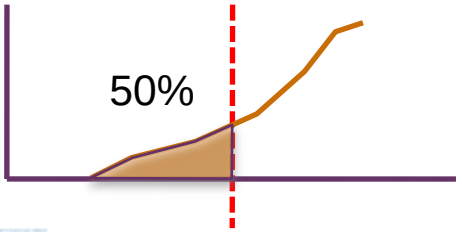


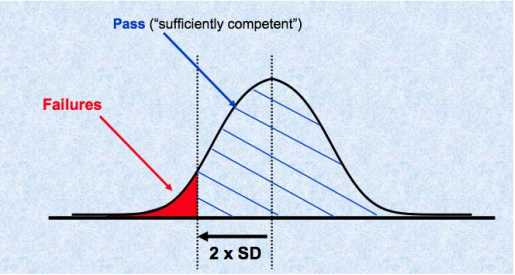
Problems with Historical Approaches to Standards

Absolute & Fixed Standard

Passing score is dependent on minimal mastery of study content but the minimum mastery point is determined *a priori (University policy),*

Ignores error variance due to variation in the quality of teaching & the difficulty of the test





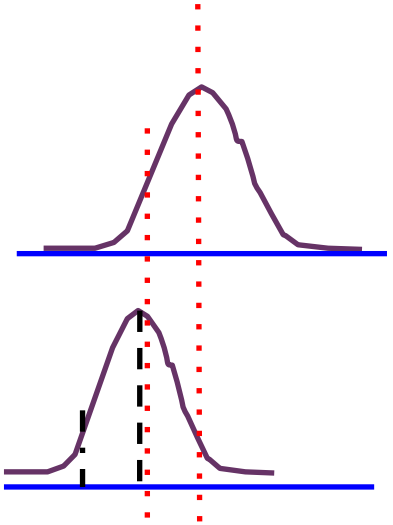
Relative & Non Fixed Standard

Passing score is dependent on performance of the reference group can correct for variation in teaching & assessment quality.

Ignores error variance due to sampling (the reference group),

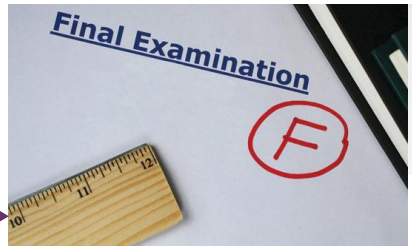
some below P2.5 might have scored over 80%

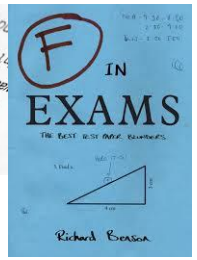
some above P2.5 may have scored even less than 35%



Current Standard Setting Practices in Med.Schools

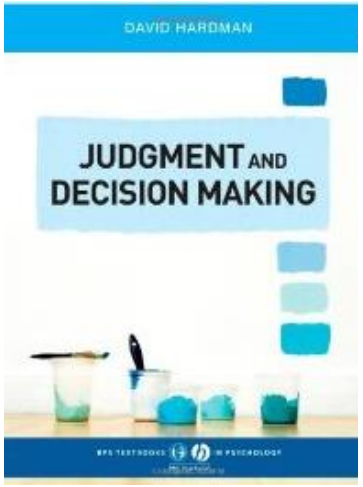
1. Test-centered





2. Examinee-centered

3. A combination of both

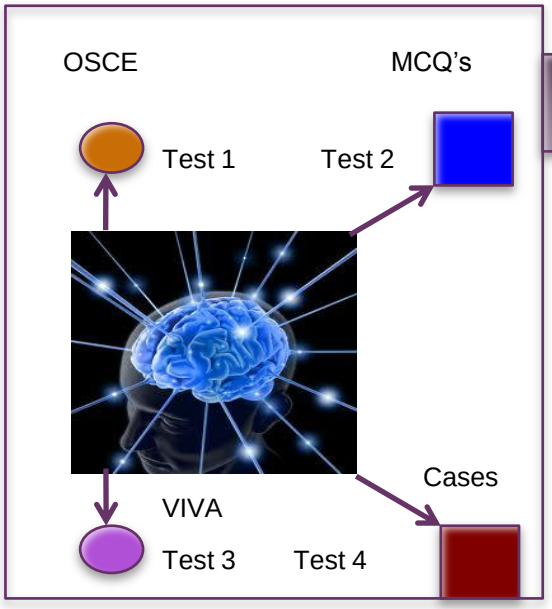


With the judgment of subject matter experts factored in the method of choice

+ Test Theory

An INFERENCE is the interpretation conclusion or meaning that one intends to make about an examinee's underlying UNOBSERVED level of knowledge skill or ability.

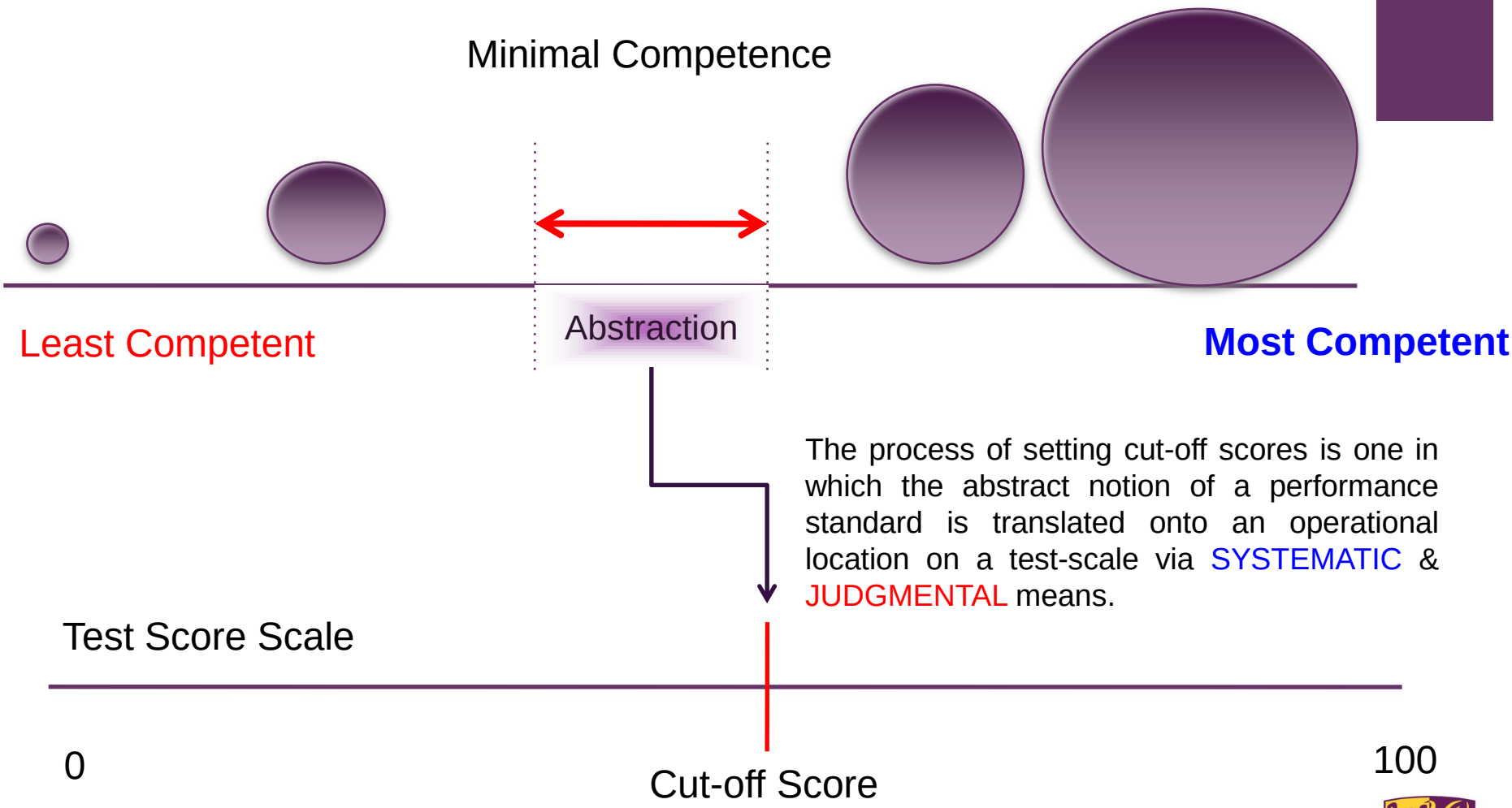
VALIDITY refers to the accuracy of the inferences that one wishes to make about the examinee usually based on observation of the examinee's performance – such as on a test / interview skills /procedural observation.



VALIDITY also requires

- Evidence to support the proposed uses of the test.
- Concern for the extent to which the interpretations are plausible and appropriate (i.e. what is the likely impact?)

Hypothetical Performance Continuum



Key Elements of the Standard Setting Process

1. Establishing Institutional Policy
   - Test Development
   - Which Method
   - Evaluation
2. Appointing the Standard Setters
   - Know Your Subject
   - Know Your Students
   - Know what is being taught – what learning is expected
   - Know TEST method and TEST elements/questions/components
3. Administration and Implementation
   - Test Administration
   - Applying Standards – setting cut-off scores
   - Identifying pass/fail students
4. Assessment of Impact
   - Impact on students
   - Feedback to teachers

Exercise 1.

Form Examination Committees (N= 4).

a) Each Committee will have a CHIEF CENSOR (Coordinate)

a) A DEPUTY CENSOR (Tabulate results and report)

b) Each Standard Setter will be identified i.e. A 1 2 3…n

The MINIMALY Competent also known as the Borderline

One who has the minimum skill and knowledge to perform tasks to an acceptable/defined degree of proficiency

Conceptualize Borderline

Please discuss amongst your group members

+ EBEL

Make 2 Judgment regarding each of the items in the test

a) Difficulty
b) Relevance

Relevance	Difficulty	Number of Items Judged to be in Category (A)	% Of Items that the Minimally Competent are expected to get correct	Product (AxB)
Essential	Easy	94	100%	9400
	Medium	0		
	Hard	0		
	Subtotal	94		
Important	Easy	106	90%	9540
	Medium	153	70%	10710
	Hard	0		
	Subtotal	259		
Acceptable	Easy	24	80%	1920
	Medium	49	60%	2940
	Hard	52	40%	2080
	Subtotal	125		
Questionable	Easy	4	70%	280
	Medium	11	50%	550
	Hard	7	30%	210
	Subtotal	22		
TOTALS		**500**		**37630**
		Passing Percentage (Cx) = 37630/500 = 75.46 %		

Exercise 2. **EBEL**

Please apply the EBEL method to categorize Item Relevance and Item Difficulty using the “Know Hong Kong History” Paper.

You may use

a) Individual Method—i.e. use your own judgment to make the relevance and difficulty rating.

a) Use consensus to arrive at conclusions

EBEL

Relevance	Difficulty	Number of Items Judged to be in Category (A)	% Of Items that the Minimally Competent are expected to get correct	Product (AxB)
Essential	Easy			
	Medium			
	Hard			
	Subtotal			
Important	Easy			
	Medium			
	Hard			
	Subtotal			
Acceptable	Easy			
	Medium			
	Hard			
	Subtotal			
Questionable	Easy			
	Medium			
	Hard			
	Subtotal			
TOTALS		**10**		
			Passing Percentage (Cx) =	

EBEL RESULTS

Group A	Group B	Group C	Group D
Actual Performance Data:			
Mean Score	SD	Range	

EBEL

Robert Ebel 1972: Essentials of Educational Measurement

DISADVANTAGES

Difficulty

Difficult to keep the 2 dimensions separate and distinct – as they may be highly correlated

Relevance

ALTERNATIVES

Judgment may be substituted by real Item Values

- *p* values

Hard	0.00 to 0.49
Medium	.50 to .79
Easy	.80 >

Items judged as Questionable should not be accepted in a certifying examination

Angoff Method

(Has been most often used in Medical Education)

A group of experts pass judgment on the proportion of minimally competent (borderline) candidates who *could correctly answer an item* (*or could correctly perform a procedure in an OSCE station*).

- Use a minimum of 8 judges

  (but 12-18 judges are needed to be safe: Margolis et al)

- Each judge must keep in mind a commonly agreed upon definition for a hypothetical borderline candidate
- Judges' estimates are averaged for each item
- Cutoff point is set as sum of these averages.

Angoff's – Absolute (competence)

	J1	J2	J3	J4	J5	J6	Average of Judges Score
Q1.	0.85	0.80	0.80	0.95	0.85	0.90	= 0.86
Q2.	0.60	0.70	0.50	0.55	0.65	0.70	= 0.62
Q3.	0.45	0.55	0.50	0.60	0.90	0.35	= 0.56
Q4.	0.90	0.95	0.90	0.95	0.85	0.90	= 0.91
Q5.	0.80	0.75	0.65	0.70	0.85	0.55	= 0.72
Q6.	0.70	0.65	0.60	0.70	0.75	0.60	= 0.67
Q7.	0.40	0.50	0.35	0.50	0.55	0.50	= 0.47
Q8.	0.75	0.65	0.60	0.70	0.75	0.60	= 0.68
Q9.	0.65	0.55	0.70	0.65	0.65	0.60	= 0.63
Q10.	0.55	0.50	0.45	0.60	0.65	0.55	= 0.55
Q11.	0.50	0.45	0.40	0.50	0.55	0.50	= 0.48
Q12	0.95	0.95	0.95	0.90	0.95	0.95	= 0.94
Cutpoint							= 8.09

Sum

Getting experts to agree & set standards is not easy

Can be time consuming for long tests

Exercise 3.

GROUP A & B

Perform the ANGOFF on the Examination paper provided.

This is a panel examination for PMUS Yr-3

Students who fail need to re-sit before they can be promoted.

Exercise 4.

Group C & D

Perform the ANGOFF on the OSCE examination stations.

This is a year-end examination.

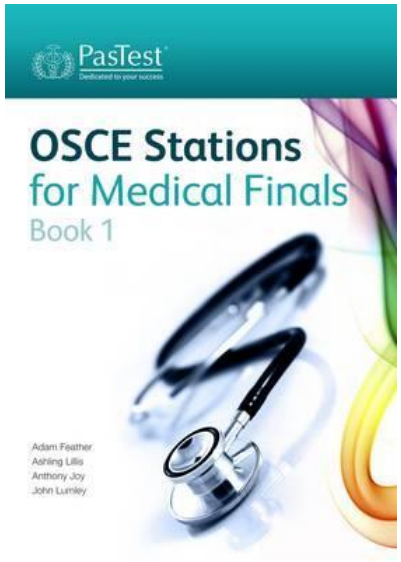
Students who fail need to attend a supplementary examination before they can be promoted.

	OSCE Stations	What proportion of Borderline Candidates are expected to pass this station ?
1	P/E - hip examination with x-ray	
2	History taking - distal thigh pain	
3	P/E - Radial nerve palsy	
4	P/E - management of sciatica	
5	Practical skills - management of unconscious patient	
6	Practical skills - assessment and management of burn wound	
7	Practical skills - hand hygiene	
8	Written - video clip of colonic tumour	
9	Written - medical devices	
10	History taking - a patient with dysphagia	
11	Practical skills - abdominal examination in a difficult surrogate patient	
12	Practical skills - suturing of banana skin	
13	History taking - a patient with acute pancreatitis	
14	Written - x-ray interpretation of adhesive intestinal obstruction	
15	Written - interpretation of ruptured HCC with abdominal CT	
16	Written - management of post-operative fast AF	
17	Written - aortic dissection	
18	Written - x-ray interpretation and management of hip and pelvic fracture	
19	Written - anterior resection	
20	Written - ethics or professional conduct	

OSCE's

- Consider Pass/Fail Cut-off for Each Station

- Borderline • Borderline regression • or Contrasting groups



Pass/Fail Cut-off for Overall OSCE Exam

- Judges Estimate the difficulty level of each station (*Angoff Style*)
- Sum of *p values* is rounded off as cut-off for number of stations required to pass the OSCE

Angoff Modifications & Extensions

- Judges make a second round of discussion
- Judges are given “performance” data on which they may base their judgments
- Judges use the YES/NO method instead of making p-value or proportionality judgments

ANGOFF RESULTS - PAPER

Group A	Group B	Group C	Group D
Actual Performance Data:			
Mean Score	SD	Range	
67		1-99	

ANGOFF RESULTS - OSCE

Group A	Group B	Group C	Group D
Actual Performance Data:			
Mean Score	SD		
73.85			

M5 Final Results-OSCE-IA				
Question	**Station**	**Dif. Level**	**Dis Index**	**Biserial**
1	P/E - hip examination with x-ray	79	13	44
2	History taking - distal thigh pain	73	11	34
3	P/E - Radial nerve palsy	79	15	48
4	P/E - management of sciatica	74	10	44
5	Practical skills - management of unconscious patient	63	19	48
6	Practical skills - assessment and management of burn wound	60	8	26
7	Practical skills - hand hygiene	81	7	34
8	Written - video clip of colonic tumour	82	14	37
9	Written - medical devices	80	18	40
10	History taking - a patient with dysphagia	68	8	22
11	Practical skills - abdominal examination in a difficult surrogate patient	90	8	45
12	Practical skills - suturing of banana skin	54	20	50
13	History taking - a patient with acute pancreatitis	75	9	25
14	Written - x-ray interpretation of adhesive intestinal obstruction	65	10	40
15	Written - interpretation of ruptured HCC with abdominal CT	89	9	30
16	Written - management of post-operative fast AF	77	10	37
17	Written - aortic dissection	65	8	30
18	Written - x-ray interpretation and management of hip and pelvic fracture	76	12	40
19	Written - anterior resection	74	11	47
20	Written - ethics or professional conduct	73	8	18
		73.85		

Angoff – Modified- The Yes-No method

	J1	J2	J3	J4	J5	J6	Average of Judges Score
Q1	1	0	1	0	0	1	0.50
Q2.							
Q3.							

1 for Yes
0 for No

Judges make a judgment about whether or not a borderline student will be able to answer each question correctly.

This modification makes it easy for judges to make judgments rather than assigning probabilities

Cutpoint = 0.809

Exercise 5

- Perform a Modified ANGOFF Standard Set using the Yes/No method.

A hypothetical Exam paper has been provided to all.

Group A	Group B	Group C	Group D
Actual Performance Data:			
Mean Score	SD	EBEL Score	
E			

+	OSCE Stations	What proportion of Borderline Candidates are expected to pass this station ?				Actual Performance
		Group A	Group B	Group C	Group D	
1	P/E - hip examination with x-ray					79
2	History taking - distal thigh pain					73
3	P/E - Radial nerve palsy					79
4	P/E - management of sciatica					74
5	Practical skills - management of unconscious patient					63
6	Practical skills - assessment and management of burn wound					60
7	Practical skills - hand hygiene					81
8	Written - video clip of colonic tumour					82
9	Written - medical devices					80
10	History taking - a patient with dysphagia					68
11	Practical skills - abdominal examination in a difficult surrogate patient					90
12	Practical skills - suturing of banana skin					54
13	History taking - a patient with acute pancreatitis					75
14	Written - x-ray interpretation of adhesive intestinal obstruction					65
15	Written - interpretation of ruptured HCC with abdominal CT					89
16	Written - management of post-operative fast AF					77
17	Written - aortic dissection					65
18	Written - x-ray interpretation and management of hip and pelvic fracture					76
19	Written - anterior resection					74
20	Written - ethics or professional conduct					73

+	OSCE Stations	What proportion of Borderline Candidates are expected to pass this station ?				Actual Performance
1	P/E - hip examination with x-ray					79
2	History taking - distal thigh pain					73
3	P/E - Radial nerve palsy					79
4	P/E - management of sciatica					74
5	Practical skills - management of unconscious patient					63
6	Practical skills - assessment and management of burn wound					60
7	Practical skills - hand hygiene					81
8	Written - video clip of colonic tumour					82
9	Written - medical devices					80
10	History taking - a patient with dysphagia					68
11	Practical skills - abdominal examination in a difficult surrogate patient					90
12	Practical skills - suturing of banana skin					54
13	History taking - a patient with acute pancreatitis					75
14	Written - x-ray interpretation of adhesive intestinal obstruction					65
15	Written - interpretation of ruptured HCC with abdominal CT					89
16	Written - management of post-operative fast AF					77
17	Written - aortic dissection					65
18	Written - x-ray interpretation and management of hip and pelvic fracture					76
19	Written - anterior resection					74
20	Written - ethics or professional conduct					73

The Hofstee or Compromise Method

- In *1979* the passing score on a test had to be lowered to 45% and even then only 55% of students passed.
- In *1980* the passing score was set to 60% and over 90% of students passed.
- *Teachers, teaching materials, tests were essentially the same*.
- The "Normative" expectations of the 1st year did not match the performance of the subsequent year – i.e. the "correctness" of the standard was questionable.
- The Hofstee method aims to strike a balance between Normative and Criterion referenced information

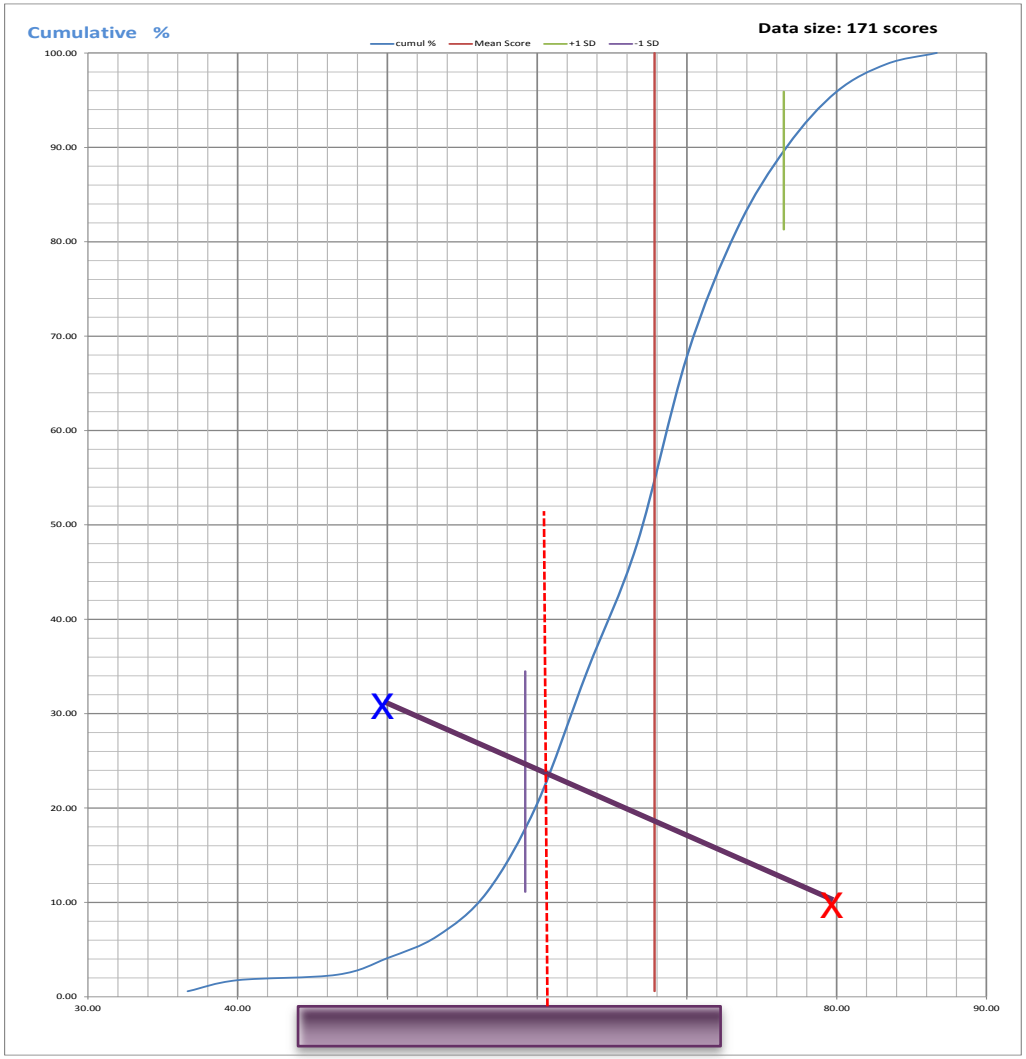
Key Tasks required of Standard Setters

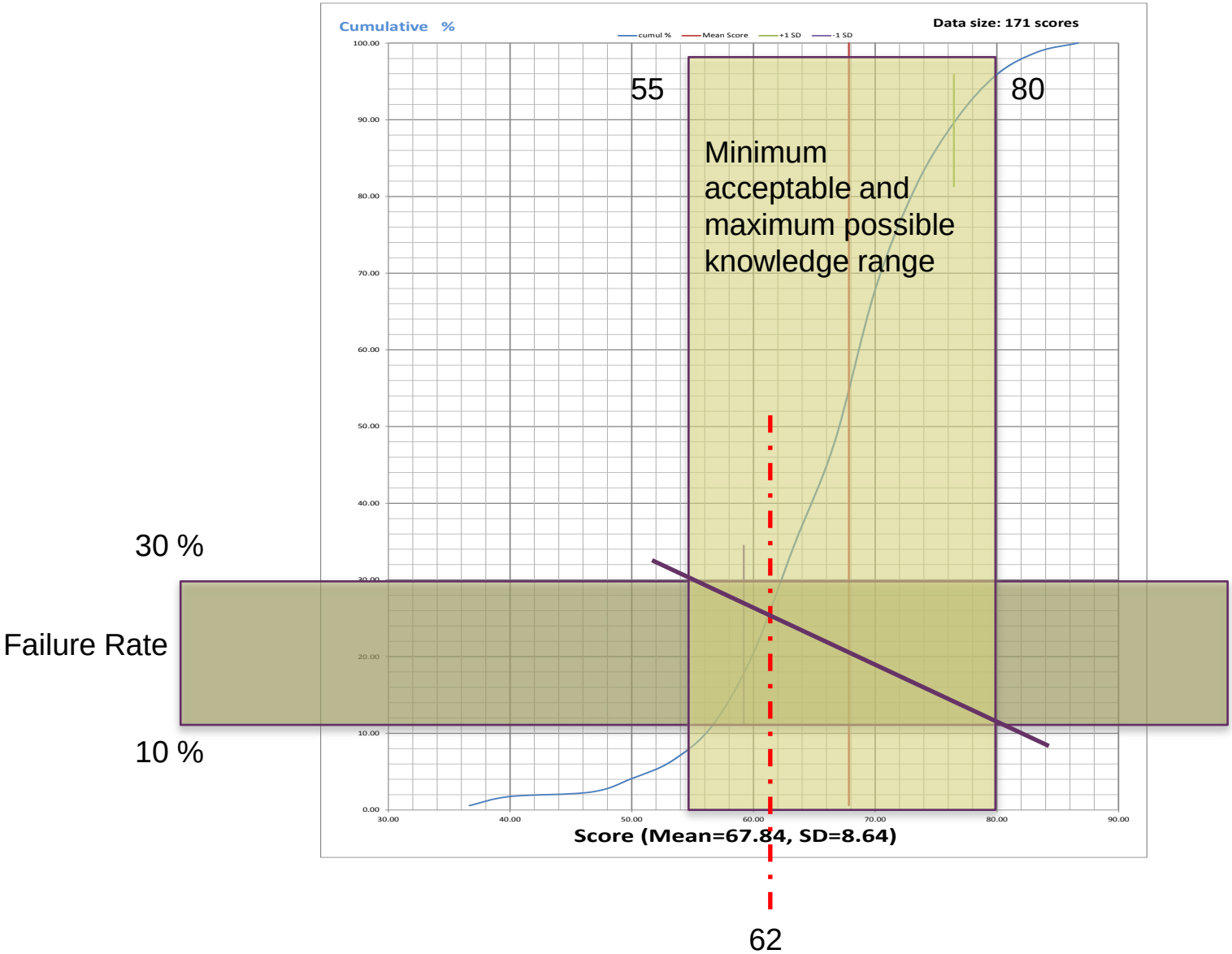
- k_{max}
  - What is the highest % correct score that would be acceptable even if every examinee attains that score –
- k_{min}
  - What is the lowest % correct score that would be acceptable even if no examinee attains that
- f_{max}
  - What is the maximum acceptable failure %?
- F_{min}
  - What is the minimum acceptable failure %?

Hoftsee Standard Setting Method:

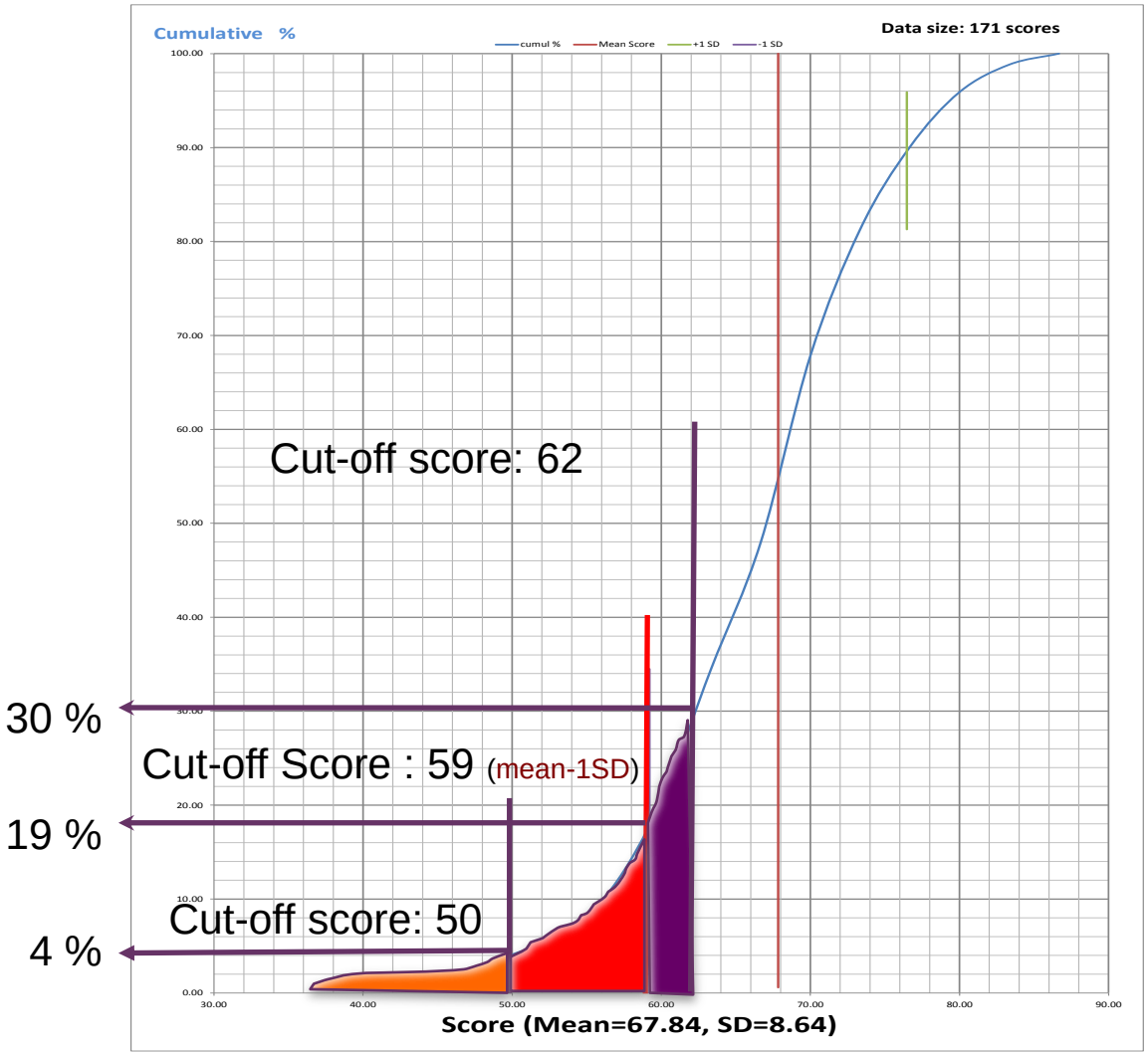
K_{max} : 80
F_{min} : 10

K_{min} : 50
F_{max} : 30

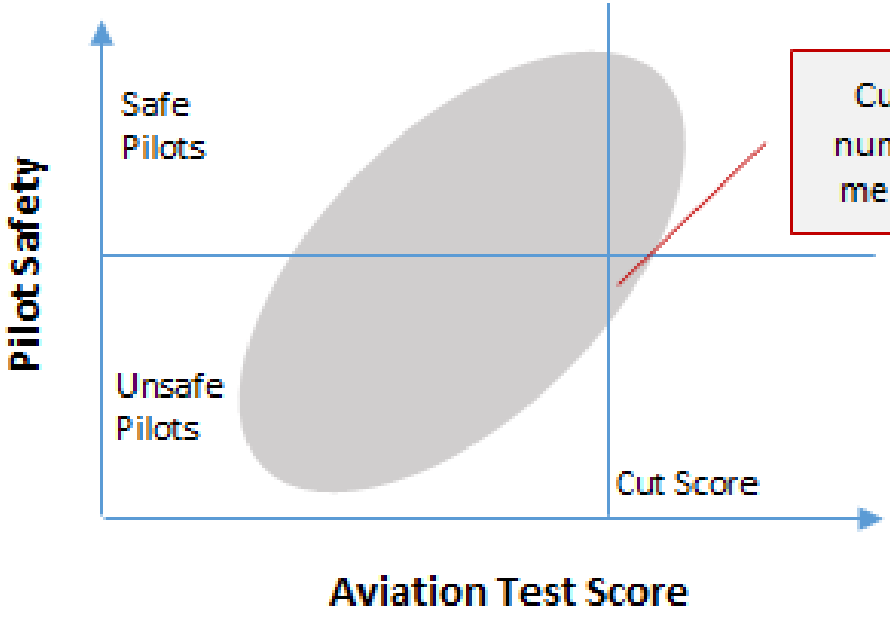



Cumulative %
Data size: 171 scores
cumul %
Mean Score
+1 SD
-1 SD
55
80
Minimum acceptable and maximum possible knowledge range
30 %
Failure Rate
10 %
Score (Mean=67.84, SD=8.64)
62

Hoftsee Applied to a Real Exam Score Distribution: CUHK PMUS-2



Aviation Safety



Cut score is set to minimize the number of unsafe pilots, though it means excluding some safe pilots

Borderline - Absolute

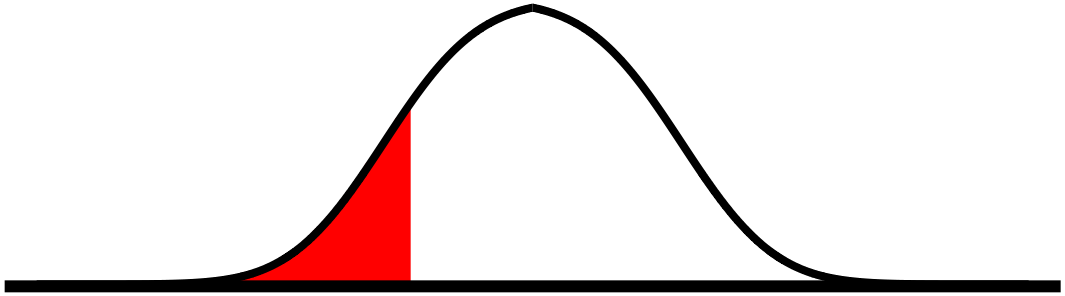
Checklist

1. Ησ σηφσ σφνησσ σφησ σφσ σφ ☑
2. Κσκσ σκσμσιθοπθλ θλθμθ θ θ θκλ ☑
3. Λαλκα κδμ δδκκ δλκλ δλλδ ☐
4. Κεψω δδ ε ρ ρρμτ τμκ ☐
5. ϑφφφκ δδ ☑
6. Ησκλ;σ σκφ σλσ σκα ακ ακλ αλδ ☑
7. Ηδηηδδη σησ αηηακκ ασ ☐

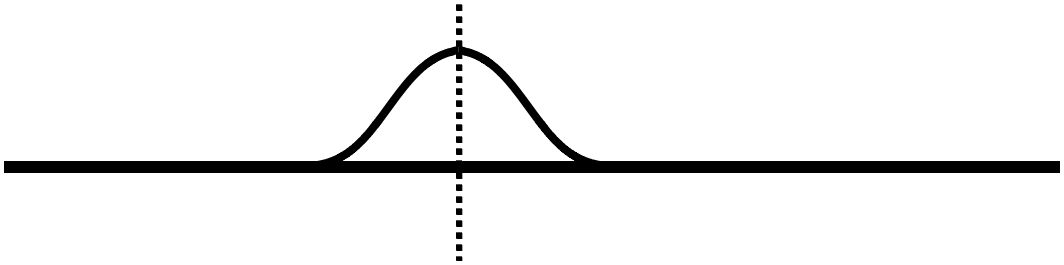
TOTAL Σ

Pass, Fail, Borderline P/B/F

Test score distribution



Borderline score distribution



Passing score

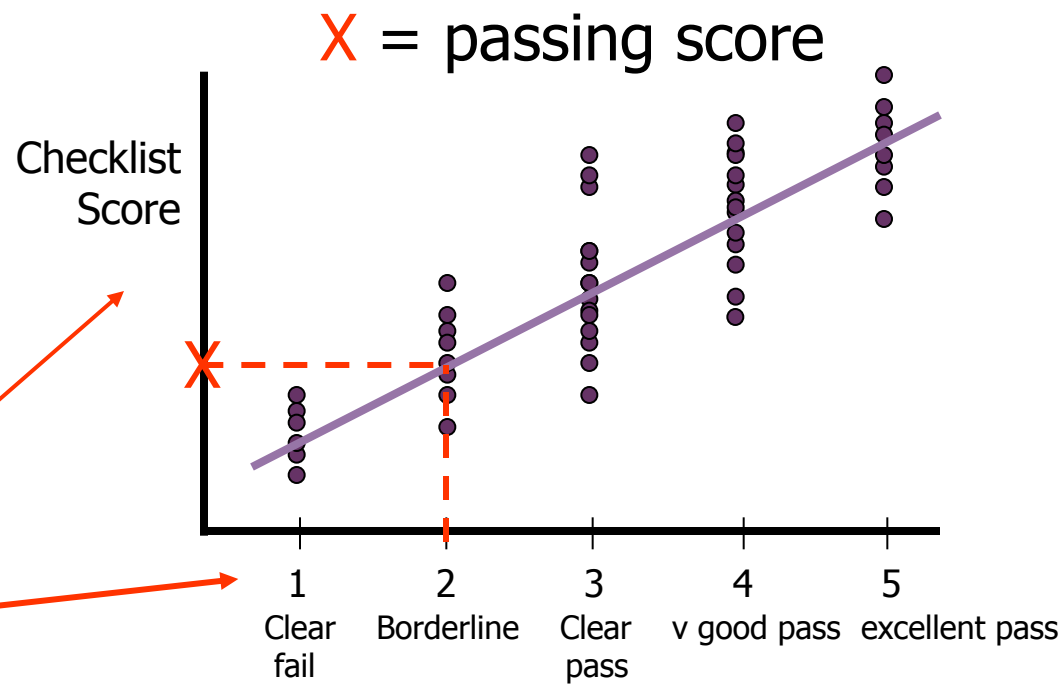
Regression – Absolute Method

Checklist

1. Ησ σηφσ σφνησσ σφησ σφσ σφ ☑
2. Κσκσ σκσμσιθοπθλ θλθμθ θ θ θκλ ☑
3. Λαλκα κδμ δδκκ δλκλ δλλδ ☐
4. Κεψω δδ ε ρ ρρμτ τμκ ☐
5. Θφφφκ δδ ☑
6. Ησκλ;σ σκφ σλσ σκα ακ ακλ αλδ ☑
7. Ηδηηδδη σησ αηηακκ ασ ☐

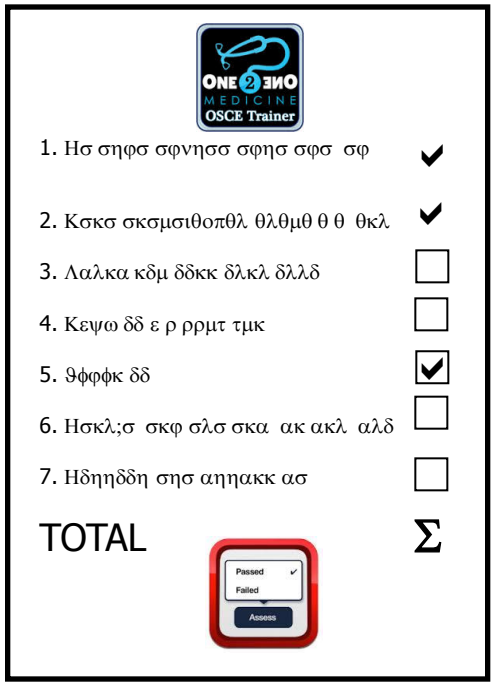
TOTAL Σ

Overall rating 1 2 3 4 5



1 = Clear fail
2 = Borderline
3 = Clear pass
4 = v good pass
5 = excellent pass

OSCE Checklists and Scoring Schemes



Do not have a good enough spread to discriminate a wide-range of performances

Borderline scores tend to be lower and thus the pass-fail cut-off is also unrealistically low

Scores need to reflect meaningful tasks and the performance characteristics required for competency should be well described

Standard Setting: A policy matter

- No Single Method is “golden”
- Policy makes decisions consistent and defensible.
- Test Development Process
  - Blue-printing
  - Standard Setting
  - Assessment of Impact
  - Post-Hoc Test & Test-Item Analysis