

# Constructing Theory Exam Questions

Kuliyah of Medicine, IIUM (1<sup>st</sup> – 2<sup>nd</sup> July 2025)

Nurhanis Syazni Roslan

Department of Medical Education, School of Medical Sciences, USM



# WORKSHOP FLOW

## **ASSESSMENT VALIDITY**

Tuesday, 0915-1000

## **PRINCIPLES OF CONSTRUCTING THEORY QUESTIONS**

Tuesday, 1000-1300

## **ASSESSMENT BLUEPRINT**

Tuesday, 1415-1500

## **ITEM ANALYSIS**

Tuesday, 1500-1700

## **STANDARD SETTING**

Wednesday, 0900-1230

# ASSESSMENT VALIDITY

0915-1000

At the end of the session, participants  
will be able to

identify key source of validity evidence  
for high stake examination.







**WHY WE  
ASSESS  
STUDENTS?**



# Assessment in Medical Education

Protect the public and patient

## FORMATIVE

- Guiding future learning
- Providing reassurance
- Promoting reflection
- Shaping values

## SUMMATIVE

Making an overall judgment

- Competence
- Fitness to practice
- Advancement to higher levels of responsibility

- Higher stake
- Lower stake

# IS MY ASSESSMENT VALID?

MMC Standards for Undergraduate  
Medical Education (2019)

2.2.1(b) Show evidence that the variety  
of the assessment methods are **valid** to  
measure the learning outcomes and  
competencies.





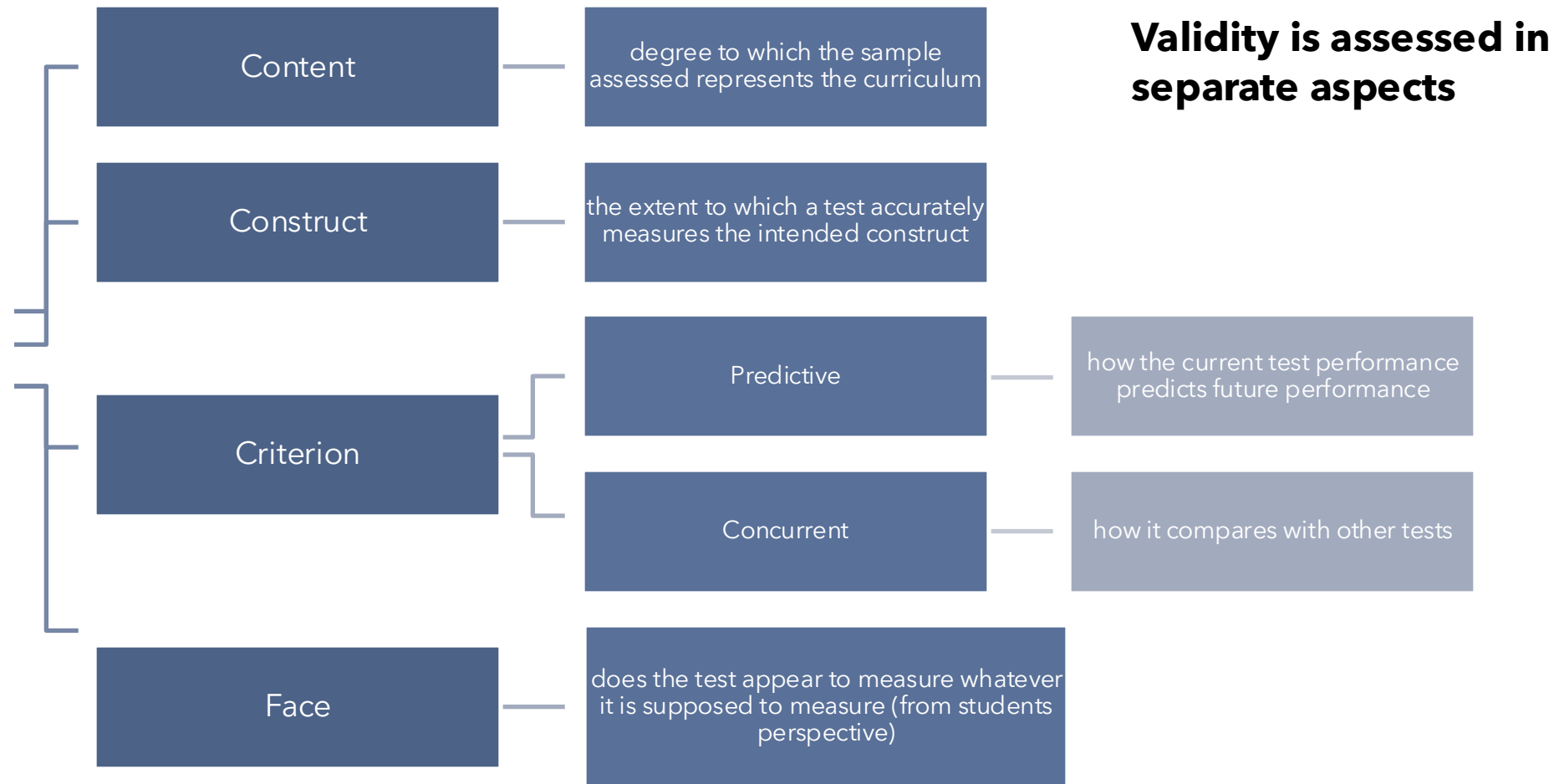


## Professional Exam 2

Test reliability: 0.88

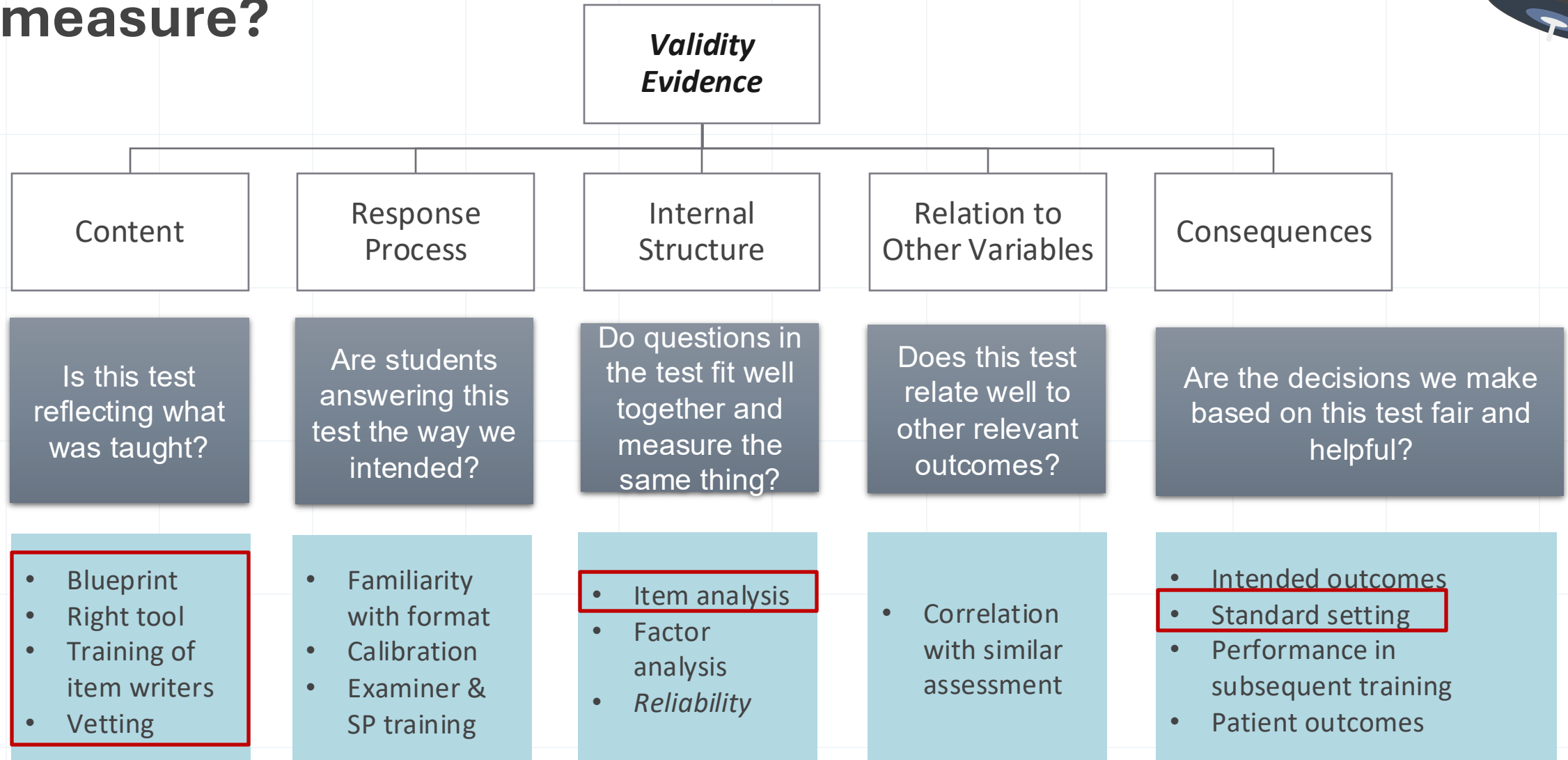
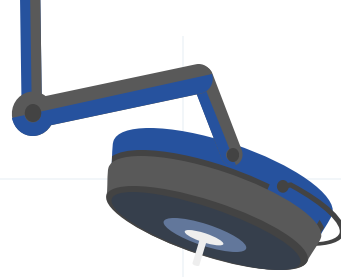
*Implied validity?*

# Traditional paradigm of assessment validity





# Is our assessment measures what it suppose to measure?



# ASSESSMENT VALIDITY: PRINCIPLES



## Unitary concept

All aspects of validity evidence have an impact on assessment validity

## Validity is not a fixed label

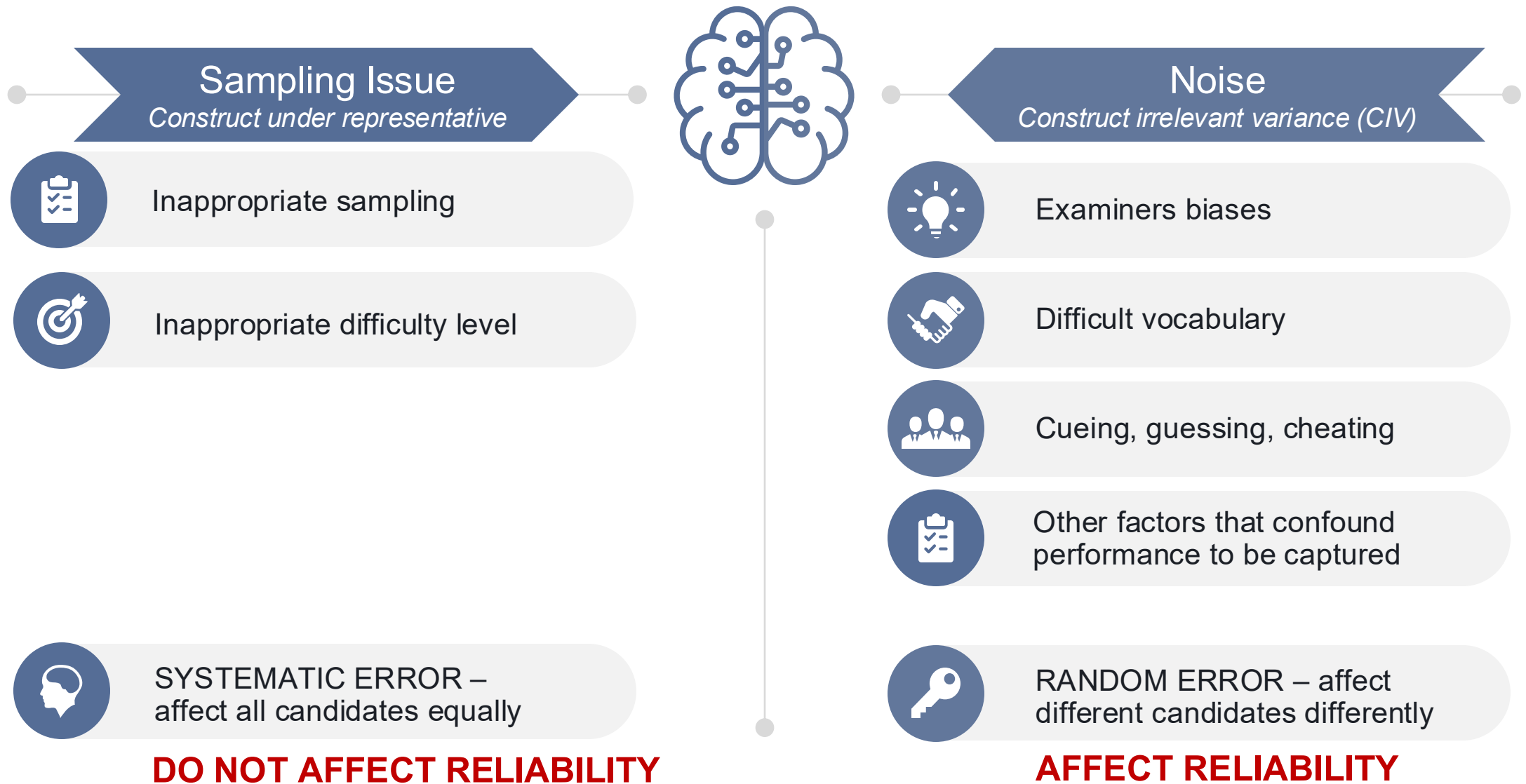
Validity  $\neq$  Tool  
Validity = Context + Evidence

## Validity is established through evidence

More evidence, more valid



# THREATS TO ASSESSMENT VALIDITY



# What reliability can and cannot capture?



## Systematic Error

Caused by consistent, repeatable flaws

Affects results in the same direction

Usually over- or underestimates a student's score

Examples: Uneven topic sampling (very hard or very easy), faulty rubric

	Qs 1	Qs 2 *very hard	Qs ... *very hard	Qs 50 *very hard
A (good)	8	4	4	5
B (borderline)	4	2	2	3
C (weak)	0	1	0	1

Reduces **validity**

Can be corrected by improving assessment design



## Random Error

Caused by unpredictable fluctuations

Affects results inconsistently

May increase or decrease a student's score

Examples: distraction, guesswork, fatigue

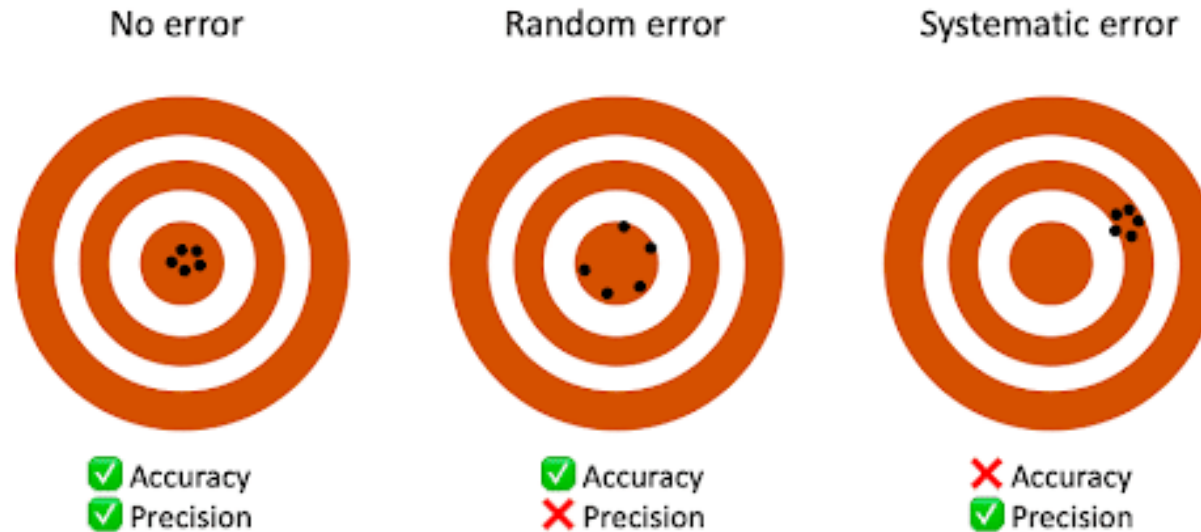
	Qs 1	Qs 2	Qs ...	Qs 50
A (good)	8	9	9	2 *tired
B (borderline)	4	5	5	7
C (weak)	0	2	2	5

Reduces **reliability AND validity**

Can be minimized by increasing number of items



# RELIABILITY DOES NOT IMPLY VALIDITY



# When validity should be investigated (at least)

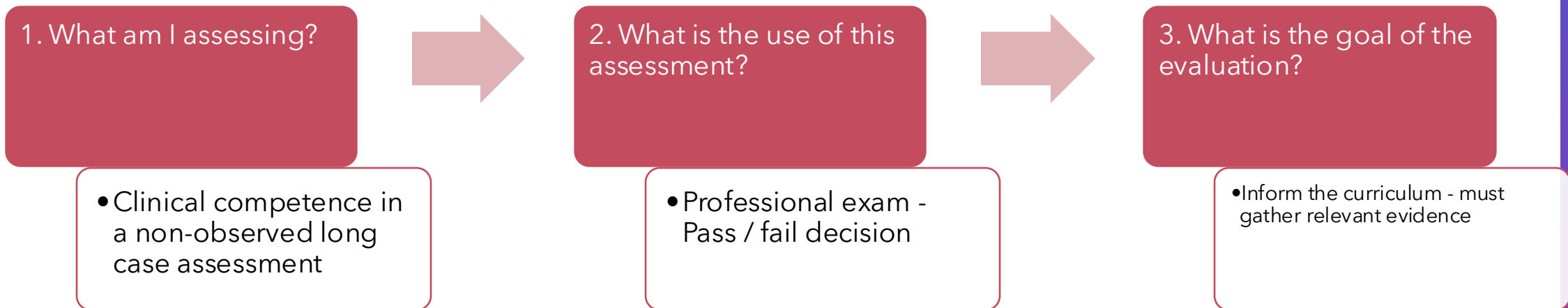
- + Ideally in every high-stake exam
- + Evaluation of new assessment tool
- + Curriculum review
- + Incoherence between students assessment scores and outcomes

NEW  
**STRAITS**ESTABLISHED 1948**TIMES**

**BUSINESS TIMES**

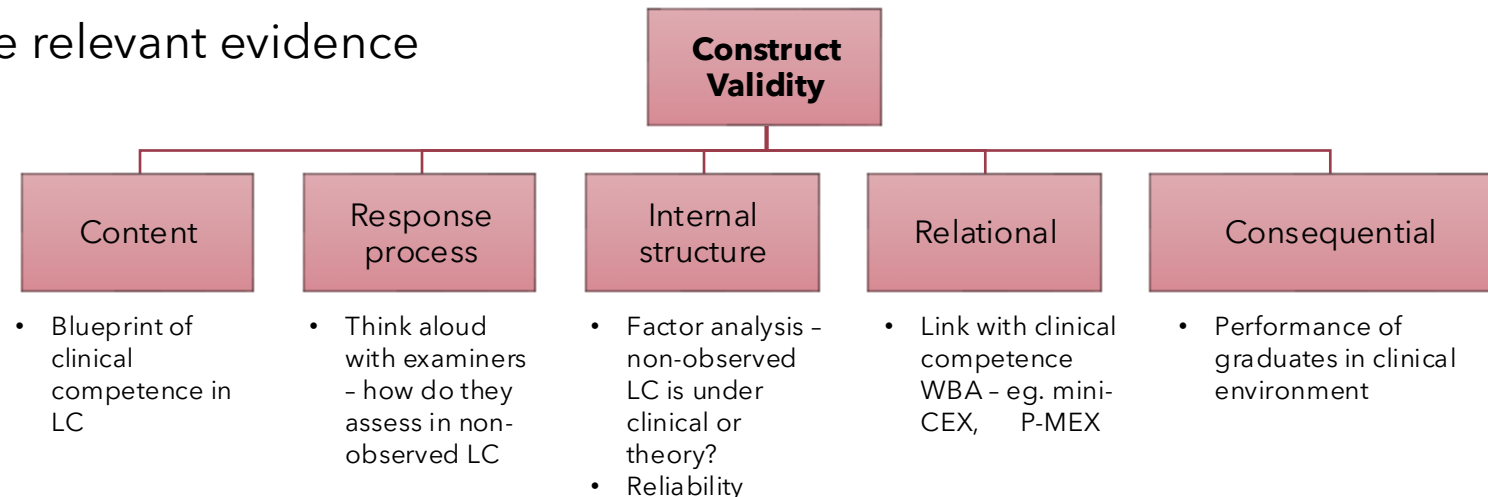
**Panel denies claims of misconduct  
in USM's exam process for  
Neurosurgery programme**

# Example



- Prior uncertainty: Is the issue genuinely in doubt? Yes
- Information yield: Decision to maintain non-observed or change to observe LC
- Cost: How expensive is the investigation in time and dollars? Secondary data
- Leverage: How critical is the information for achieving consensus in the relevant audience? Very important

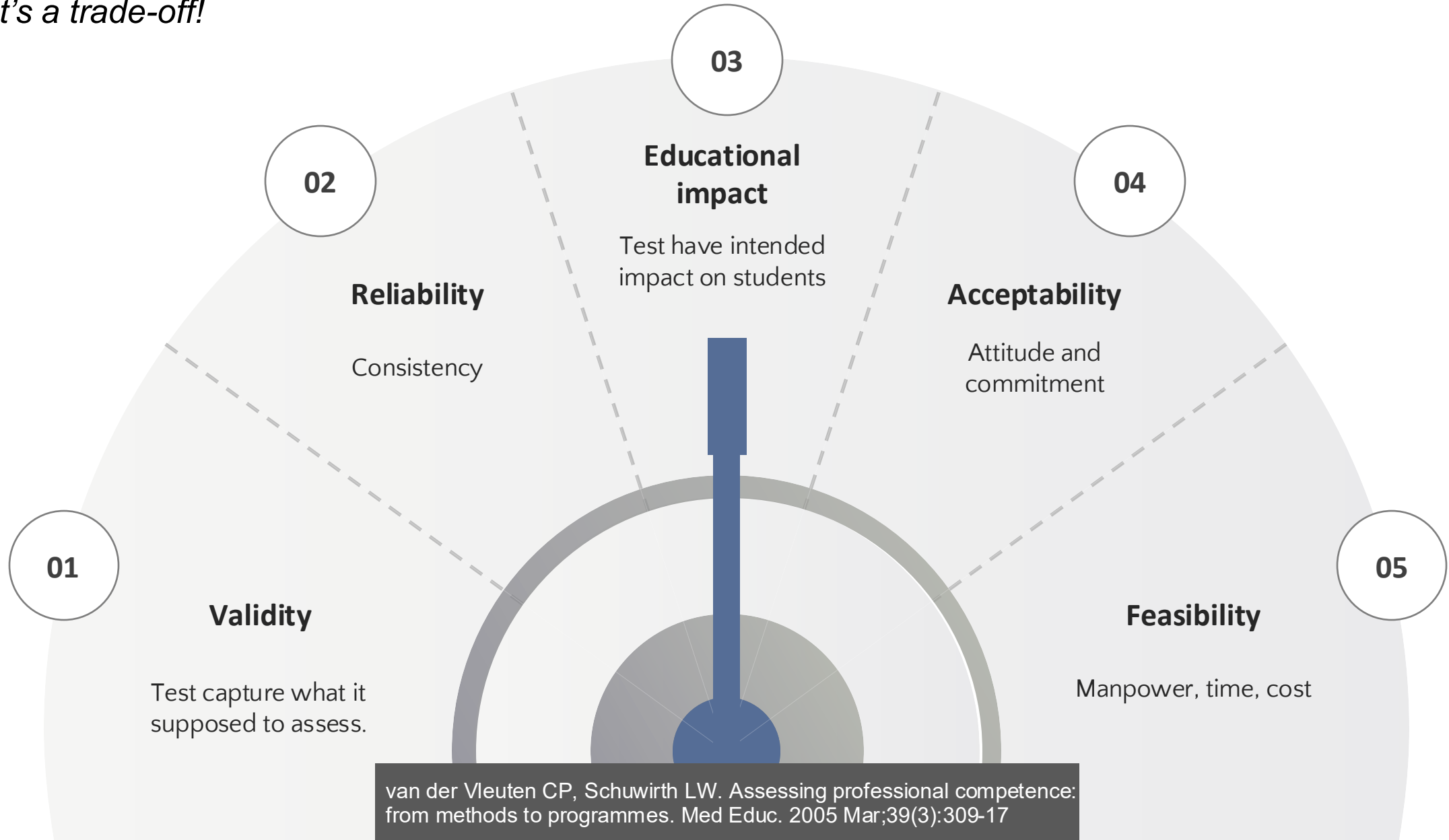
## 4. What are the relevant evidence to gather?



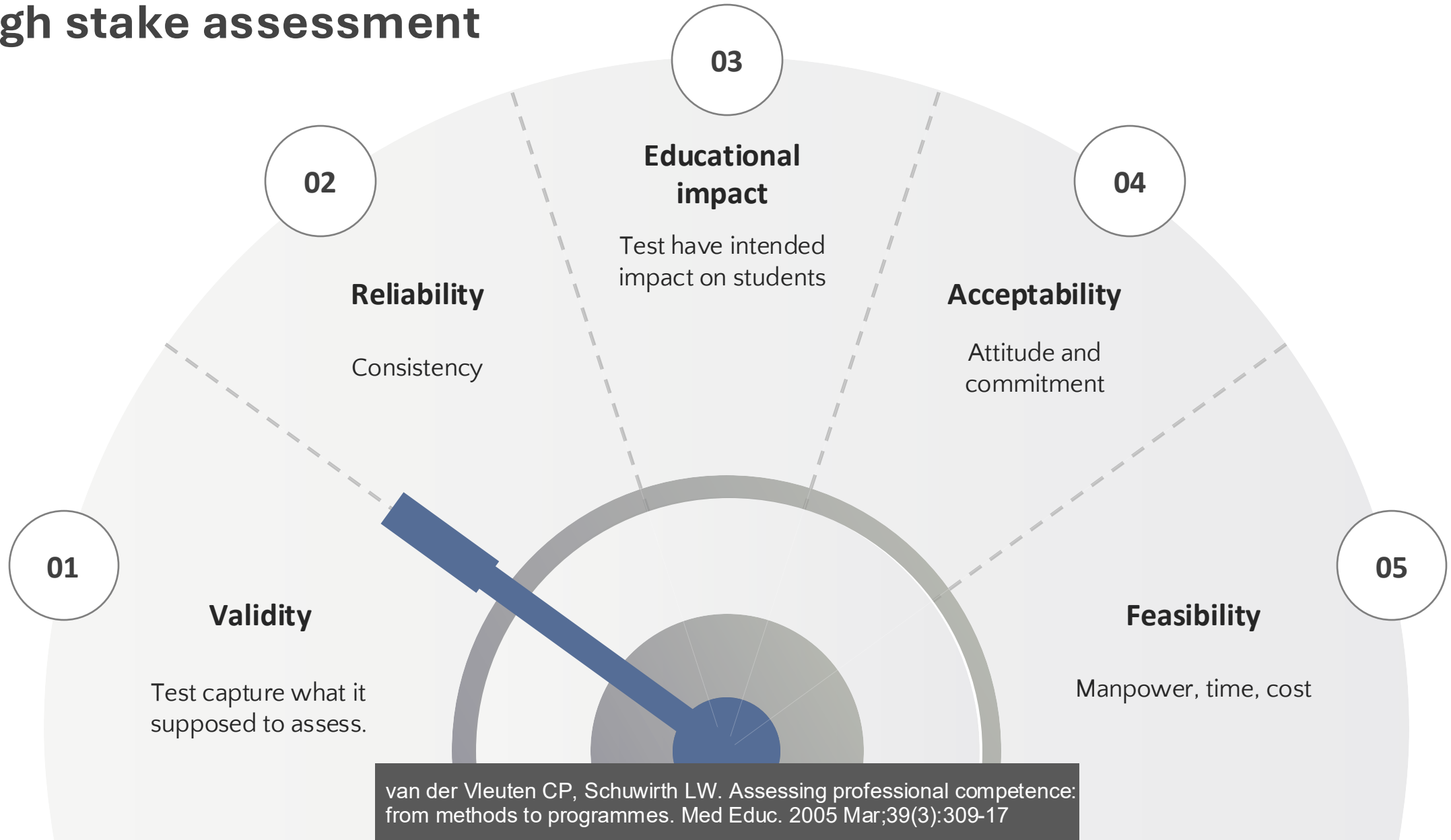


# Desirable criteria of assessment

*But it's a trade-off!*



# Desirable criteria of assessment: High stake assessment



# PRINCIPLES OF CONSTRUCTING THEORY QUESTIONS

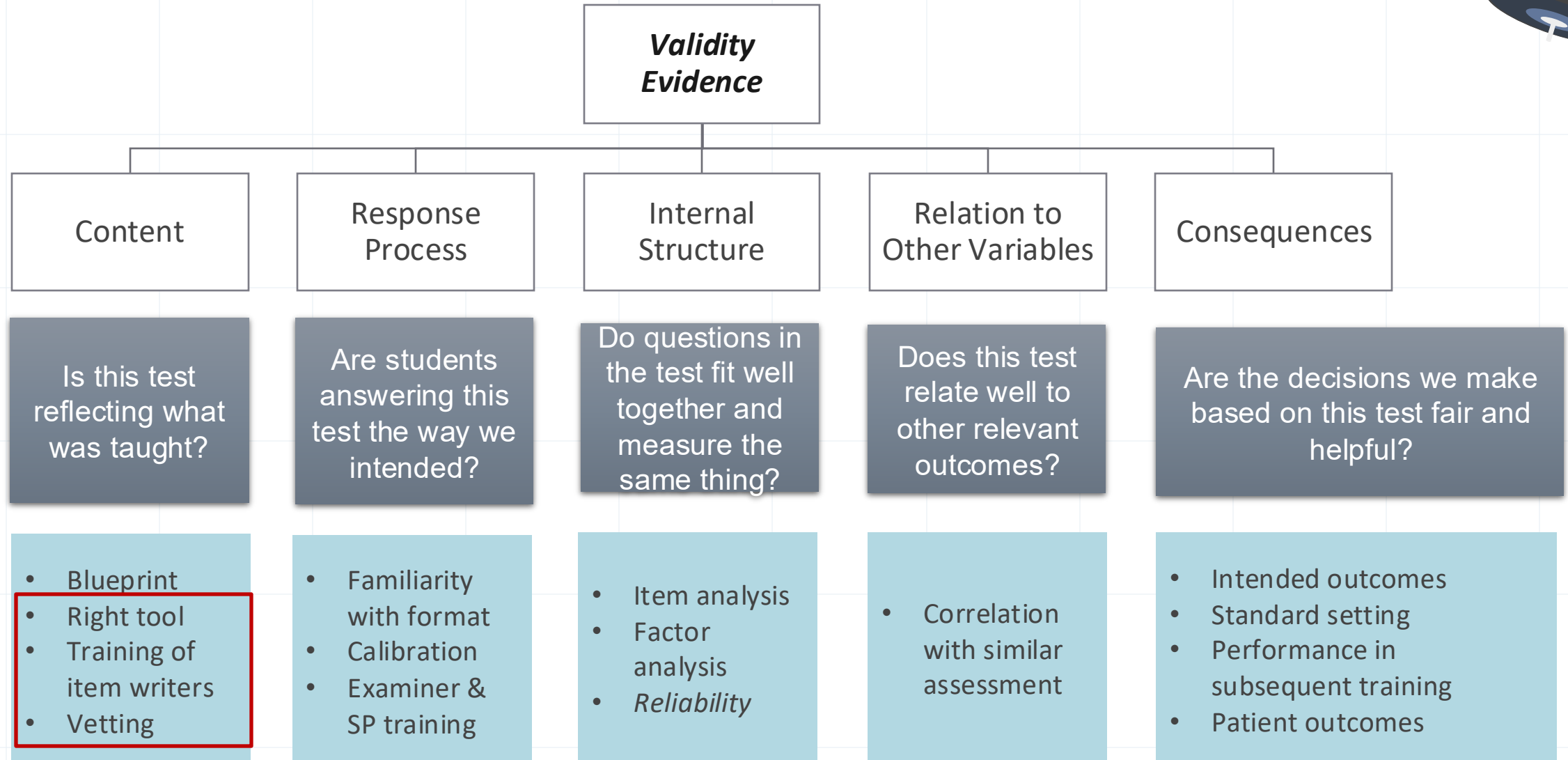
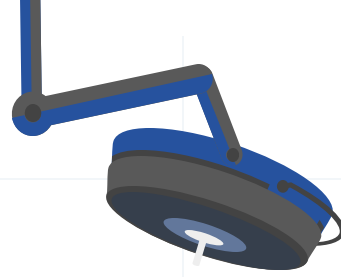
1000-1300

At the end of the session, participants will be able to

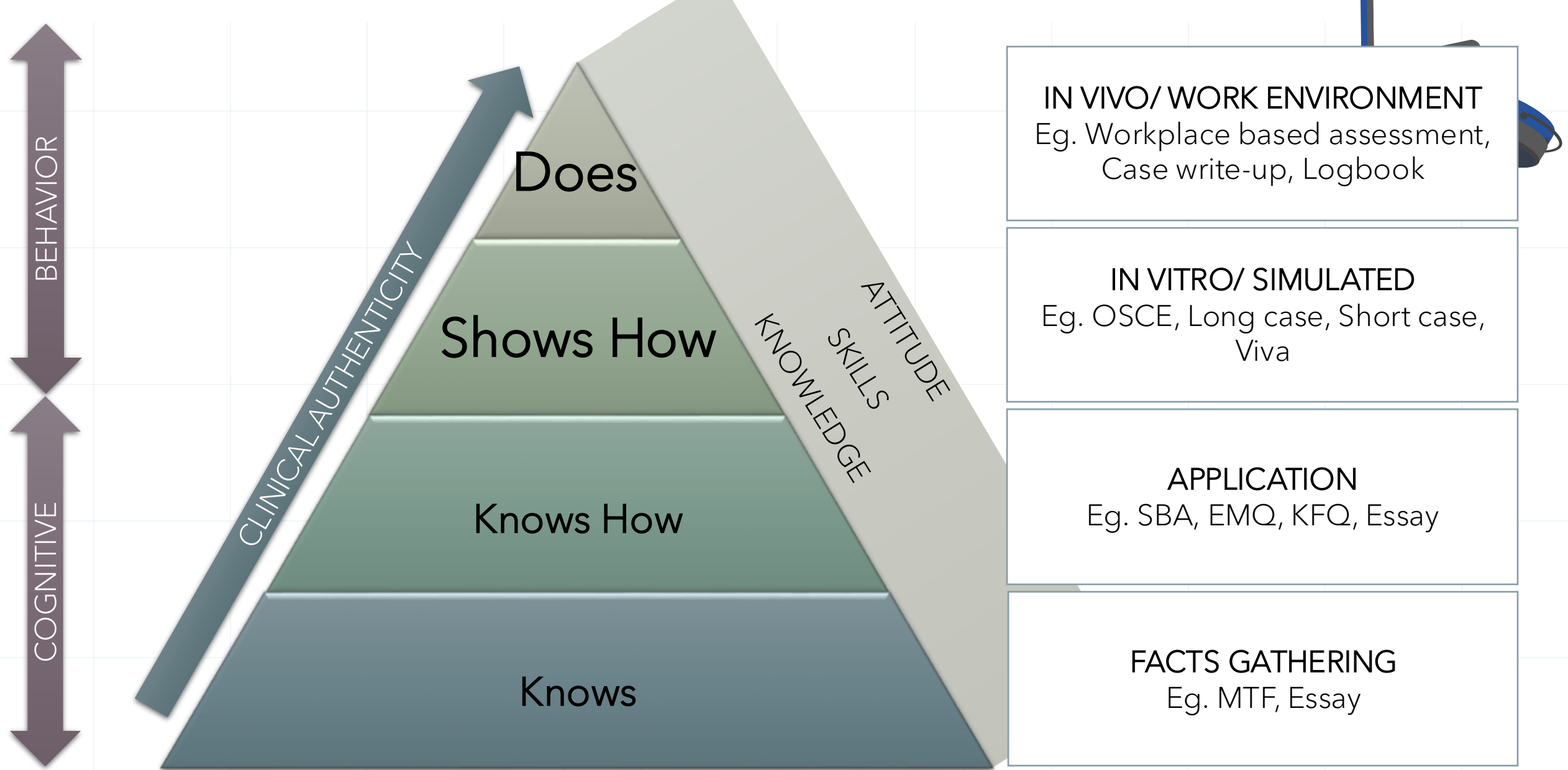
construct MTF, OBA, EMQ, SBQ and KFQ based on best practice guidelines.



# Proposing validity in assessment



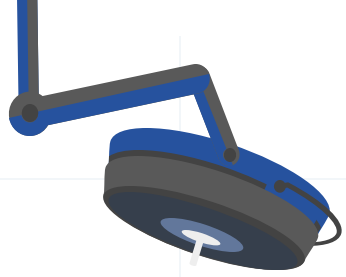




Miller's Pyramid of Clinical Competence (Miller, 1990)

# Group work (1010-1120)

<https://tinyurl.com/2025iium>



1. Check your group (MTF, OBA, EMQ, SBQ, KFAQ)

2. Go through the IIUM draft guideline and resources in Google Drive. You may also use other resources.

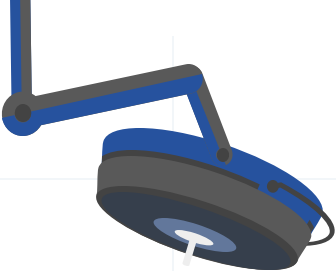
3. Prepare a 10-min presentation

- structure of the format with GOOD example
- best practice (including which level of taxonomy)
- common flaws.

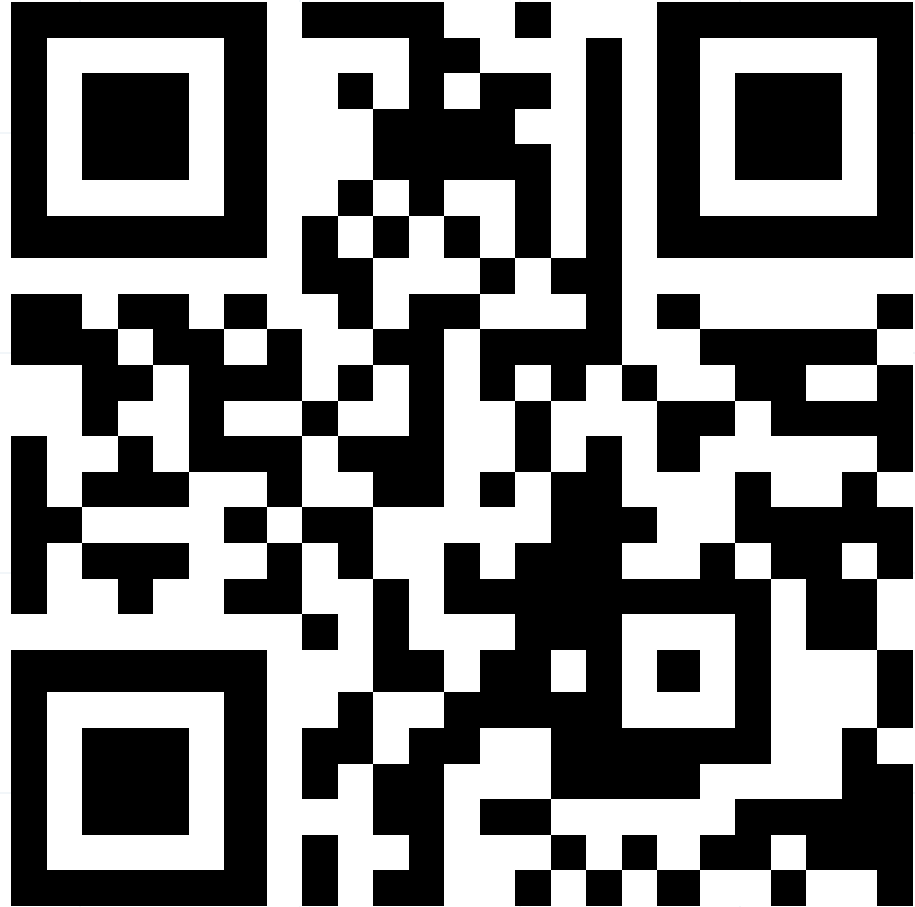
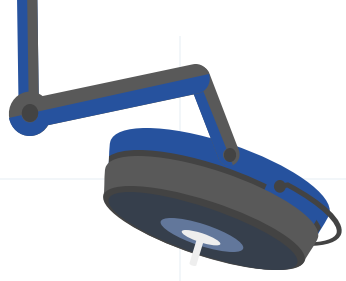
Include one FLAWED example and let the other groups vet the questions.

4. Presentation will begin at 1120

# Selecting the right tool



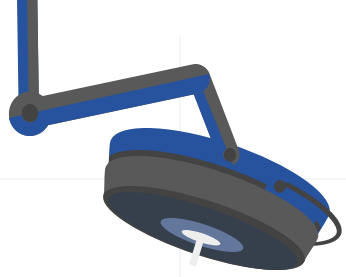
Cognitive Level	Taxonomy	Verbs Used	Examples of Outcome Measured	Suitable Assessment Tools
LOTS	Remember	Define, List	Facts	
	Understand	Explain, Describe, Review	Concept, Problem identification	
HOTS	Apply	Interpret, Apply, Organize	Investigation	
	Analyse	Distinguish, Analyze, Compare & Contrast	Differential diagnosis	
	Evaluate	Evaluate, Choose	Comparing options, Evaluating management, Prognosis	
	Create	Plan, Design	Solving a problem, Management	



<https://tinyurl.com/2025iium>



# Multiple True False (MTF)



- Known as Type X (Simple true false)
- Assesses C1 (recall) and C2 (understand)
- Options are correct OR incorrect

Regarding anatomy of the heart:

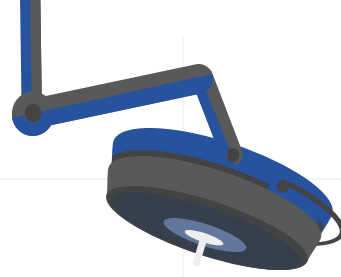
STEM

OPTIONS

True (T) / False (F)

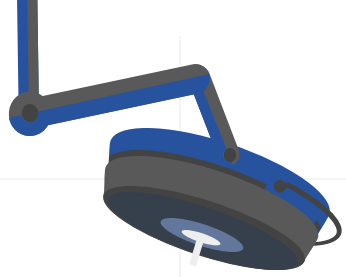
- A. The left ventricle has thicker walls than the right ventricle.
- B. The tricuspid valve is located between the left atrium and left ventricle.
- C. The pulmonary artery carries deoxygenated blood.
- D. The sinoatrial (SA) node is located in the right atrium.

# Multiple True False (MTF)



Strengths	<ul style="list-style-type: none"><li>• Easy to construct</li><li>• High sampling</li><li>• Efficient to score</li></ul>
Common flaws	<ul style="list-style-type: none"><li>• Negative phrased lead in (acceptable in options)</li><li>• <b>Testing two facts in one option</b></li><li>• Grammatical cues</li><li>• Dependent options (eg. knowing A can make student guess B)</li><li>• Using vague terms (eg. common, can, possibly, usually) - normally TRUE</li></ul>

# Identify areas to correct

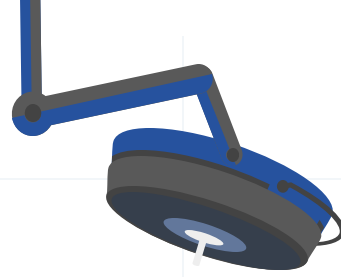


**Which of the following statements about cardiac physiology is incorrect?**

True (T) / False (F)

- A. The sinoatrial node controls heart rate and is located in the left atrium.
- B. If the sinoatrial node functions normally and oxygen demand decreases during exercise, cardiac output decreases regardless of venous return.
- C. The heart usually increases its oxygen demand during exercise.
- D. The pulmonary veins always carry deoxygenated blood to the heart.
- E. The Frank-Starling mechanism is a commonly accepted physiological explanation used in various settings to describe how venous return affects stroke volume.

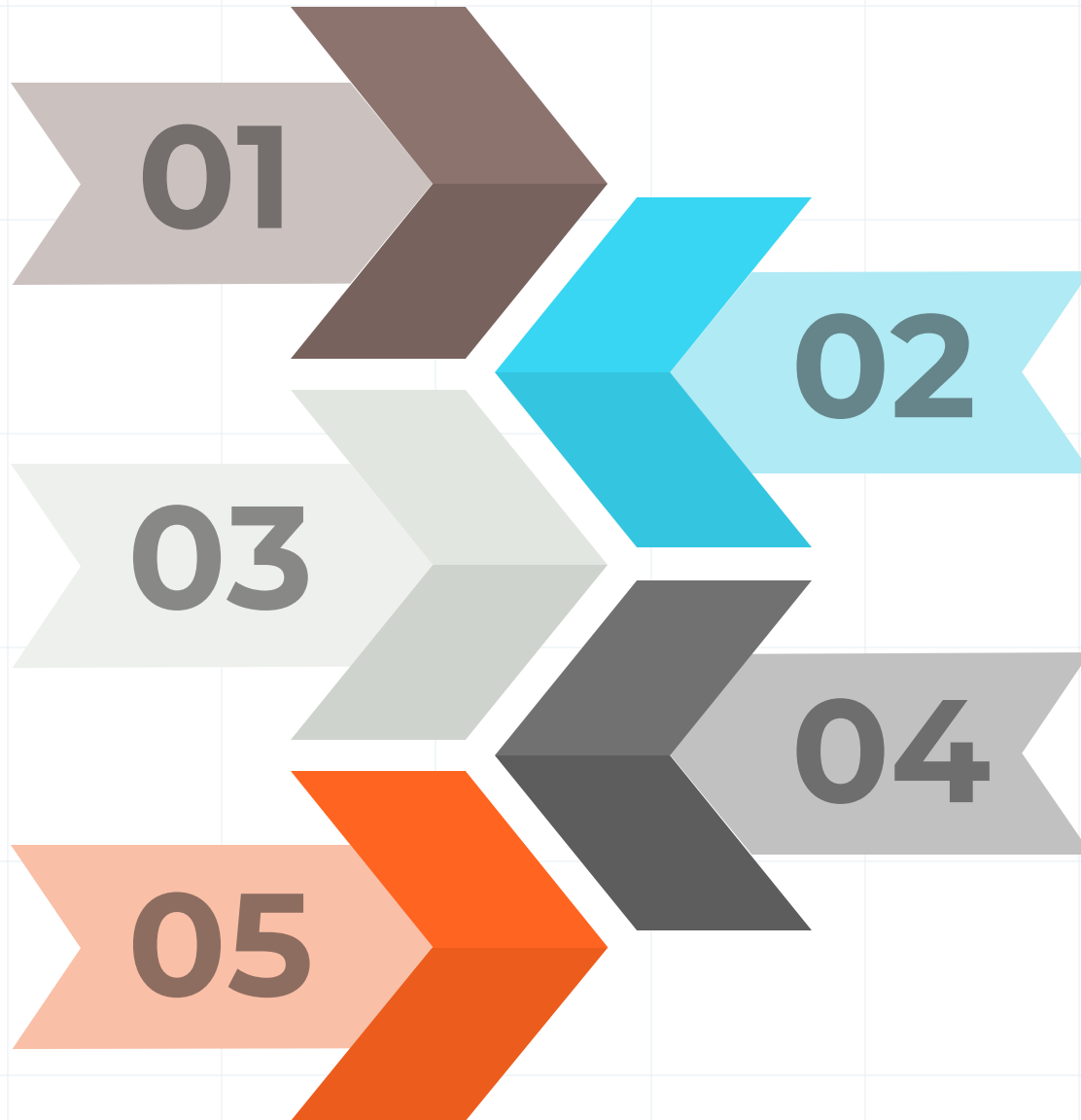
# Is negative marking the solution of guessing?



Random guessing on well-written questions is **overestimated**

The **probability of passing** from random guessing alone is **extremely low**

Best way to control guessing -  
**Good item construction**



Need to differentiate with **informed elimination** of wrong answers with partial knowledge.

Negative marking **do not** solve guessing issues – it changes who still guess (risk taking behavior)



*“Thus, the best way to control blind guessing is to write effective test questions, not to attempt to manipulate the examinee’s psyche or to transform test scores post hoc by using formula scoring”*

Most educational measurement specialists recommend ‘number- correct’ scoring

Standard setting can also incorporate guessing possibility.

Holt A. An analysis of negative marking in multiple-choice assessment.  
Available at: [www.citrenz.ac.nz/conferences/2006/papers/115.pdf](http://www.citrenz.ac.nz/conferences/2006/papers/115.pdf)

Foley, B. P., (2016) “Getting Lucky: How Guessing Threatens the Validity of Performance Classifications”, *Practical Assessment, Research, and Evaluation* 21(1): 3

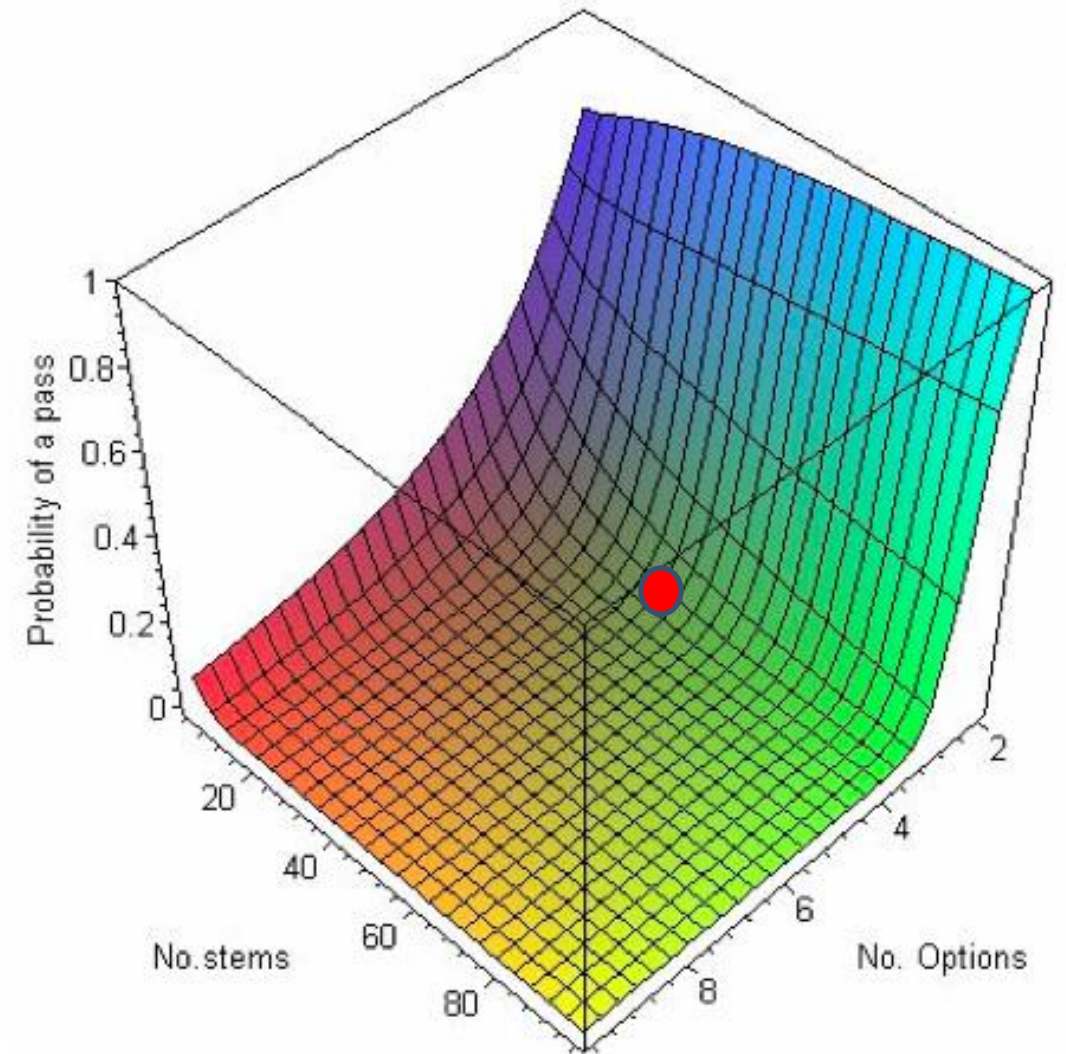
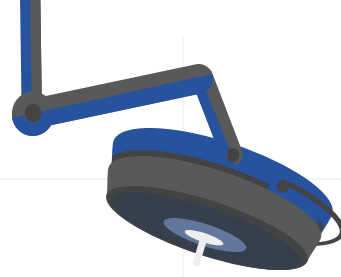


Figure 1 The probability of passing using a pure guessing strategy

# One Best Answer (OBA)



- Known as Type A
- Assesses C3 (apply), C4 (analyse) or C5 (evaluate)
- Options number range 3 to 7. But 3 is the most optimal.
- Incorrect options can be partially or wholly incorrect – but plausible to the weak students

## VIGNETTE

A 30-year-old, gravida 2 para 1 at 37 weeks gestation with one previous Caesarean section 18 months ago currently in active labour. Four hours later, she complains of persistent abdominal pain associated with shortness of breath. Her blood pressure is 90/60 mmHg and pulse rate of 108 bpm. Abdomen examination shows scar tenderness and obvious fetal part is palpable. Cardiotocography shows fetal bradycardia.

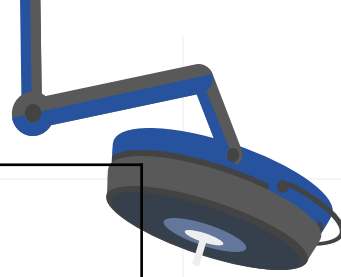
Which is the most likely diagnosis?

## LEAD-IN

## OPTIONS

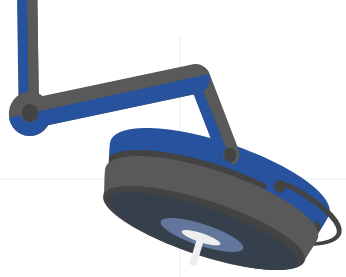
- |                            |                                |
|----------------------------|--------------------------------|
| A. Abruptio placenta       | <i>Distractor</i>              |
| B. Amniotic fluid embolism | <i>Distractor</i>              |
| C. Pulmonary embolism      | <i>Distractor</i>              |
| D. Uterine rupture         | <i>The most correct answer</i> |

# One Best Answer (OBA)



Strengths	<ul style="list-style-type: none"><li>• Test application of knowledge</li><li>• Can accommodate contextual and non-contextual items (without vignette)</li><li>• Relatively higher sampling than essays</li><li>• Efficient to score</li></ul>
Common flaws	<p>Vignette</p> <ul style="list-style-type: none"><li>• Not clear. Good OBA can be answered without looking at options</li><li>• Too lengthy and required more than given time</li><li>• Contains unnecessary information (<i>window dressing</i>)</li></ul> <p>Lead-in</p> <ul style="list-style-type: none"><li>• Stand alone – can be answered even without reading the vignette</li></ul> <p>Options</p> <ul style="list-style-type: none"><li>• Not homogenous</li><li>• Non-functioning distractor (answer stands out)</li></ul>

# Identify areas to correct



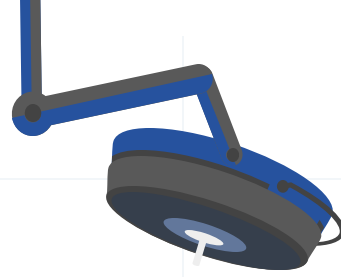
A 25-year-old male medical student presents to your clinic complaining of fatigue for the past two weeks. He has been studying late for exams and admits to skipping meals. He also reports having vivid dreams, occasional palpitations, and a few headaches in the past week. He smokes socially and drinks coffee 3–4 times daily.

On examination, his blood pressure is 128/82 mmHg, pulse 88 bpm, and BMI is 22. Cardiovascular and respiratory exams are unremarkable. He recently adopted a stray cat and spends most of his time indoors studying.

What is the most common complication associated with caffeine use?

- A. Gastrointestinal bleeding
- B. Tachycardia
- C. Anxiety
- D. Peptic ulcer disease
- E. Bradycardia

# Identify areas to correct



A 72-year-old man with a past medical history of ischemic heart disease (status post stent placement 5 years ago), hypertension, atrial fibrillation (on apixaban), and stage 4 chronic kidney disease (eGFR 22 mL/min/1.73m<sup>2</sup>) presents with increasing dyspnea, orthopnea, and reduced exercise tolerance for the past month. He also notes early satiety and mild abdominal distension. He denies chest pain or palpitations.

On examination, his blood pressure is 138/85 mmHg, heart rate is 82 bpm and irregularly irregular, and respiratory rate is 22/min. Oxygen saturation is 95% on room air. He has elevated JVP, bilateral basal crepitations, a displaced apex beat, and 2+ pedal edema. There is mild hepatomegaly on abdominal exam.

Blood tests reveal hemoglobin of 10.2 g/dL, sodium 134 mmol/L, potassium 5.2 mmol/L, BUN 28 mmol/L, creatinine 280 µmol/L, and NT-proBNP of 3000 pg/mL. Echocardiogram shows global hypokinesia, LVEF 30%, and mild tricuspid regurgitation. There is no pericardial effusion.

He is currently on bisoprolol 2.5 mg daily, furosemide 40 mg twice daily, and amlodipine 5 mg daily.

What is the most appropriate next step in optimizing this patient's long-term heart failure management?

- A. Increase bisoprolol to target dose and continue diuretic therapy
- B. Initiate an angiotensin receptor–neprilysin inhibitor (ARNI) after switching off amlodipine
- C. Add mineralocorticoid receptor antagonist (MRA) while monitoring potassium
- D. Refer for evaluation for cardiac resynchronization therapy (CRT)
- E. Stop beta-blocker and start digoxin due to advanced CKD and atrial fibrillation



# Extended Matching Question (EMQ)



**Theme: Choice of Antihypertensive Agents in Special Clinical Situations**

**Lead-in:**

For each of the following stem, select the **most appropriate antihypertensive agent** from the list below. Each option may be used once, more than once, or not at all.

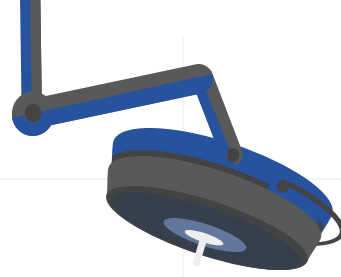
**Options:**

- A. Amlodipine
- B. Lisinopril
- C. Methyldopa
- D. Prazosin
- E. Atenolol
- F. Clonidine
- G. Losartan
- H. Hydrochlorothiazide
- I. Nifedipine

**Stems:**

1. A 30-year-old woman in her second trimester of pregnancy presents for routine antenatal care. Her blood pressure has been consistently elevated over the past 2 weeks, now reading 150/95 mmHg. She has no proteinuria or signs of pre-eclampsia. Her obstetrician plans to initiate antihypertensive treatment. She is otherwise healthy and not on any medications. **Answer: C**
2. A 65-year-old man with a long-standing history of hypertension reports frequent episodes of dizziness when standing up from a seated position. His blood pressure is 140/85 mmHg while sitting and drops to 115/70 mmHg upon standing. He is currently taking lisinopril and furosemide. He also has mild cognitive impairment and a history of non-adherence to medications with complex regimens. **Answer: F**
3. A 70-year-old man with poorly controlled hypertension presents with complaints of increased urinary frequency, especially at night, and a sensation of incomplete bladder emptying. His digital rectal exam shows an enlarged, smooth prostate. His current antihypertensive regimen includes amlodipine, which has been only partially effective. He is keen to avoid polypharmacy if possible. **Answer: D**

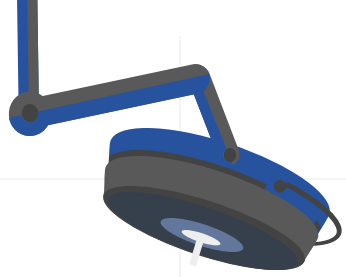
# Extended Matching Question (EMQ)



- Known as Type R
- Assesses C4 (analyse) or C5 (evaluate) – clinical reasoning
- Options number  $\geq 3$  per question. Mimic reality and less prone to scoring.
- Incorrect options can be partially or wholly incorrect – but plausible to the weak students

Strengths	<ul style="list-style-type: none"><li>• Test application of knowledge</li><li>• Relatively higher sampling than essays</li><li>• Reduce chances of guessing (higher number of options than MTF and OBA)</li><li>• Efficient to score</li><li>• Studies shows it is more discriminative than OBA</li></ul>
Common flaws	<ul style="list-style-type: none"><li>• Test trivial topics or nice to know</li><li>• Missing lead-in</li><li>• Not homogenous</li><li>• Too lengthy and required more than given time</li></ul>

# Identify areas to correct



## Theme: Paediatric Presentations in the Emergency Department

### Options:

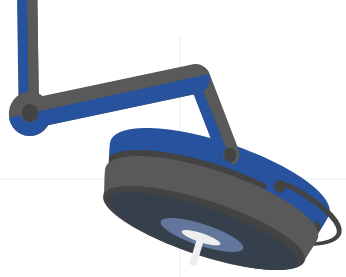
- A. Acute viral bronchiolitis
- B. Pneumonia
- C. Foreign body aspiration
- D. Gastroenteritis
- E. Febrile seizure
- F. Urinary tract infection
- G. Meningitis
- H. Sepsis

### Stems:

1. A 3-month-old infant with a 3-day history of cough, rhinorrhoea, and difficulty feeding. On examination, the baby is tachypnoeic with subcostal recession and diffuse crackles throughout both lung fields. Oxygen saturation is 91% on room air, and the child has had reduced urine output for the past 24 hours. Mother is worried as the baby is her first child and was born prematurely at 33 weeks of gestation.

2. A previously well 2-year-old boy brought in by his nanny who reports that he was playing with coins and suddenly started coughing and gagging. He appears well now but has intermittent wheeze with no fever. The parents are not around and no reliable past medical history is available. There is concern due to a vague history of a similar episode a few weeks ago. Chest X-ray is pending and staff are debating whether to sedate for bronchoscopy immediately.

# Scenario Based Question (SBQ)



- Variant of MEQ – shorter and allow testing of more topics within given time
- Assesses C3 (apply), C4 (analyse) or C5 (evaluate) on clinical scenarios
- CFA on USM data – SBQ loaded on both theory and clinical constructs

## SCENARIO

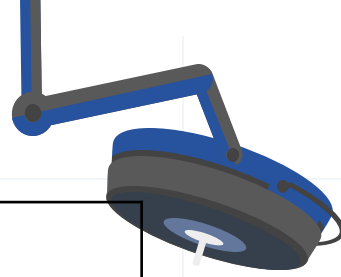
A 28-year-old right-handed woman presented to the Emergency Department with sudden onset of right-sided weakness and slurred speech for the past two hours. She has no prior medical history and is not on any medication. She is a non-smoker and denies any illicit drug use.

On examination, her Glasgow Coma Scale (GCS) is 15. Neurological exam reveals right facial droop, 0/5 power in the right upper and lower limbs, and expressive aphasia. Her blood pressure is 160/90 mmHg and pulse is regular at 88 bpm. An urgent non-contrast CT brain shows no evidence of haemorrhage. CT angiography reveals occlusion of the left middle cerebral artery (MCA).

## QUESTIONS

- a. State the most likely diagnosis. (2 marks)
- b. List TWO (2) modifiable risk factors for the above condition. (2 marks)
- c. State TWO (2) immediate investigations required before initiating definitive treatment. (2 marks)
- d. List FOUR (4) key principles in the acute management of this patient. (4 marks)

# Scenario Based Question (SBQ)



Strengths	<ul style="list-style-type: none"><li>• Assesses knowledge application and clinical reasoning</li><li>• Higher sampling than MEQ</li><li>• Can accommodate clinical images, investigation results</li></ul>
Common flaws	<p>Scenario</p> <ul style="list-style-type: none"><li>• Scenario does not mimic actual clinical presentation – not clinically authentic</li><li>• Too lengthy (window dressing)</li></ul> <p>Questions</p> <ul style="list-style-type: none"><li>• Test recall rather than application</li><li>• Verb does not direct candidates on the level of detail required for answers</li><li>• Accept more answers than allowed</li><li>• Model answer is not enough to reduce intra-rater and inter-rater variability</li></ul>



# SBQ – model answer



a. State the most likely diagnosis. (2 marks)

Acute ischaemic stroke (left MCA territory) – 2m

Accept: “Acute ischaemic stroke

Stroke or CVA – 1m

- A list of correct responses
- Assigned marks for each response
- Clear guidelines on what to include or exclude as correct responses when relevant

b. List TWO (2) modifiable risk factors for the above condition. (2 marks)

Hypertension, Smoking, Diabetes, Hyperlipidaemia (Any 2 = 2m)

Non-modifiable factors – 0 m

c. State TWO (2) immediate investigations required before initiating definitive treatment. (2 marks)

Non-contrast CT brain, Blood glucose, CT angiography, Coagulation profile (Any 2 = 2m)

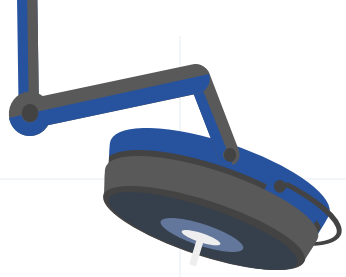
MRI, HbA1c, Lipid profile – 0 m

d. List FOUR (4) key principles in the acute management of this patient. (4 marks)

IV thrombolysis (if within 4.5 hrs), Mechanical thrombectomy, BP control, Swallowing assessment, DVT prophylaxis (Any 4 = 4m)

Long-term prevention (e.g., statins) – 0 m

# Identify areas to correct

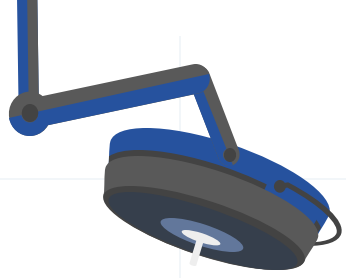


A 35-year-old freelance musician presented to the clinic after a weekend music festival where he had consumed large quantities of fried food and alcohol. He says he started having some “weird” upper belly pain that radiates to the back, which he thought was due to poor sleeping posture in his car. He also reports feeling bloated, hasn’t eaten well since, and passed loose stools. He remembers having something similar last year after a food fair.

On examination, he’s mildly febrile (37.8°C), HR 105 bpm, BP 110/72 mmHg, and mild epigastric tenderness is noted. His friend thinks it’s probably food poisoning.

- a. Mention possible diagnoses. (2 marks)
- b. State two common causes of acute pancreatitis. (2 marks)
- c. What tests would you like to order? (2 marks)
- d. Mention your next steps in management. (4 marks)

# Key Feature Question (KFQ)



- Focus on **challenging or critical** aspect in the diagnosis and management that are **error-prone**
- Assesses C3 (apply), C4 (analyse) or C5 (evaluate) on clinical scenarios
- 2-3 key feature per case maximises test reliability

# Key Feature Question (KFQ)

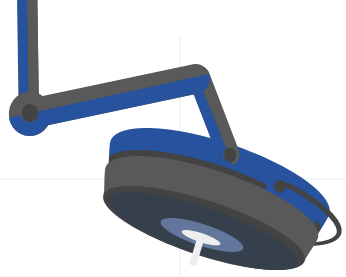
- Assesses C3 (apply), C4 (analyse) or C5 (evaluate)
- Case: Focus on **challenging or critical** aspect in the diagnosis and management that are **error-prone**
- Key feature: 2-3 KF per case maximises test reliability

## *Identifying case and items for KFQ*

Given an adult brought to the emergency room with multiple seizures and without having regained full consciousness, the graduating medical student (new intern) will:

1. Generate a provisional diagnosis of status epilepticus;
2. Begin immediate initial management: secure airways, vitamin complex, bolus of hyper-tonic glucose, and anti-epileptic medication; and
3. Order immediate investigation to identify potentially treatable causes of the seizures: alcohol level, arterial blood gases, brain CT or MRI, serum calcium, serum, and drug screening. Note: Although appropriate, electrolytes and serum glucose were not included in the KF because they are part of routine orders in this situation and not likely to discriminate among interns.

*Source: Adapted, with permission, from an example from the Medical Council of Canada's Guidelines for the Development of Key Feature Problems and Test Cases, 2012*



## Clinical Scenario

An unknown man in his thirties is brought to the emergency room by ambulance because he collapsed on to the sidewalk while waiting for the bus. A witness immediately called an ambulance and reported that before falling to the ground, the man seemed confused, agitated, and arguing with himself. After falling, he began to twitch for a short while, his face became blue, and then he began to have jerky movements all over his body for about a minute. He then partly recovered consciousness but remained confused. During the 12-minute ambulance ride, he presented a similar incident, without recovering full consciousness, and was given lorazepam 2 mg IV by the ambulance personnel who also installed a normal saline IV line.

On arrival in the emergency room, he had a third incident that you witnessed. His vital signs are: Pulse is 74/minute, regular; respiration rate is 16/minute, non-labored; blood pressure is 122/74 mmHg; temperature is 37.8 °C; and an oxygen saturation of 89% on room air. He looks neglected and is unconscious. No relatives or friends accompanied him. His capillary glucose level is 4.6 mmol/l.

### Part 1

**Question 1:** What is (are) your leading working diagnosis(es) at this point in time? You may not list more than two.

1. \_\_\_\_\_
2. \_\_\_\_\_

**Points**

**1**

**Keyed responses**

Status epilepticus. Note: Both elements are required; epilepsy alone is not acceptable.

**0**

Wrote more than two diagnoses.

**Question 2:** What is your immediate management (excluding investigation) at this point in time? Be specific; you may not list more than six.

1. \_\_\_\_\_
2. \_\_\_\_\_
3. \_\_\_\_\_
4. \_\_\_\_\_
5. \_\_\_\_\_
6. \_\_\_\_\_

**Question 2: KF-2:** *Begin immediate initial management: secure airways, vitamin B complex, bolus hypertonic glucose, and anti-epileptic medication.*

Points	Keyed responses
0.25	Intubation, mechanical ventilation, or secure airways. Note: Oxygen alone is <i>not</i> acceptable.
0.25	Vitamin B, B1, or thiamine.
0.25	Glucose, hypertonic, bolus. Note: All three elements are required.
0.25	[lorazepam or diazepam or clonazepam] AND [phenytoin <i>or valproate sodium or levetiracetam</i> ]
0	Listed more than five responses or wrote “none.”



**Question 3:** You have not been able to contact anyone who might know him. What investigation will you order at this point? You may select as many as you feel appropriate. Select option 35 if you do not wish to order any investigation at this time.

1. Alanine aminotransferase (ALT)

2. Alcohol level

3. Aldolase, serum

4. Alkaline phosphatase, serum

5. Amylase, serum

6. Arterial blood gases (ABG)

7. Aspartate aminotransferase (AST)

8. Brain CT-scan

9. Brain MRI

10. Brain PET-scan

11. Calcium, serum

12. Carotid US-doppler

13. Cerebral angiography

14. Cerebro-spinal fluid examination

15. Complete blood count (CBC)
16. C-Reactive protein

17. Creatine phosphokinase, s

18. Creatinine, serum

19. Drug screening, serum

20. Drug screening, urine

21. Echovirus, serology

22. EEG recording

23. Electrolytes (Na, K, Cl)

24. g-Glutamyl transferase

25. Glucose, serum

26. Lactate dehydrogenase, serum

27. Lyme disease, serology

28. Protein electrophoresis, plasma

29. T4, Free

30. Temporal artery biopsy

31. Thyroid-stimulating hormone

32. Total protein, plasma

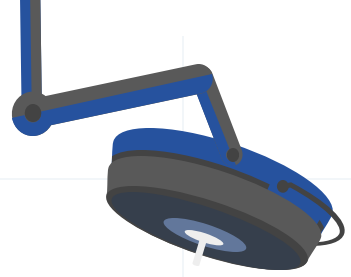
33. Urea, serum

34. VDRL (Venereal Disease Research Laboratory), serum

35. No investigation needed at this point in time

Question 3: KF-3: Order immediate investigation to identify potentially treatable causes of the seizures: Alcohol level, ABG, brain CT or MRI, drug-screening test, and serum calcium.	
Points	Keyed responses
0.20	2. Alcohol level
0.20	6. Arterial blood gases (ABG)
0.20	8. Brain CT or 9. Brain MRI
0.20	11. Calcium, serum
0.20	19. Drug screening, serum or 20. Drug screening, urine
0	Selected more than eight options (i.e., over-investigation) or selected option 35.

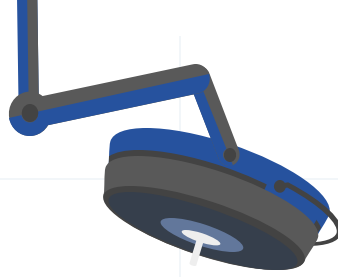
# Key Feature Question (KFQ)



- Assesses C3 (apply), C4 (analyse) or C5 (evaluate)
- Case: Focus on **challenging or critical** aspect in the diagnosis and management that are **error-prone**
- Key feature: 2-3 KF per case maximises test reliability
- Key feature: pick N item, EMQ, fill-in-the-blank, match, MTF
- Can be marked by computer in a good scoring system

Common flaws	<p>Case</p> <ul style="list-style-type: none"><li>• Does not test critical areas that is error-prone</li></ul> <p>Key feature item</p> <ul style="list-style-type: none"><li>• Format does not match what happens in practice (eg. using short menu is more suitable in selecting investigation than suggesting diagnosis)</li><li>• KF does not highlight the critical aspect (eg. What investigation you will order at this time?)</li></ul>
--------------	--

# Identify areas to correct

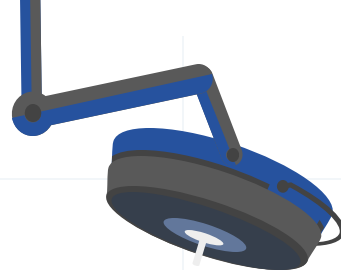


A 63-year-old man with a 20-year history of poorly controlled type 2 diabetes presents to the clinic with a foot wound. He reports that it began as a small blister one week ago and has become increasingly painful. He has no fever but feels "tired." He walks with a limp.

On examination, there is a 4 cm ulcer over the plantar surface of the right foot with surrounding redness. There is minimal discharge. Pedal pulses are difficult to palpate. The wound has a mild odour.

1. Select your diagnosis (*1 mark*)
2. What investigations would you consider for this patient? (Select all that apply) (*2 marks*)
3. What are the management steps? (*2 marks*)

# Composing questions: Considerations



In general, MCQ are preferred for its high reliability (high sampling of topics).

Reserve SBQ and KFQ for the HOTs and topics that fit the purpose of assessment – test clinical reasoning or critical points that are error-prone

*Table 1 Reliability estimates of different assessment instruments as a function of testing time*

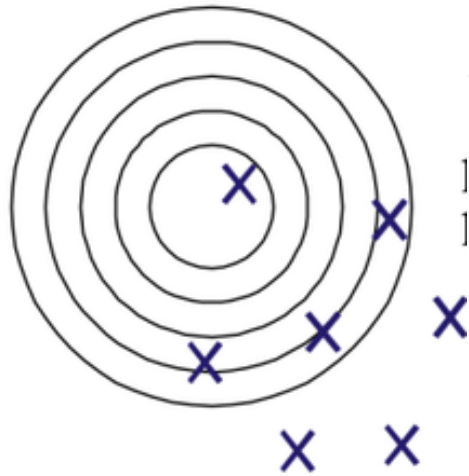
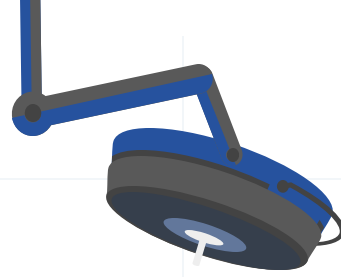
Instrument	Description	Reliability for different testing times			
		1 hour	2 hours	4 hours	8 hours
Multiple choice <sup>*42</sup>	Short stem and short menu of options	0.62	0.76	0.93	0.93
Patient management problem <sup>*42</sup>	Simulation of patient, full scenarios	0.36	0.53	0.69	0.82
Key feature case (write-in) <sup>*43</sup>	Short patient case vignette followed by write-in answer	0.32	0.49	0.66	0.79
Oral examination <sup>†44</sup>	Oral examination based on patient cases	0.50	0.69	0.82	0.90
Long case examination <sup>†45</sup>	Oral examination based on previously unobserved real patient	0.60	0.75	0.86	0.90
OSCE <sup>*46</sup>	Simulated realistic encounters in round robin format	0.54	0.69	0.82	0.90
Mini-clinical exercise (mini-CEX) <sup>‡47</sup>	Short follow-up oral examination based on previously observed real patient	0.73	0.84	0.92	0.96
Practice video assessment <sup>†16</sup>	Selected patient–doctor encounters from video recordings in actual practice	0.62	0.76	0.93	0.93
Incognito standardised patients <sup>‡48</sup>	Real consultations scored by undetected simulated patients	0.61	0.76	0.82	0.86

\* One-facet all random design with items crossed with persons (pxi).

† Two-facet all random design with judges (examiners) nested within items within persons (j:i:p).

‡ One-facet all random design with items nested within persons (i:p).

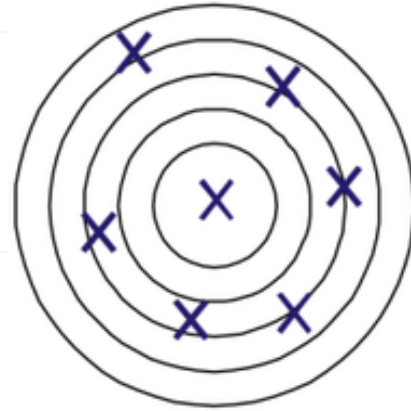
# Validity and reliability: Essay



LOTs  
Incomplete  
model  
answer

Less valid

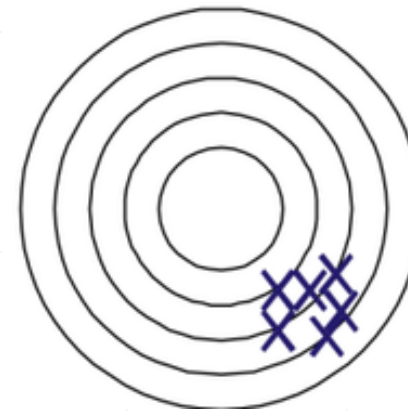
Less



HOTs  
Incomplete  
model  
answer

More valid

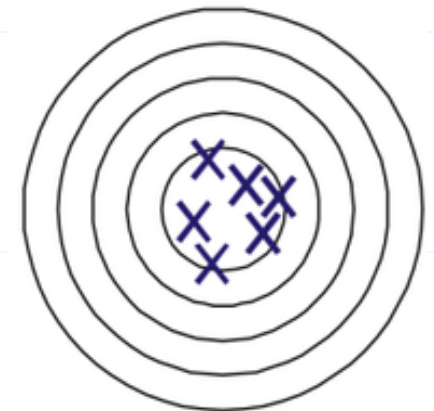
Less reliable



LOTs  
Complete  
model  
answer

Less valid

More

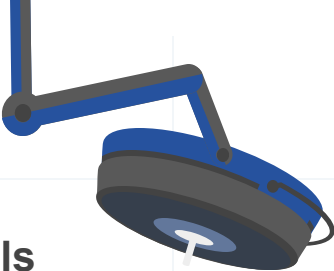


HOTs  
Complete  
model  
answer

More valid

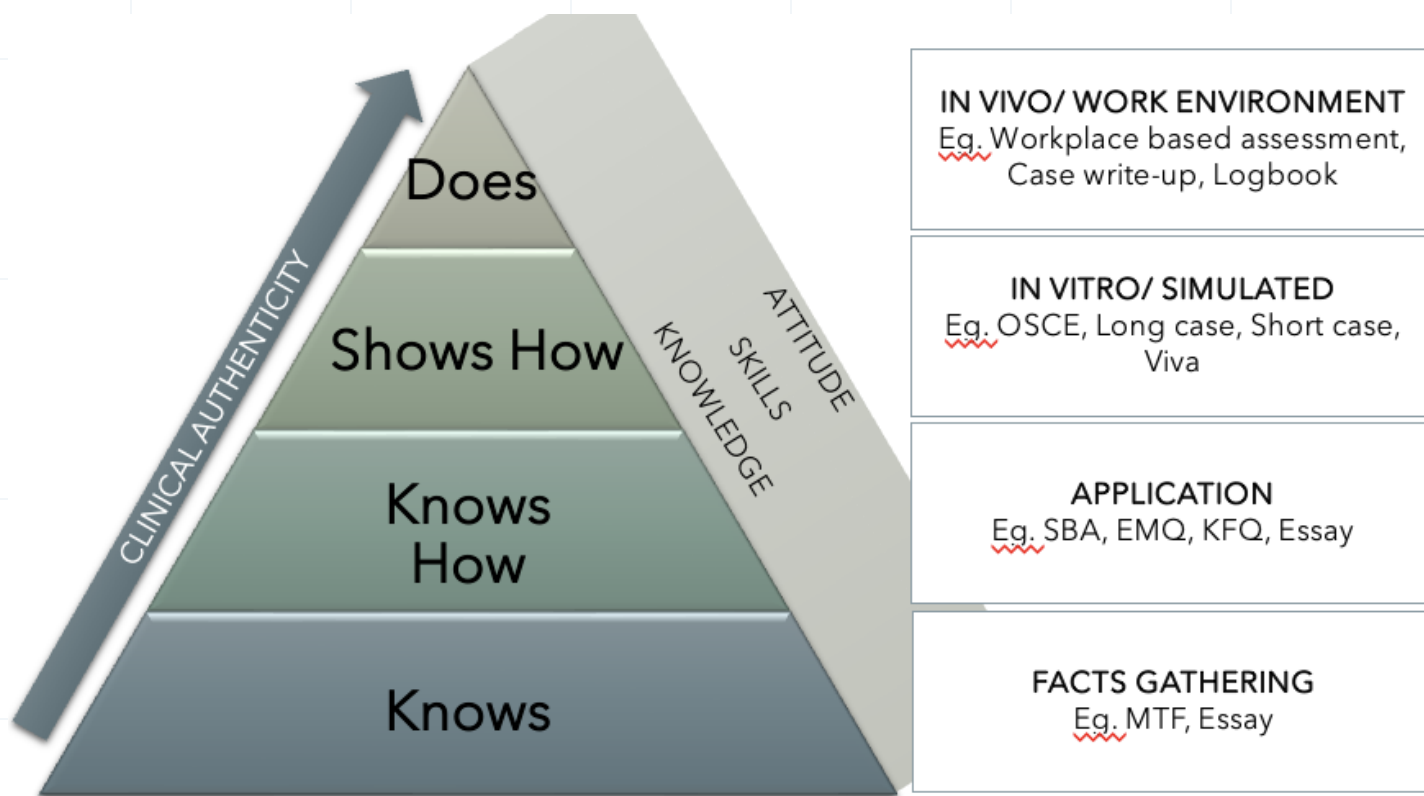
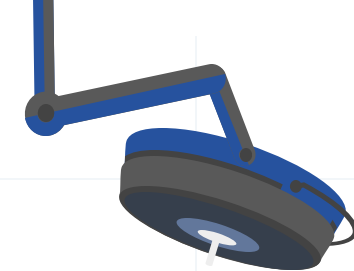
More

# Selecting the right tool



Cognitive Level	Taxonomy	Verbs Used	Examples of Outcome Measured	Suitable Assessment Tools
LOTS	Remember	Define, List	Facts	• MTF
	Understand	Explain, Describe, Review	Concept, Problem identification	• MTF
HOTS	Apply	Interpret, Apply, Organize	Investigation	
	Analyse	Distinguish, Analyze, Compare & Contrast	Differential diagnosis	• OBA • EMQ • SBQ • KFQ
	Evaluate	Evaluate, Choose	Comparing options, Evaluating management, Prognosis	
	Create	Plan, Design	Solving a problem, Management	





- No single method can do it all
- Competence is specific, not generic → blueprint is important to cover important areas
- All methods of assessment can have ‘built-in’ validity → writing guidelines, training, vetting

# ASSESSMENT BLUEPRINT

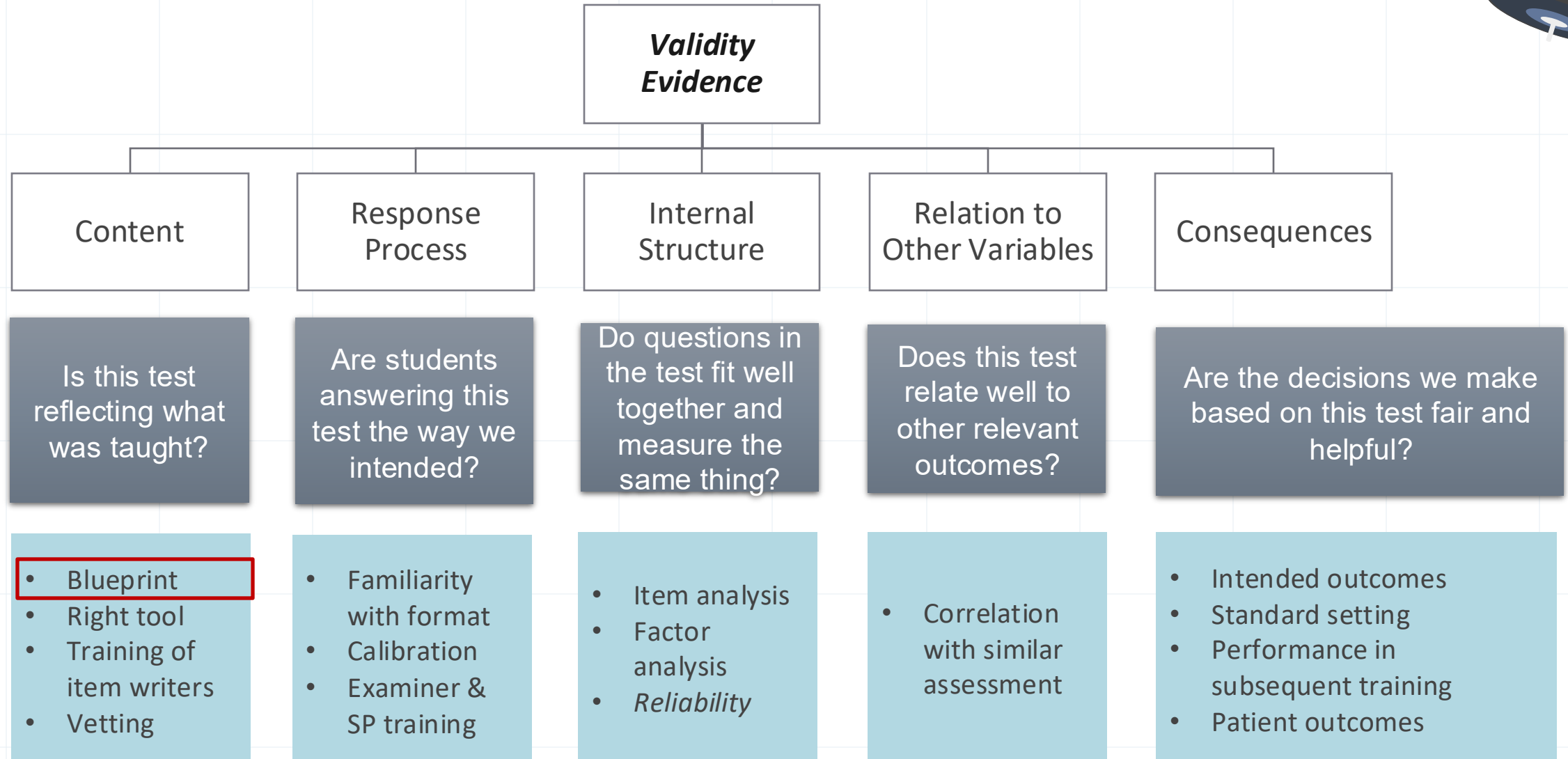
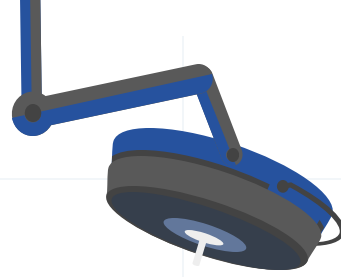
1415-1500

At the end of the session, participants will be able to

construct assessment blueprint based on best practice guidelines.



# Proposing validity in assessment



# Working example: Medical Professionalism (Year 1)

(USM MD Phase 1 Guidebook, 2020)

2 Credit Hours

BIL.	HASIL PEMBELAJARAN KURSUS	PO	LT	SS	KAEDAH PENILAIAN
1	Memahami asas profesionalisme perubatan	PLO1	C2	-	SBA, SBQ, Reflective Essay
2	Mengaplikasi asas profesionalisme perubatan dalam kes klinikal	PLO2	C3	CTPS	SBA, SBQ, PBL, Reflective Essay
3	Mempamer asas profesionalisme perubatan dari sudut komunikasi secara lisan dan bukan lisan	PLO	A3	CS	Group presentation, OSCE
4	Menjelas kes klinikal dari sudut etika dan profesionalisme	PO6	A3	EM	SBA, SBQ, PBL, Reflective Essay

# Working example: Medical Professionalism (Year 1)

(USM MD Phase 1 Guidebook, 2020)

BI L.	TAJUK/SUB TAJUK	Mod Pengajaran	Memahami asas profesionalisme perubatan	Mengaplikasi asas profesionalisme perubatan dalam kes klinikal	Mempamer asas profesionalisme perubatan dari sudut komunikasi secara lisan dan bukan lisan	Menjelas kes klinikal dari sudut etika dan profesionalisme
1.	Introduction to Medical Professionalism	Lecture	/			
2.	Communication Skills in Medicine	Tutorial	/	/	/	
3.	Introduction to Clinical Skills (History Taking, Physical Examination, Vital Signs)	Tutorial & Practical		/	/	
4.	Patient Autonomy	Tutorial & PBL	/	/	/	/
5.	Patient Confidentiality	Tutorial	/	/	/	/
6.	Patient Rights	Tutorial & PBL	/	/	/	/
7.	Interprofessionalism in Medicine	Tutorial & Hospital visit	/			/
8.	Introduction to Medical Negligence	Lecture	/			
9.	Equity & Social Justice	Lecture	/			/
10.	Academic Integrity	Workshop	/		/	
11.	Professional Resilience	Workshop	/			
12.	Reflection Skills	Workshop			/	/

Is this a blueprint?

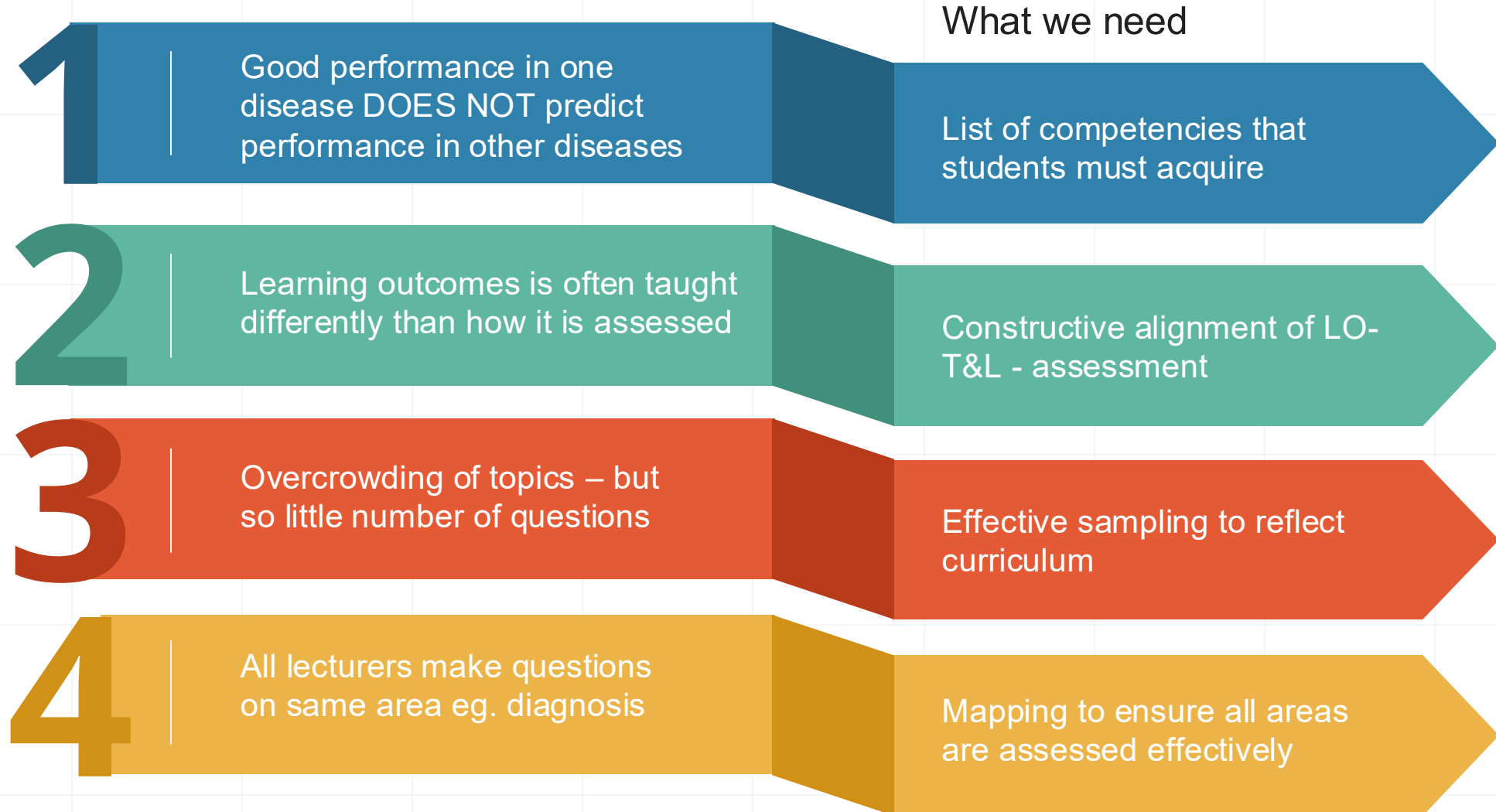
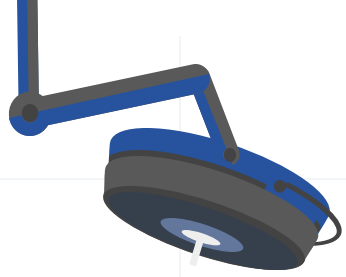
Do not inform course coordinators on

- weightage
- assessment plan

Do not guide students on learning

(McLaughlin, Lemaire & Coderre, 2005)

# Why blueprint is a must?





## 2.2. Assessment Methods

- 2.2.1. The medical school must ensure:
- that there are a variety of methods and tools that are appropriate for the assessment of learning outcomes and competencies.
  - it assesses medical students against the learning outcomes at appropriate points, and make sure they achieve all outcomes upon graduation.
  - that students who graduate have demonstrated that they are competent in all the outcomes.
  - that the assessments are open to scrutiny by external expertise.

*Annotation: A variety of methods and tools: Medical school must use a valid and reliable assessment tool to assess different learning domains. It is best shown by assessment **blueprint**.*



**quality indicators**

# No specific template

(McLaughlin, Lemaire & Coderre, 2005)

## One format in one course

2009; 31: 322-324



### TWELVE TIPS

## Twelve tips for blueprinting

SYLVAIN CODERRE, WAYNE WOLOSCHUK & KEVIN MCLAUGHLIN  
Office of Undergraduate Medical Education, University of Calgary, Canada

## Many formats in one course

### EDUCATIONAL RESOURCE

Volume 12 Issue 1 2020

DOI: 10.21315/eimj2020.12.1.8

#### ARTICLE INFO

Submitted: 26-12-2019

Accepted: 31-01-2020

Online: 10-04-2020

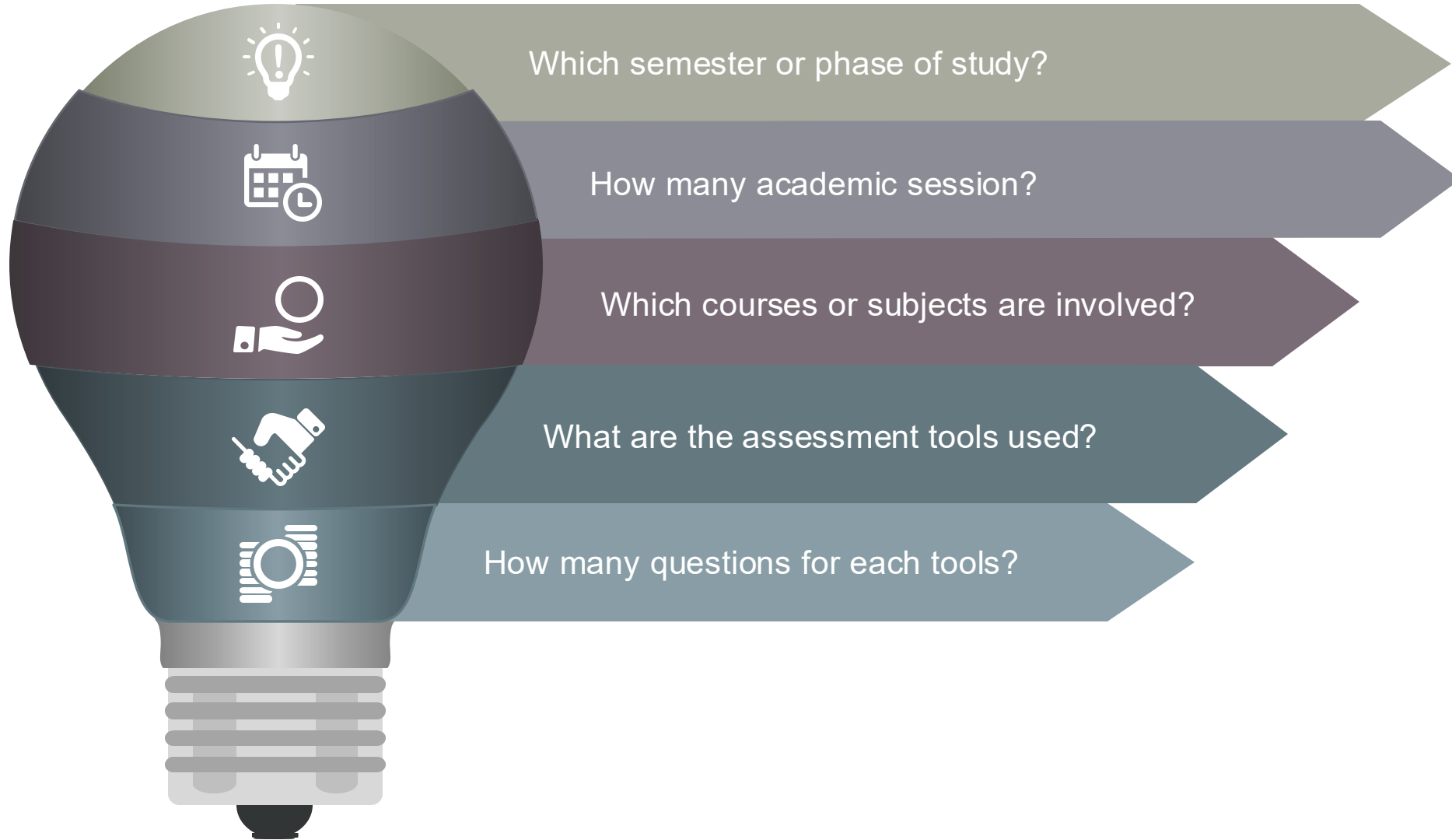
## Seven Steps to Construct an Assessment Blueprint: A Practical Guide

Muhd Al-Aarifin Ismail, Mohamad Najib Mat Pa, Jamilah Al-  
Muhammady Mohammad, Muhamad Saiful Bahri Yusoff

*Department of Medical Education, School of Medical Sciences,  
Universiti Sains Malaysia, Kelantan, MALAYSIA*

# STEP 1: Define blueprinting purpose and scope

(Ismail, Mat Pa, Mohammad, Yusoff, 2020)



# STEP 2: Tabulate curricular content

(Ismail, Mat Pa, Mohammad, Yusoff, 2020)

## Topics

Course
MICROBIOLOGY
Classification of medically important bacteria
Classification and development of medically important fungi
Classification and replication of viruses
Microbial genetics and principles of antimicrobial resistance mechanisms
Transport media, growth and microbial culture
Classification of medically important parasites
Microscopic examination of bacteria and fungi
Host-pathogen interactions
Basic medical microbiology
IMMUNOLOGY
Cells and organs of the immune system
T & B cell development
Complement and cytokines
Immune response and regulation
Principle of antigen and antibody interactions

## Diseases

Care of Chronic disease / Therapeutic
Diabetes
Hypertension
Thyroid disease (Hyperthyroidism/hypothyroidism)
Cardiovascular accident/stroke/ Dyslipidaemia
Ischemic heart disease (L)
COPD
Asthma
Chronic kidney disease
Anaemia
Gouty Arthritis
Pulmonary Tuberculosis

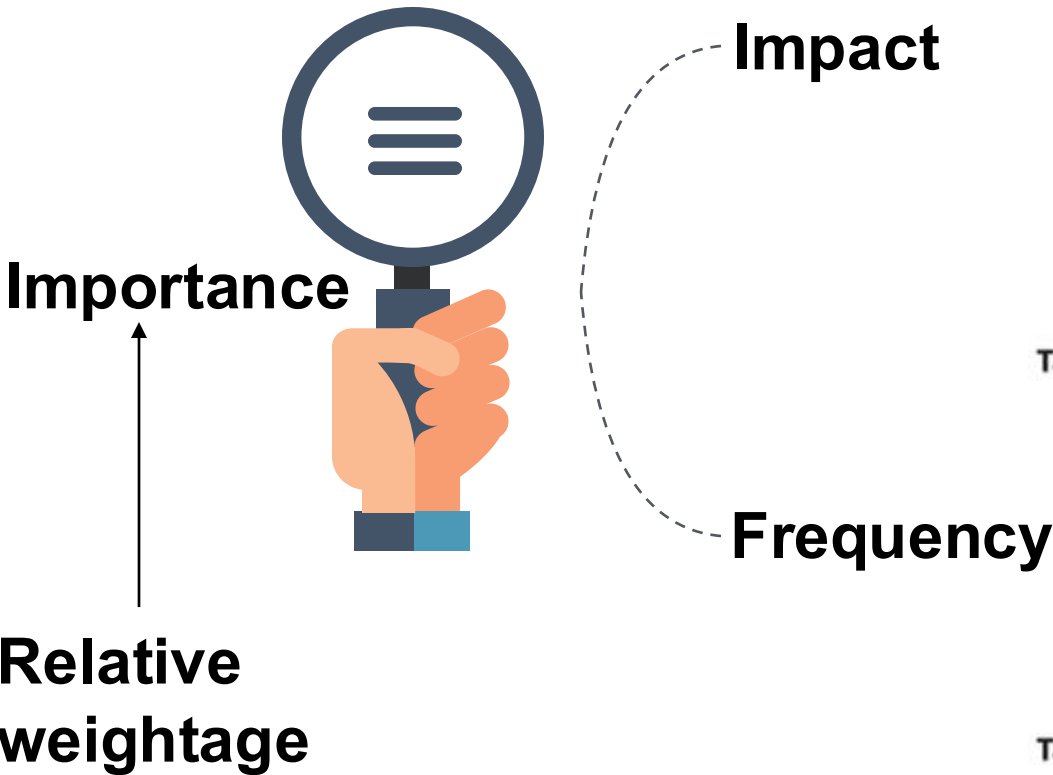
## Clinical presentation

### Presentation

Hypematremia  
Hyponatremia  
Hyperkalemia  
Hypokalemia  
Acidosis  
Alkalosis  
ARF  
CRF  
Hematuria  
Proteinuria  
Edema  
Scrotal mass  
Urinary retention  
Hypertension  
Polyuria  
Renal colic  
Dysuria  
Incontinence  
TOTAL

# STEP 3: Identify IF

(Ismail, Mat Pa, Mohammad, Yusoff, 2020)



**Table 4:** Weighting criteria for impact and frequency of the clinical presentations for undergraduate renal course at the University of Calgary

Impact		Frequency	
1	Non-urgent, little prevention potential	1	Rarely seen
2	Serious, but not immediately life threatening	2	Relatively common
3	Life threatening emergency and/or high potential for prevention impact	3	Very common

**Table 2:** Weighting criteria for impact and frequency of the curricular contents for undergraduate Phase 1 at the School of Medical Sciences

Impact		Frequency	
1	Less important for Phase 2	1	Rarely applied in Phase 2
2	Important for Phase 2	2	Commonly applied in Phase 2
3	Very important for Phase 2	3	Frequently applied in Phase 2

**Table 3:** Weighting criteria for impact and frequency of the curricular contents for undergraduate Phase 2 at the School of Medical Sciences

Impact		Frequency	
1	Less important for house officers	1	Rarely applied in clinical practices
2	Important for house officers	2	Commonly applied in clinical practices
3	Very important for house officers	3	Frequently applied in clinical practices

# STEP 4: Categorize curricular content based on weightage

(Ismail, Mat Pa, Mohammad, Yusoff, 2020)

**Table 2:** Weighting criteria for impact and frequency of the curricular contents for undergraduate Phase 1 at the School of Medical Sciences

Impact		Frequency	
1	Less important for Phase 2	1	Rarely applied in Phase 2
2	Important for Phase 2	2	Commonly applied in Phase 2
3	Very important for Phase 2	3	Frequently applied in Phase 2

NO	TOPICS	TEACHING MODE	IMPACT	FREQUENCY
1.	Introduction to Medical Professionalism	Lecture	3	3
2.	Communication Skills in Medicine	Tutorial	3	3
3.	Introduction to Clinical Skills (History Taking, Physical Examination, Vital Signs)	Tutorial & Practical	3	3
4.	Patient Autonomy	Tutorial & PBL	3	2
5.	Patient Confidentiality	Tutorial	3	3
6.	Patient Rights	Tutorial & PBL	3	2
7.	Interprofessionalism in Medicine	Tutorial & Hospital visit	2	2
8.	Introduction to Medical Negligence	Lecture	1	1
9.	Equity & Social Justice	Lecture	2	2
10.	Academic Integrity	Workshop	3	3
11.	Professional Resilience	Workshop	3	2
12.	Reflection Skills	Workshop	2	3




# STEP 5: Decide on % of questions for each category

(Ismail, Mat Pa, Mohammad, Yusoff, 2020)

**USM’s approach: Suitable if you use various assessment tools.** (Ismail, Mat Pa, Mohammad, Yusoff, 2020)

Get a consensus on curriculum profile

Eg. USM MD  
Must know 60%  
Should know 30%  
Nice to know 10%



**Table 5:** Classification of curricular content based on their weightage

I x F	Category	Percentage	SEMESTER I			PRO. I		SUPP. PRO. I	
			MTF	SEQ	OSCE	MTF	SEQ	MTF	SEQ
6–9	Must know	60 %	13	3	2	6	1	6	1
3–4	Should know	30 %	6	2	1	1	0	1	0
1–2	Nice to know	10 %	1	0	0	0	0	0	0
Total		100 %	20	5	3	7	1	7	1

**Table 3:** Weighting criteria for impact and frequency of the curricular contents for undergraduate Phase 2 at the School of Medical Sciences

Impact		Frequency	
1	Less important for house officers	1	Rarely applied in clinical practices
2	Important for house officers	2	Commonly applied in clinical practices
3	Very important for house officers	3	Frequently applied in clinical practices

# University of Calgary's approach: Suitable if you use only ONE assessment tool. And love a bit of Maths.

(Coderre, Woloschuk, McLaughlin, 2009)

1. Tabulate curricular content

3. Sample opinion on weighting

4. Decide number of items

5. Decide task for each topic

Table 1. Blueprint for the undergraduate renal course at the University of Calgary.

Column #: 1	2	3	4	5	6	7	8	9	10
Presentation	Impact	Frequency	I x F	Weight	Number of items	Diagnosis	Investigation	Treatment	Basic science
Hypematremia	2	1	2	0.025	1.50	1	0	0	1
Hyponatremia	3	2	6	0.075	4.50	2	0	1	1
Hyperkalemia	3	3	9	0.1125	6.75	3	1	2	1
Hypokalemia	2	2	4	0.05	3.00	2	0	0	1
Acidosis	3	2	6	0.075	4.50	2	0	1	1
Alkalosis	2	2	4	0.05	3.00	2	0	0	1
ARF	3	3	9	0.1125	6.75	5	1	1	0
CRF	2	3	6	0.075	4.50	3	1	1	0
Hematuria	2	2	4	0.05	3.00	2	1	0	0
Proteinuria	2	3	6	0.075	4.50	2	0	0	2
Edema	1	3	3	0.0375	2.25	1	0	1	0
Scrotal mass	2	2	4	0.05	3.00	2	1	0	0
Urinary retention	1	3	3	0.0375	2.25	1	0	1	0
Hypertension	2	3	6	0.075	4.50	2	1	1	0
Polyuria	1	1	1	0.0125	0.75	1	0	0	0
Renal colic	1	3	3	0.0375	2.25	1	0	1	0
Dysuria	1	2	2	0.025	1.50	1	0	1	0
Incontinence	1	2	2	0.025	1.50	1	0	1	0
TOTAL			80	1	60	34	6	12	8

= (IXF)/  
Sum IF

= Weight X  
No of items

2. Provide relative weighting

Table 2. Weighting for impact and frequency of the clinical presentations.

Impact	Weight	Frequency	Weight
Non-urgent, little prevention potential	1	Rarely seen	1
Serious, but not immediately life threatening	2	Relatively common	2
Life threatening emergency and/or high potential for prevention impact	3	Very common	3

# STEP 6: Decide on number of item for each tool / task

(Ismail, Mat Pa, Mohammad, Yusoff, 2020)

						*USM MD Must know 60% Should know 30% Nice to know 10%	Summative assessment (40%) 5 single best answer 3 scenario-based question 2 OSCE			Continuous assessment Essay (20%) PBL (20%) Group video (20%)		
NO	TOPICS	MODE	IMPACT	FREQ.	I X F	Semester Exam			Pro 1 Exam			Cont.
						SBA	SBQ	OSCE	SBA	SBQ	OSCE	
1.	Introduction to Medical Professionalism	Lecture	3	3	9	/			/			Video
2.	Communication Skills in Medicine	Tutorial	3	3	9				/			
3.	Introduction to Clinical Skills	Tutorial & Practical	3	3	9	/			/			
4.	Patient Autonomy	Tutorial & PBL	3	2	6	/					PBL	
5.	Patient Confidentiality	Tutorial	3	3	9	/			/			PBL
6.	Patient Rights	Tutorial & PBL	3	2	6	/			/			
7.	Inter professionalism in Medicine	Tutorial & Hosp visit	2	2	4	/			/			
8.	Introduction to Medical Negligence	Lecture	1	1	1	/						
9.	Equity & Social Justice	Lecture	2	2	2	/						
10.	Academic Integrity	Workshop	3	3	9							Essay
11.	Professional Resilience	Workshop	3	2	6							Essay
12.	Reflection Skills	Workshop	2	3	6							Essay

(Ismail, Mat Pa, Mohammad, Yusoff, 2020)

NO	TOPICS	MODE	IMPACT	FREQ.	I X F	Semester Exam			Pro 1 Exam			Cont.	Lect.	
						SBA	SBQ	OSCE	SBA	SBQ	OSCE			
1.	Introduction to Medical Professionalism	Lecture	3	3	9	/			/				NAY	
2.	Communication Skills in Medicine	Tutorial	3	3	9	/			/			Video	NAY	
3.	Introduction to Clinical Skills	Tutorial & Practical	3	3	9	/		/			/		MNM	
4.	Patient Autonomy	Tutorial & PBL	3	2	6	/						PBL	MI	
5.	Patient Confidentiality	Tutorial	3	3	9		/		/				HVRA	
6.	Patient Rights	Tutorial & PBL	3	2	6	/				/		PBL	HVRA	
7.	Inter professionalism in Medicine	Tutorial & Hosp visit	2	2	4	/			/			Video	MRA	
8.	Introduction to Medical Negligence	Lecture	1	1	1	/								RAA
9.	Equity & Social Justice	Lecture	2	2	2	/								NAAH
10.	Academic Integrity	Workshop	3	3	9								Essay	KAB
11.	Professional Resilience	Workshop	3	2	6								Essay	NSR
12.	Reflection Skills	Workshop	2	3	6								Essay	NAO

**Table 6:** The summary of seven steps in constructing a blueprint

Steps	Description
1. Define the blueprinting purpose and scope	Identify its purpose and scope. For which semester or phase of study? Which academic session? Which courses? What assessment tools? How many questions?
2. Tabulate curricular content	Curricular contents – course learning outcomes, clinical presentations or topics – are listed based on curricular setting.
3. Identify impact and frequency	The impact and frequency for each curricular content are identified based on the selected criteria.
4. Categorise curricular content based on relative weightage	The curricular contents are classified as “must know”, “should know” and “nice to know” knowledge.
5. Decide on percentage of questions for each category	Determine how many percent questions should be constructed from “must know”, “should know” and “nice to know” knowledge.
6. Decide on number of item for each assessment task	Decide on how many questions should be constructed for each category of curricular content.
7. Assign questions to lecturers for items preparation	Identify question makers for items preparation. Practically, the one who teaches the curricular content should prepare the questions.

(Ismail, Mat Pa, Mohammad, Yusoff, 2020)

## Twelve tips from Calgary’s

1. Tabulate curricular content
2. Provide relative weightage
3. Sample opinion on weightage from stakeholders
4. Decide number of items
5. Decide tasks  
(investigation/diagnosis/management)
6. Create evaluation forms
7. Create item banks
8. **Revise learning objectives**
9. **Revise learning experience**
10. Distribute blueprint to educators
11. **Monitor content validity**
12. **Distribute blueprint to learners**

(Coderre, Woloschuk, McLaughlin, 2009)

# ITEM ANALYSIS

**1415-1500**

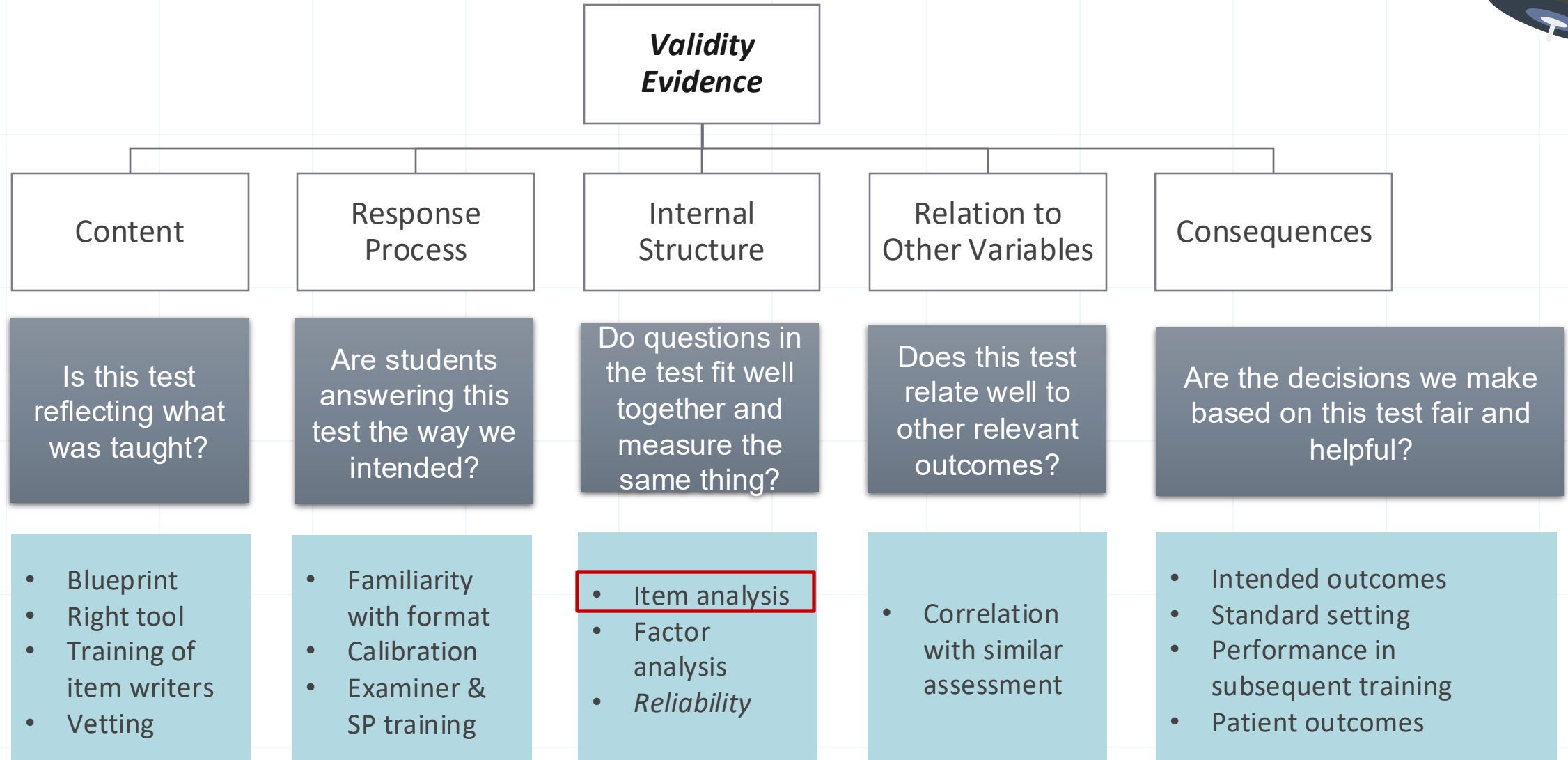
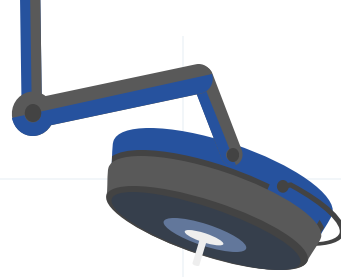
At the end of the session, participants will be able to

interpret basic item analysis in MCQ and select good items for banking.





# Proposing validity in assessment



# Examination process – how are your practice here?



# Assessment that fits purposes

01

Assessment that **facilitates learning**

02

Assessment that **supports high stake decisions**

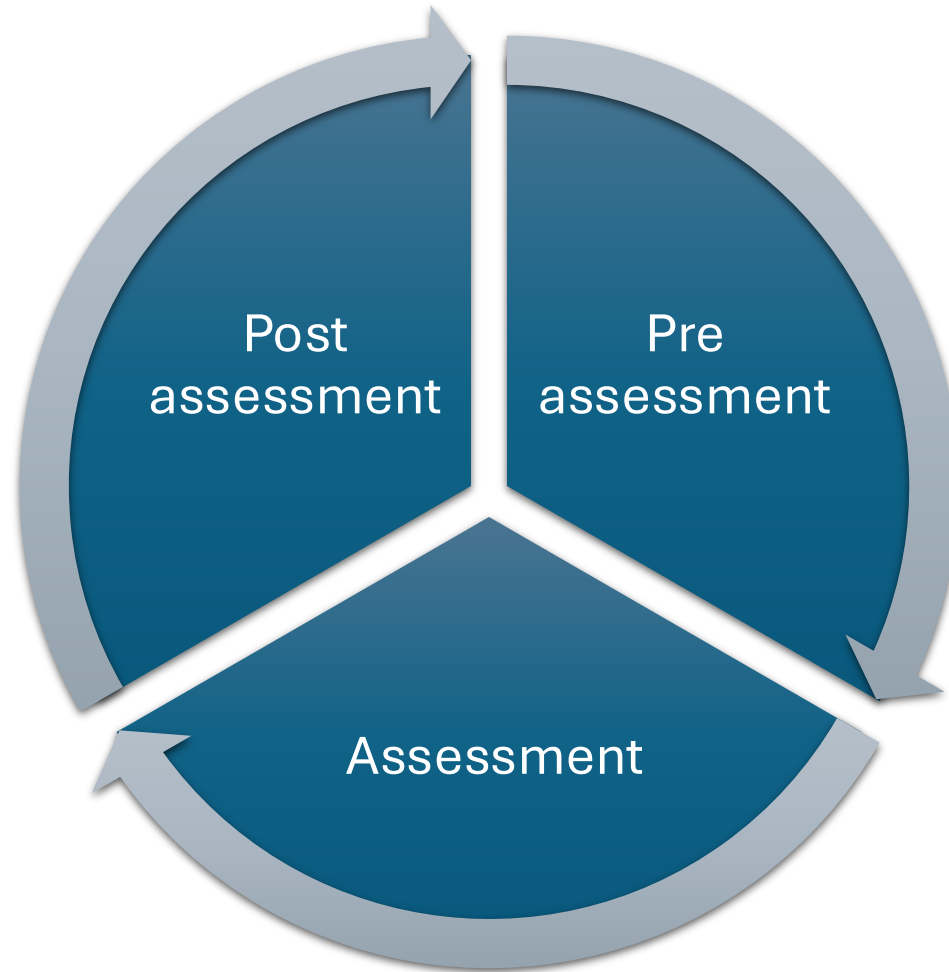
03

Assessment that **informs curriculum**

van der Vleuten CP, Schuwirth LW, Driessen EW, Dijkstra J, Tigelaar D, Baartman LK, van Tartwijk J. A model for programmatic assessment fit for purpose. Med Teach. 2012;34(3):205-14

# Assessment is a *loop*.

- Standard-setting
- Data entry
- Item analysis
- Examination analysis
- Reporting
- Student appeal procedures
- Feedback



- Blueprinting
- Question construction
- Vetting
- Student preparation
- Question security

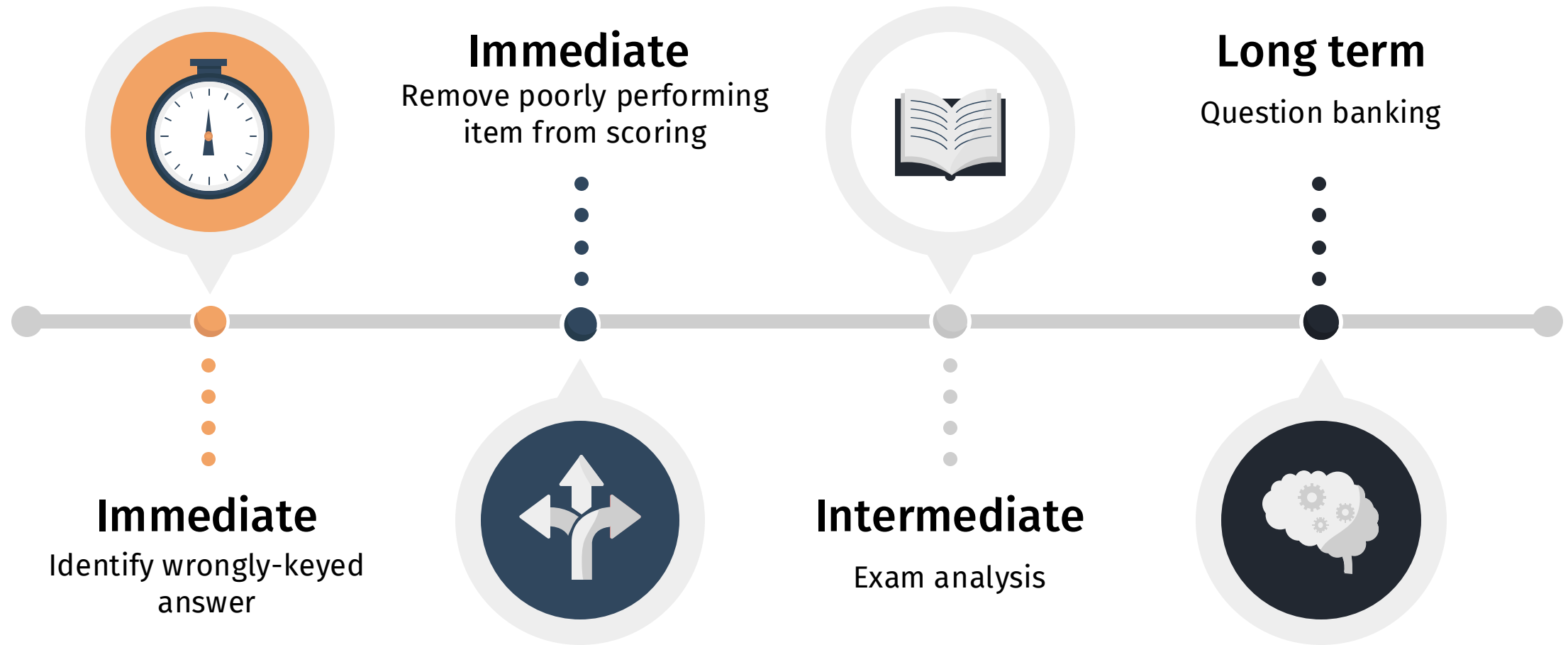
- Examination environment
- Standardization

# What makes a *good* question?



# Item Analysis

process of evaluating **individual** test questions (items) to determine their quality and ability to differentiate between different levels of student performance







## Item Analysis

Objective format - OMR machine  
Essays or other formats - manual

Condensed Test Report

Generated By Remark® From Gravic, Inc.

						2	Test Item Statistics Report									
						3	Generated By Remark® From Gravic, Inc.									
						4										
						5										
						6										
						7										
						8										
						9										
						10										
						11										
						12										
						13										
						14										
						15										
						16										
						17										
						18										
						19										
						20										
						21										
						22										
						23										
						24										
						25										
						26										
						27										
						28										
						29										
						30										
						31										
						32										
						33										
						34										
						35										
						36										
						37										
						38										
						39										
						40										
						41										
						42										
						43										
						44										
						45										
						46										
						47										
						48										
						49										
						50										
						51										
						52										
						53										
						54										
						55										
						56										
						57										
						58										
						59										
						60										
						61										
						62										
						63										
						64										
						65										
						66										
						67										
						68										
						69										
						70										
						71										
						72										
						73										
						74										
						75										
						76										
						77										
						78										
						79										
						80										
						81										
						82										
						83										
						84										
						85										
						86										
						87										
						88										
						89										
						90										
						91										
						92										
						93										
						94										
						95										
						96										
						97										
						98										
						99										
						100										

# 3 main input from Item Analysis

01

Difficulty index – **how many students can answer?**

02

Discrimination index – **can the question differentiate good and weak students?**

03

Distractor analysis – **are all the options functioning?**

# IIUM format – OBA

## ITEM ANALYSIS REPORT

Form Type :			New Form Single Response			Total Student :			37		Difficulty index		Discrimination index	
Center Name :			PAED			Total Question :			Invalid Question :		Marks / Question			
Subject Code :			2433			Mark :								
Examination Date :			09-Jun-2025											
Que	Master key	A	B	C	D	E	Blank	Multipl	Correct	Wrong	Passing Index (PI)	PI Grade	Discrimination Index (DI)	DI Grade
1	B	3	28	4	2	0	0	0	28	9	0.76	Easy	0.13	Very Difficult
		8.11	75.68	10.81	5.41	0.00	0.00	0.00	75.68	24.32				
2	D	0	2	0	35	0	0	0	35	2	0.95	Very Easy	-0.07	Very Difficult
		0.00	5.41	0.00	94.59	0.00	0.00	0.00	94.59	5.41				
3	D	3	19	2	13	0	0	0	13	24	0.35	Moderately Difficult	0.47	Moderately Difficult
		8.11	51.35	5.41	35.14	0.00	0.00	0.00	35.14	64.86				
4	A	28	4	1	4	0	0	0	28	9	0.76	Easy	0.00	Extreamly Difficult
		75.68	10.81	2.70	10.81	0.00	0.00	0.00	75.68	24.32				
5	C	9	2	21	5	0	0	0	21	16	0.57	Moderately Easy	0.47	Moderately Difficult
		24.32	5.41	56.76	13.51	0.00	0.00	0.00	56.76	43.24				
6	C	6	1	23	7	0	0	0	23	14	0.62	Moderately Easy	0.20	Difficult
		16.22	2.70	62.16	18.92	0.00	0.00	0.00	62.16	37.84				

\*EMQ cannot be captured as options here is only until E

# IIUM format – MTF

## ITEM ANALYSIS REPORT

Correct  
answer

Type : New Form Multiple True/False  
Name : O&G  
Code : 2433  
Examination Date : 09-Jun-2025

Total Student :  
Total Question :  
Invalid Question :

Difficulty  
index

Discrimination  
index

MTF 1

Que	Master key	True	False	Blank	Multipl	Correct	Wrong	Passing Index (PI)	PI Grade	Discrimination Index (DI)	DI Grade
1A.	T	40 100.00	0 0.00	0 0.00	0 0.00	40 100.00	0 0.00	1.00	Extreamly Easy	0.06	Very Difficult
1B.	F	1 2.50	39 97.50	0 0.00	0 0.00	39 97.50	1 2.50	0.98	Extreamly Easy	0.06	Very Difficult
1C.	F	1 2.50	38 95.00	1 2.50	0 0.00	38 95.00	1 2.50	0.95	Extreamly Easy	-0.06	Very Difficult
1D.	T	25 62.50	7 17.50	8 20.00	0 0.00	25 62.50	7 17.50	0.63	Moderately Easy	0.13	Very Difficult
1E.	F	30 75.00	5 12.50	5 12.50	0 0.00	5 12.50	30 75.00	0.13	Very Difficult	-0.13	Very Difficult
2A.	F	7 17.50	31 77.50	2 5.00	0 0.00	31 77.50	7 17.50	0.78	Easy	0.31	Moderately Difficult
2B.	F	20 50.00	20 50.00	0 0.00	0 0.00	20 50.00	20 50.00	0.50	Moderately Easy	0.56	Moderately Easy
2C.	F	3 7.50	24 60.00	13 32.50	0 0.00	24 60.00	3 7.50	0.60	Moderately Easy	0.13	Very Difficult



# Difficulty (*Easiness*) Index

- Also known as Facility Index or *p* in OMR machine
- Percentage of candidates who answered the item correctly
- Range 0.0 - 1.0 (the higher the easier)




Que	Master key	A	B	C	D	E	Blank	Multipl	Correct	Wrong	Passing Index (PI)	PI Grade	Discrimination Index (DI)	DI Grade
1	B	3 8.11	28 75.68	4 10.81	2 5.41	0 0.00	0 0.00	0 0.00	28 75.68	9 24.32	0.76	Easy	0.13	Very Difficult
2	D	0 0.00	2 5.41	0 0.00	35 94.59	0 0.00	0 0.00	0 0.00	35 94.59	2 5.41	0.95	Very Easy	-0.07	Very Difficult
3	D	3 8.11	19 51.35	2 5.41	13 35.14	0 0.00	0 0.00	0 0.00	13 35.14	24 64.86	0.35	Moderately Difficult	0.47	Moderately Difficult
4	A	28 75.68	4 10.81	1 2.70	4 10.81	0 0.00	0 0.00	0 0.00	28 75.68	9 24.32	0.76	Easy	0.00	Extreamly Difficult
5	C	9 24.32	2 5.41	21 56.76	5 13.51	0 0.00	0 0.00	0 0.00	21 56.76	16 43.24	0.57	Moderately Easy	0.47	Moderately Difficult
6	C	6 16.22	1 2.70	23 62.16	7 18.92	0 0.00	0 0.00	0 0.00	23 62.16	14 37.84	0.62	Moderately Easy	0.20	Difficult

# Difficulty Index – essay?

## DIFFICULTY (*easiness*) INDEX

1. Decide what is accepted as correct. (eg. 6/10)
2. Formula =  
percentage of  
candidates  
exceeded 'correct'

	 Name	 Q1	 Q1_PB	 Q2	 Q3	 Q4	 Totalscore
1	Muthu	7.00	1.00	9.00	8.00	9.00	33.00
2	Lina	9.00	1.00	9.00	6.00	9.00	33.00
3	Ali	8.00	1.00	6.00	9.00	9.00	32.00
4	Yen	6.00	1.00	8.00	7.00	10.00	31.00
5	Sani	6.00	1.00	6.00	7.00	7.00	26.00
6	Diva	9.00	1.00	7.00	7.00	3.00	26.00
7	Tan	8.00	1.00	5.00	6.00	6.00	25.00
8	Abu	6.00	1.00	6.00	5.00	7.00	24.00
9	Lam	5.00	.00	7.00	5.00	7.00	24.00
10	Tini	6.00	1.00	3.00	5.00	6.00	20.00
11	John	4.00	.00	4.00	4.00	4.00	16.00
12	Greg	3.00	.00	4.00	2.00	4.00	13.00

Say we take 6 and  
above as correct


Diff index (Q1)  
=9/12  
=0.75



# Discrimination Index

- Ability of an item to differentiate between high ability and low ability students
- 27% formula (d) or point biserial (r)
- Range -1.0 to 1.0 (the higher the better)
- Most important - reflects the item ability to contribute to assessment objective

Discrimination  
index



Que	Master key	A	B	C	D	E	Blank	Multipl	Correct	Wrong	Passing Index (PI)	PI Grade	Discrimination Index (DI)	DI Grade
1	B	3 8.11	28 75.68	4 10.81	2 5.41	0 0.00	0 0.00	0 0.00	28 75.68	9 24.32	0.76	Easy	0.13	Very Difficult
2	D	0 0.00	2 5.41	0 0.00	35 94.59	0 0.00	0 0.00	0 0.00	35 94.59	2 5.41	0.95	Very Easy	-0.07	Very Difficult
3	D	3 8.11	19 51.35	2 5.41	13 35.14	0 0.00	0 0.00	0 0.00	13 35.14	24 64.86	0.35	Moderately Difficult	0.47	Moderately Difficult
4	A	28 75.68	4 10.81	1 2.70	4 10.81	0 0.00	0 0.00	0 0.00	28 75.68	9 24.32	0.76	Easy	0.00	Extremely Difficult
5	C	9 24.32	2 5.41	21 56.76	5 13.51	0 0.00	0 0.00	0 0.00	21 56.76	16 43.24	0.57	Moderately Easy	0.47	Moderately Difficult
6	C	6 16.22	1 2.70	23 62.16	7 18.92	0 0.00	0 0.00	0 0.00	23 62.16	14 37.84	0.62	Moderately Easy	0.20	Difficult

# Discrimination Index

1. First, sum the candidate score for the test (imply ability in the test)
2. Rank the sum score

## 27% formula (d)

= Average difficulty index of top 27% - average difficulty index of bottom 27%

27% is the minimum group size to compare performance without having overlap

## Point biserial (r)

No need to rank candidates.

Correlating each students question performance (1=correct, 0= wrong) with total score.

\*Utilize data from whole class. Not just top and bottom 27%

Que	Master key	A	B	C	D	E	Blank	Multipl	Correct	Wrong	Passing Index (PI)	PI Grade	Discrimination Index (DI)	DI Grade
1	B	3 8.11	28 75.68	4 10.81	2 5.41	0 0.00	0 0.00	0 0.00	28 75.68	9 24.32	0.76	Easy	0.13	Very Difficult
2	D	0 0.00	2 5.41	0 0.00	35 94.59	0 0.00	0 0.00	0 0.00	35 94.59	2 5.41	0.95	Very Easy	-0.07	Very Difficult
3	D	3 8.11	19 51.35	2 5.41	13 35.14	0 0.00	0 0.00	0 0.00	13 35.14	24 64.86	0.35	Moderately Difficult	0.47	Moderately Difficult
4	A	28 75.68	4 10.81	1 2.70	4 10.81	0 0.00	0 0.00	0 0.00	28 75.68	9 24.32	0.76	Easy	0.00	Extreamly Difficult
5	C	9 24.32	2 5.41	21 56.76	5 13.51	0 0.00	0 0.00	0 0.00	21 56.76	16 43.24	0.57	Moderately Easy	0.47	Moderately Difficult
6	C	6 16.22	1 2.70	23 62.16	7 18.92	0 0.00	0 0.00	0 0.00	23 62.16	14 37.84	0.62	Moderately Easy	0.20	Difficult

# Discrimination Index – essay?

## 27% formula (d)

= Difficulty index of top 27% -  
difficulty index of bottom 27%  
=  $3/3 - 1/3$   
=  $1 - 0.33$   
= 0.67

## Point biserial (r)

Correlating each students question  
performance (1=correct, 0= wrong) with total  
score.  
Utilize data from whole class. Not just top  
and bottom 27%

## Item-total correlation (r)

Correlating each students question raw  
score with total score.  
Utilize data from whole class. Not just top  
and bottom 27%

	Name	Q1	Q1_PB	Q2	Q3	Q4	Totalscore
1	Muthu	7.00	1.00	9.00	8.00	9.00	33.00
2	Lina	9.00	1.00	9.00	6.00	9.00	33.00
3	Ali	8.00	1.00	6.00	9.00	9.00	32.00
4	Yen	6.00	1.00	8.00	7.00	10.00	31.00
5	Sani	6.00	1.00	6.00	7.00	7.00	26.00
6	Diva	9.00	1.00	7.00	7.00	3.00	26.00
7	Tan	8.00	1.00	5.00	6.00	6.00	25.00
8	Abu	6.00	1.00	6.00	5.00	7.00	24.00
9	Lam	5.00	.00	7.00	5.00	7.00	24.00
10	Tini	6.00	1.00	3.00	5.00	6.00	20.00
11	John	4.00	.00	4.00	4.00	4.00	16.00
12	Greg	3.00	.00	4.00	2.00	4.00	13.00

# Can AI ease my work here?

if i upload excel of my students essay scores, can you calculate difficulty index and discrimination index (27% approach and biserial)

Yes, absolutely! Please go ahead and upload your Excel file with your students' essay scores. Once I have it, I can help you calculate:

1. **Difficulty Index**
2. **Discrimination Index using:**
  - **27% Upper-Lower Group Method**
  - **Point-Biserial Correlation**

Let me know if there's anything specific in the file (e.g. which sheet or columns) I should look at, or you can just upload it and I'll take a look.

# Distractor Analysis

- Analyse the effectiveness of incorrect options (distractor) in MCQ
- Function distractor if 5% or more candidates selected it.
- Non-functioning distractor if <5% of candidates selected it** (Shakurnia et al, 2022)
  - Could indicate cueing effect or ambiguity issue

Que	Master key	A	B	C	D	E	Blank	Multipl	Correct	Wrong	Passing Index (PI)	PI Grade	Discrimination Index (DI)	DI Grade
1	B	3 8.11	28 75.68	4 10.81	2 5.41	0 0.00	0 0.00	0 0.00	28 75.68	9 24.32	0.76	Easy	0.13	Very Difficult
2	D	0 0.00	2 5.41	0 0.00	35 94.59	0 0.00	0 0.00	0 0.00	35 94.59	2 5.41	0.95	Very Easy	-0.07	Very Difficult
3	D	3 8.11	19 51.35	2 5.41	13 35.14	0 0.00	0 0.00	0 0.00	13 35.14	24 64.86	0.35	Moderately Difficult	0.47	Moderately Difficult
4	A	28 75.68	4 10.81	1 2.70	4 10.81	0 0.00	0 0.00	0 0.00	28 75.68	9 24.32	0.76	Easy	0.00	Extremely Difficult
5	C	9 24.32	2 5.41	21 56.76	5 13.51	0 0.00	0 0.00	0 0.00	21 56.76	16 43.24	0.57	Moderately Easy	0.47	Moderately Difficult
6	C	6 16.22	1 2.70	23 62.16	7 18.92	0 0.00	0 0.00	0 0.00	23 62.16	14 37.84	0.62	Moderately Easy	0.20	Difficult

# Item Analysis: Interpretation

Most informative test item

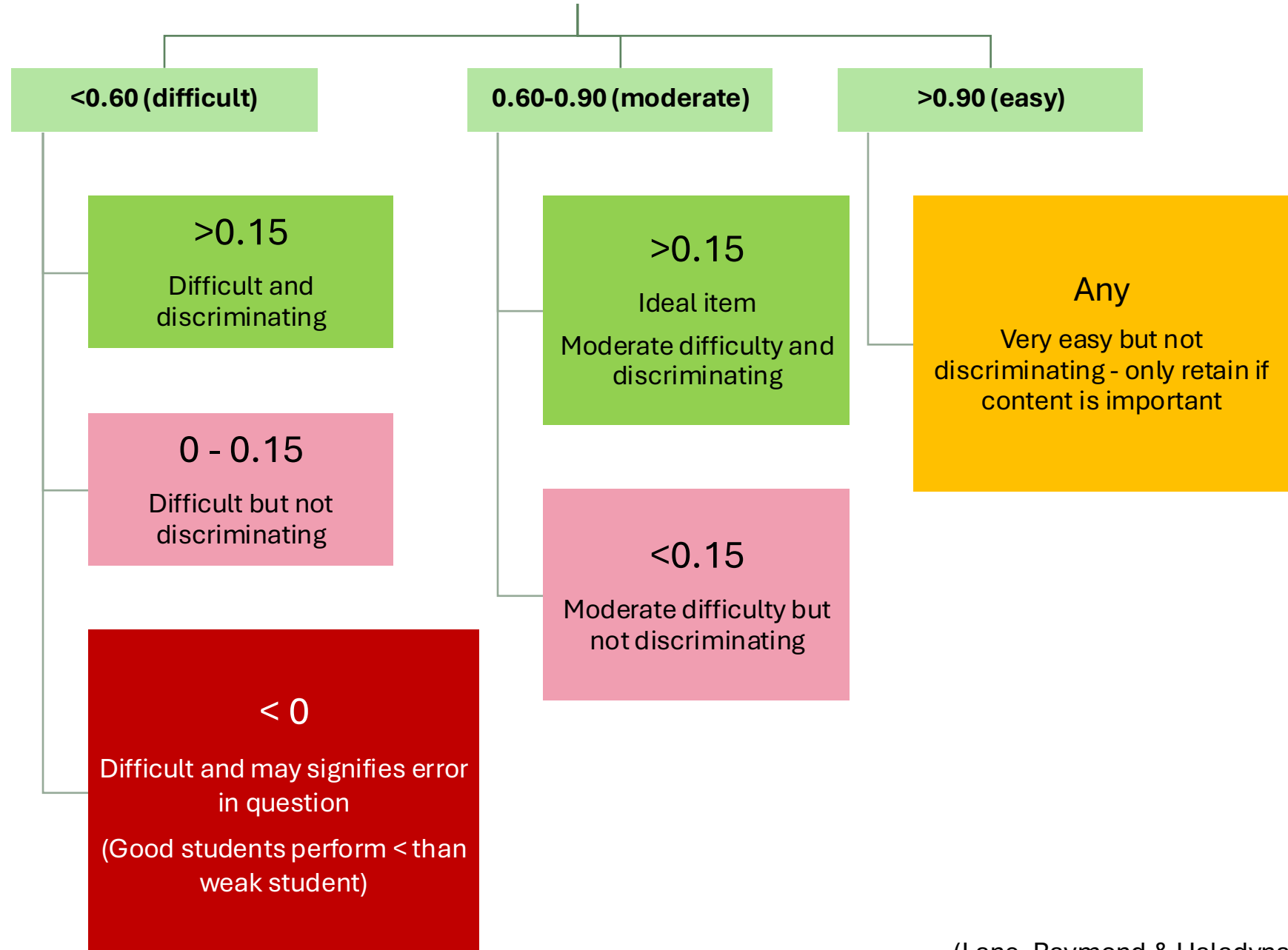
- Moderate difficulty
- Discriminate highly

Table 20.3 Some Arbitrary Standards for Evaluating Item Statistics for SR and CROS Items

Type	Difficulty	Disc. <i>Some use 0.20</i>	Comment
1	0.60 to 0.90	> 0.15	Ideal item; moderate difficulty and high discrimination
2	0.60 to 0.90	< 0.15	Poor discrimination
3	Above 0.90	Disregard	High performance item; usually not very discriminating
4	< 0.60	> 0.15	Difficult but very discriminating
5	< 0.60	< 0.15	Difficult and nondiscriminating
6	< 0.60	< 0.00	Identical to type 5 except that one of the distractors has a pattern like type 1, which signifies a key error

# 1. What is the facility / difficulty index?

## 2. What is the discrimination index?





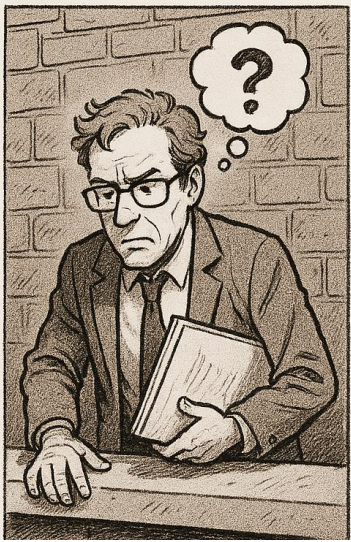
Que	Master key	A	B	C	D	E	Blank	Multipl	Correct	Wrong	Passing Index (PI)	PI Grade	Discrimination Index (DI)	DI Grade	Moderate difficulty but not discriminating
1	B	3 8.11	28 75.68	4 10.81	2 5.41	0 0.00	0 0.00	0 0.00	28 75.68	9 24.32	0.76	Easy	0.13	Very Difficult	Check content – very important?
2	D	0 0.00	2 5.41	0 0.00	35 94.59	0 0.00	0 0.00	0 0.00	35 94.59	2 5.41	0.95	Very Easy	-0.07	Very Difficult	
3	D	3 8.11	19 51.35	2 5.41	13 35.14	0 0.00	0 0.00	0 0.00	13 35.14	24 64.86	0.35	Moderately Difficult	0.47	Moderately Difficult	Good Qs - bank
4	A	28 75.68	4 10.81	1 2.70	4 10.81	0 0.00	0 0.00	0 0.00	28 75.68	9 24.32	0.76	Easy	0.00	Extreamly Difficult	
5	C	9 24.32	2 5.41	21 56.76	5 13.51	0 0.00	0 0.00	0 0.00	21 56.76	16 43.24	0.57	Moderately Easy	0.47	Moderately Difficult	Good Qs - bank
6	C	6 16.22	1 2.70	23 62.16	7 18.92	0 0.00	0 0.00	0 0.00	23 62.16	14 37.84	0.62	Moderately Easy	0.20	Difficult	

1. Which item you need to check for error immediately?
2. Which item that were too easy?
3. Which items contains non-functioning distractors?
4. Which items you can bank now for future use?

Que	Master key	A	B	C	D	E	Blank	Multipl	Correct	Wrong	Passing Index (PI)	PI Grade	Discrimination Index (DI)	DI Grade
2	D	0 0.00	2 5.41	0 0.00	35 94.59	0 0.00	0 0.00	0 0.00	35 94.59	2 5.41	0.95	Very Easy	-0.07	Very Difficult
3	D	3 8.11	19 51.35	2 5.41	13 35.14	0 0.00	0 0.00	0 0.00	13 35.14	24 64.86	0.35	Moderately Difficult	0.47	Moderately Difficult
4	A	28 75.68	4 10.81	1 2.70	4 10.81	0 0.00	0 0.00	0 0.00	28 75.68	9 24.32	0.76	Easy	0.00	Extreamly Difficult
5	C	9 24.32	2 5.41	21 56.76	5 13.51	0 0.00	0 0.00	0 0.00	21 56.76	16 43.24	0.57	Moderately Easy	0.47	Moderately Difficult
6	C	6 16.22	1 2.70	23 62.16	7 18.92	0 0.00	0 0.00	0 0.00	23 62.16	14 37.84	0.62	Moderately Easy	0.20	Difficult
7	D	9 24.32	9 24.32	12 32.43	7 18.92	0 0.00	0 0.00	0 0.00	7 18.92	30 81.08	0.19	Difficult	0.33	Moderately Difficult
8	A	12 32.43	1 2.70	1 2.70	23 62.16	0 0.00	0 0.00	0 0.00	12 32.43	25 67.57	0.32	Moderately Difficult	0.13	Very Difficult
9	A	19 51.35	13 35.14	3 8.11	2 5.41	0 0.00	0 0.00	0 0.00	19 51.35	18 48.65	0.51	Moderately Easy	0.33	Moderately Difficult
10	A	7 18.92	0 0.00	2 5.41	28 75.68	0 0.00	0 0.00	0 0.00	7 18.92	30 81.08	0.19	Difficult	-0.07	Moderately Difficult

1. Which item you need to check for error immediately? – **Item 10**
2. Which item that were too easy? - **Item 2**
3. Which items contains non-functioning distractors? – **Item 2, 4, 6, 8, 10**
4. Which items you can bank now for future use? – **Item 3, 5, 7, 9**

Que	Master key	A	B	C	D	E	Blank	Multipl	Correct	Wrong	Passing Index (PI)	PI Grade	Discrimination Index (DI)	DI Grade
2	D	0 0.00	2 5.41	0 0.00	35 94.59	0 0.00	0 0.00	0 0.00	35 94.59	2 5.41	0.95	Very Easy	-0.07	Very Difficult
3	D	3 8.11	19 51.35	2 5.41	13 35.14	0 0.00	0 0.00	0 0.00	13 35.14	24 64.86	0.35	Moderately Difficult	0.47	Moderately Difficult
4	A	28 75.68	4 10.81	1 2.70	4 10.81	0 0.00	0 0.00	0 0.00	28 75.68	9 24.32	0.76	Easy	0.00	Extremely Difficult
5	C	9 24.32	2 5.41	21 56.76	5 13.51	0 0.00	0 0.00	0 0.00	21 56.76	16 43.24	0.57	Moderately Easy	0.47	Moderately Difficult
6	C	6 16.22	1 2.70	23 62.16	7 18.92	0 0.00	0 0.00	0 0.00	23 62.16	14 37.84	0.62	Moderately Easy	0.20	Difficult
7	D	9 24.32	9 24.32	12 32.43	7 18.92	0 0.00	0 0.00	0 0.00	7 18.92	30 81.08	0.19	Difficult	0.33	Moderately Difficult
8	A	12 32.43	1 2.70	1 2.70	23 62.16	0 0.00	0 0.00	0 0.00	12 32.43	25 67.57	0.32	Moderately Difficult	0.13	Very Difficult
9	A	19 51.35	13 35.14	3 8.11	2 5.41	0 0.00	0 0.00	0 0.00	19 51.35	18 48.65	0.51	Moderately Easy	0.33	Moderately Difficult
10	A	7 18.92	0 0.00	2 5.41	28 75.68	0 0.00	0 0.00	0 0.00	7 18.92	30 81.08	0.19	Difficult	-0.07	Moderately Difficult



# Fitting Item Analysis into your busy schedule

Immediate

Check for **negative discrimination index**

Are they key errors?

Intermediate CO-  
PO  
CQI

Check for items with **low difficulty index**

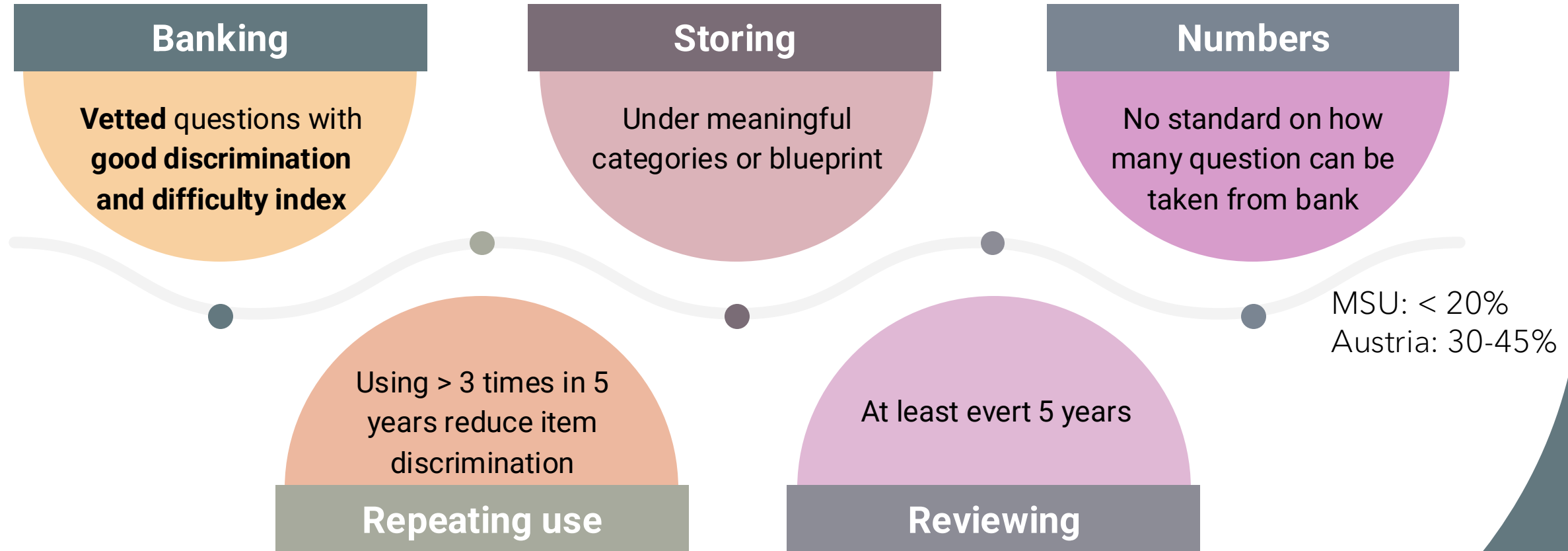
Which topics  
students did not  
perform?  
Was it student or  
question factor?

Long term

Select items with **moderate difficulty and good discrimination**

Bank this items.  
Review items with  
non-functioning  
distractor

# Question Banking: Some guides



## How can we explain MCQ that has poor difficulty / discriminator index?

### *Constructive alignment*

Link to the LO?

### *Quality*

Is the question clear?

For MCQ, involve 6 or less thinking steps?

Plausible / Functional distractors?  
(more than 5% candidates selected it)

For SBA, can item have answered without looking at options?

In modern political theory, democratic governments are defined by certain principles that emphasize the importance of individual rights, popular sovereignty, and equality before the law. The hallmark of a democracy is that its leaders are accountable to the people, either through periodic elections or other forms of public oversight.

Which of the following best describes a characteristic of a democratic government?

- A. Power is inherited and passed down through royal bloodline
- B. The government operates through a single-party system
- C. Citizens have the right to vote in free and fair elections, and they are protected by laws that ensure freedom of speech and political participation
- D. Political decisions are made solely by military leaders

### *Test-wiseness, cueing*

Avoid items under similar domain next to each other?

Avoid grammatical cues?

Answer choices have around similar length?

For answer choices with numbers, listed in ascending order?

Avoid 'not' or 'except' (or bold them when used)?

Avoid all of the above (AOTA) or NOTA?

Avoid K-type  
a. I, II, and IV?

## How can we explain OSCE that has poor difficulty / discriminator index?



Indices that may be affected	Difficulty Index	Discrimination Index
Domain <ul style="list-style-type: none"><li>Is OSCE assessing skills or theoretical component?</li></ul>	/	/
Authenticity <ul style="list-style-type: none"><li>Is the case clinically authentic?</li><li>Is the level of difficulty appropriate?</li></ul>	/	
Instruction <ul style="list-style-type: none"><li>Is the instruction clear and concise?</li><li>Does the instruction orientate the candidate towards the task?</li></ul>	/	/
Checklist <ul style="list-style-type: none"><li>Include discriminating item</li><li>Avoid mark for nonspecific thoroughness</li></ul>	/	/
Time <ul style="list-style-type: none"><li>Is there enough time to understand instruction?</li><li>Is there enough time to perform task?</li></ul>	/	



# Any limitation on this item analysis?



## Sample dependent

- 200 – stable
- <100 – with caution
- <30 – with caution but still can use

## Circular dependency

- Item difficulty affect discrimination index

Relies on total score as performance indicator

# Main take home message.

- Item analysis
  - Immediate – detect key error
  - May inform CO-PO and CQI
  - Long run – select good questions for banking
- Traditional item analysis (CTT) has values, somewhat comparable to IRT and feasible to interpret



# STANDARD SETTING

0900-1230

At the end of the session, participants  
will be able to

apply Modified Angoff standard setting  
method to theory examination.



## **ASSESSMENT VALIDITY**

Tuesday, 0915-1000

## **PRINCIPLES OF CONSTRUCTING THEORY QUESTIONS**

Tuesday, 1000-1300

## **ASSESSMENT BLUEPRINT**

Tuesday, 1415-1500

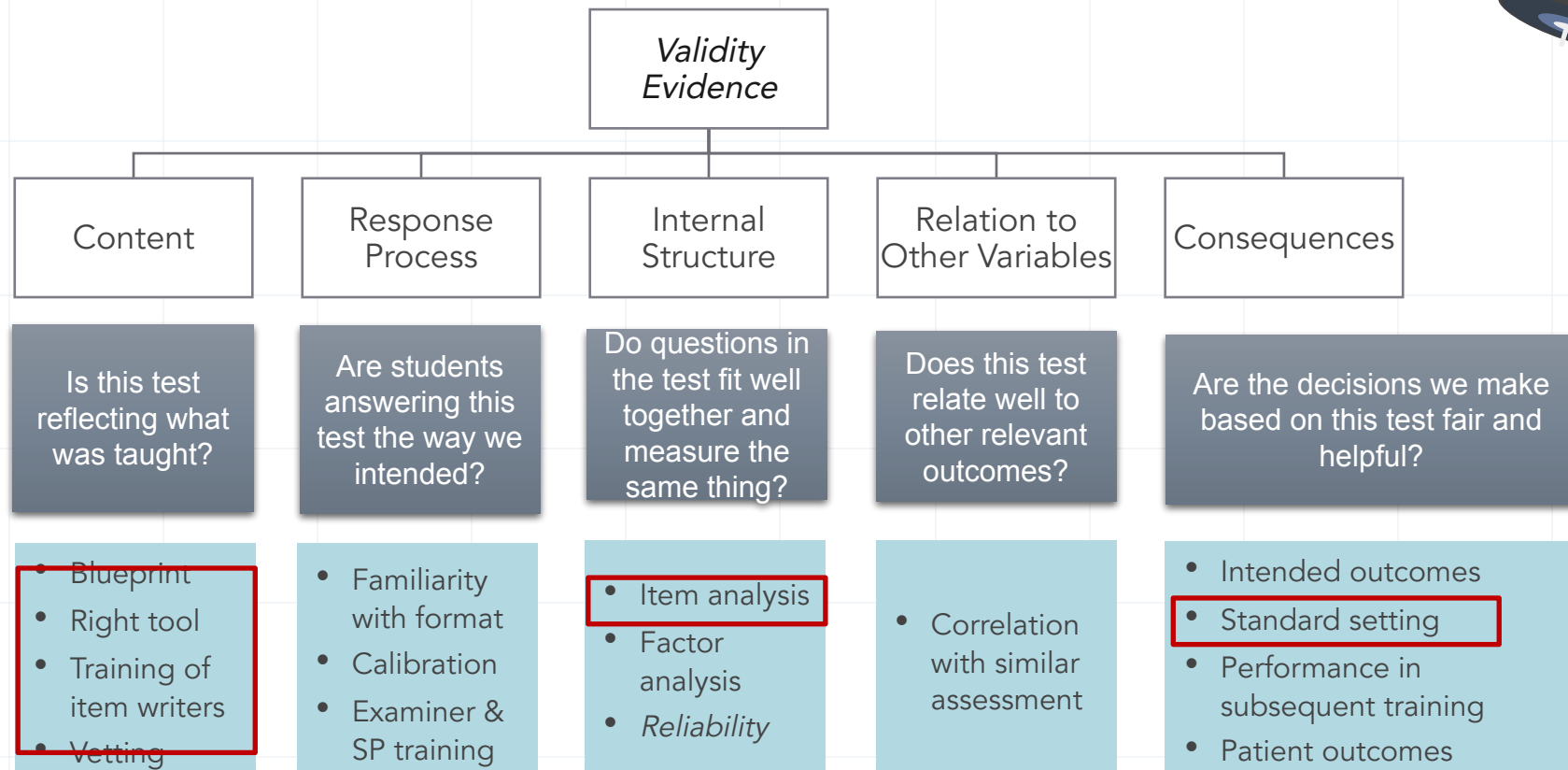
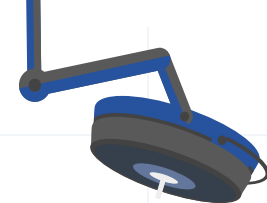
## **ITEM ANALYSIS**

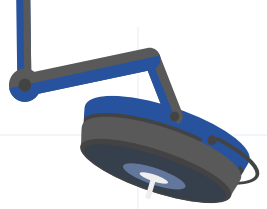
Tuesday, 1500-1700

## **STANDARD SETTING**

Wednesday, 0900-1230

# Proposing validity in assessment



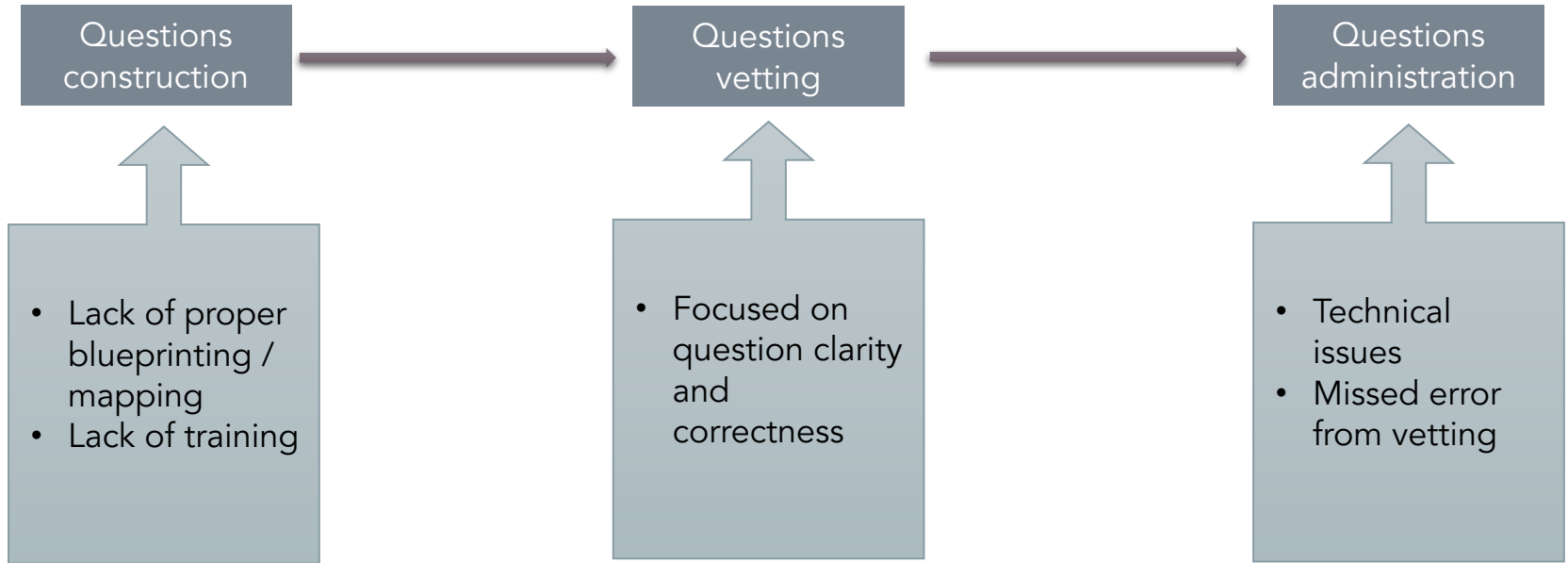


# MMC Standards for Undergraduate Medical Students (2019)

## 2.2.3.

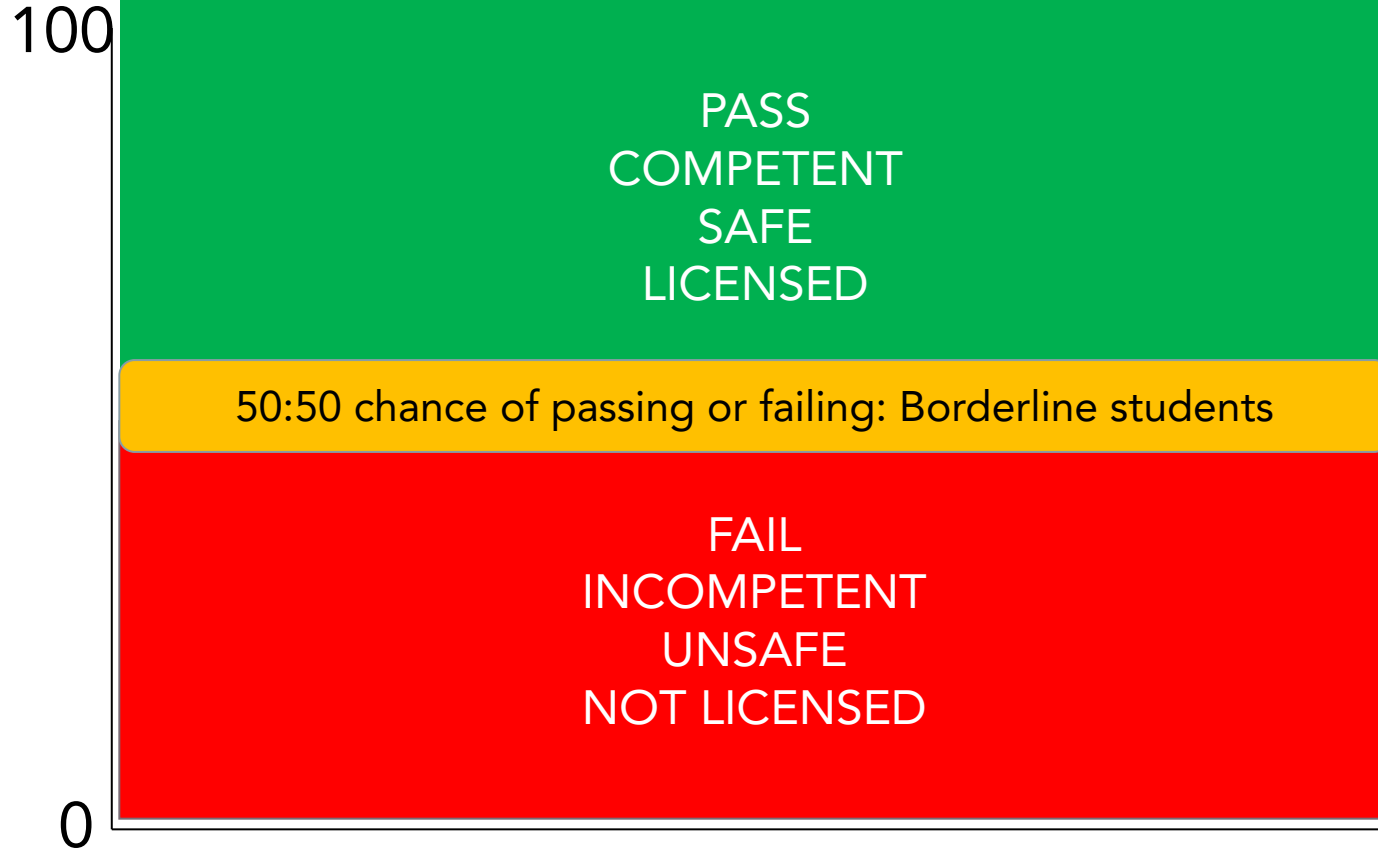
The medical school must document and communicate to students the frequency, methods, and criteria of student assessment - including the grading system, **the criteria for setting pass marks**, grade boundaries, rules of progression, number of allowed retakes and appeal policies.

# How to justify 50%?

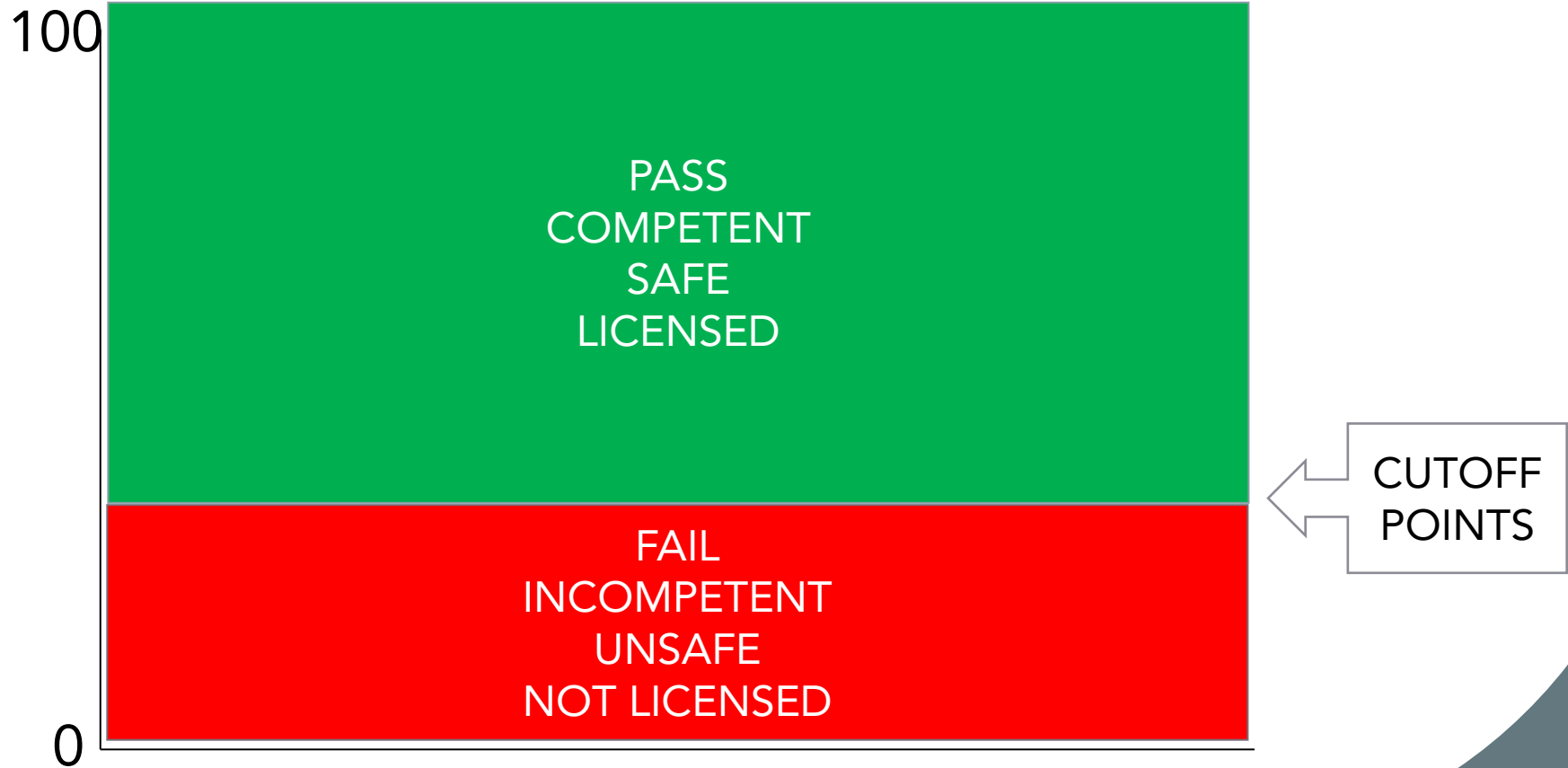




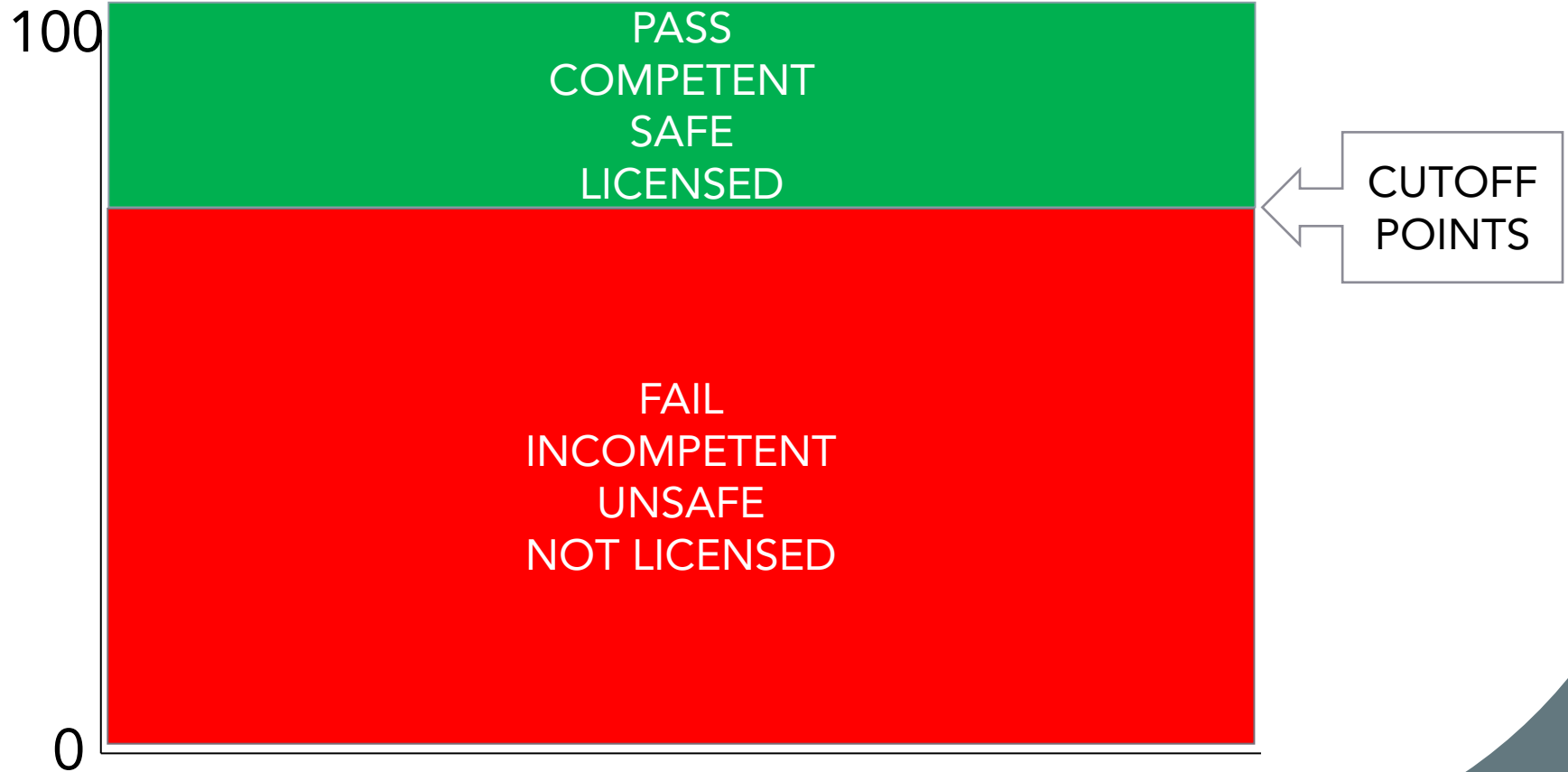
## STANDARD SETTING: Why?



## STANDARD SETTING: Difficult assessment?



## STANDARD SETTING: Easy assessment?



“The proper following of a prescribed, rational system of rules or procedures resulting in the assignment of a number to differentiate between two or more states or degrees of performance”

(Cizek, 1993)

COMPETENCE → PASSING SCORE

(Kane, 1994; Norcini, 1994)



## Education and training standard setting

The specific type of standard setting used in the RCPCH is a **Modified Angoff Method**. Following each exam diet a panel of professional judges who are subject matter experts grade the difficulty of the exam, by item, which results in a cut-score or pass mark.

Prior to grading the exam, judges are asked to conceptualise the notion of the 'borderline candidate'. The borderline candidate is a fictitious person who has a 50% chance of passing the exam. Once this has been established, the judges are asked to independently grade each item of the exam and rate out of 100 borderline candidates the percentage of how many will answer the question correctly. Once completed these ratings are compiled and prepared for a panel discussion.

The judges will then come together as a panel to examine the ratings for each item in sequence and discuss each item's rating. Judges will discuss disparity in ratings and different viewpoints. Once these have been considered, if they wish, judges will have an opportunity to re-rate items based on the discussions, taking into account:

- the judgements of the entire panel
- the definition of the borderline candidate
- the comments of the panel.

## Assessment Strategy for 2021 Anaesthetics Curriculum

**Published:** 12/01/2023

Introduction

Executive summary

Purpose of programme of assessment

How the programme of assessment is to be used with the curriculum

Critical Progression Points

Overarching assessment blueprint

RCoA Assessments

FRCA Examinations

The Primary FRCA examinations

The Final FRCA examinations

Linking curriculum content to the examinations – Blueprints and examination syllabus

Validity of the Examinations

### Standard setting

Standard setting: Primary and Final MCQ examinations

Standard setting: Primary FRCA OSCE

Standard setting: Primary and Final FRCA OSCE

### Standard setting

The FRCA examinations are high-stake summative assessments that have the potential to impact on trainee careers and patient safety. The format of the examinations was approved by the GMC in September 2009, and they have continued to be used as a means of ensuring anaesthetic anaesthetists in training have the appropriate minimum level of knowledge and skills to progress to the next stage of their training programme. With this purpose, the processes that underpin pass/fail decisions must be robust, consistent and fair.

The principle of standard setting is to set the pass mark for an examination against a criterion-referenced standard by determining the minimum level of knowledge and/or skills required to pass an examination. There are several recognised standard setting processes used in high-stake medical examinations, and different examination formats lend themselves to different standard setting methods. The FRCA examinations use a mixture of test-centred and examinee-centred standard setting methods.

# STANDARD SETTING: International practice

roca.ac.uk/documents/standard-setting-primary-final-mcq-examinations



Introduction

Executive summary

Purpose of programme of assessment

How the programme of assessment is to be used with the curriculum

Critical Progression Points

Overarching assessment blueprint

RCoA Assessments

FRCA Examinations

The Primary FRCA examinations

The Final FRCA examinations

Linking curriculum content to the examinations – Blueprints and examination syllabus

Validity of the Examinations

Standard setting

**Standard setting: Primary and Final MCQ examinations**

Standard setting: Primary FRCA OSCE

Standard setting: Primary and Final FRCA SOE

Standard setting: Fairness

Standard setting: Quality assurance

FRCA Examinations: Feedback

Standard setting: Monitoring and Reviewing

Assessments in the workplace: Formative assessment

## Standard setting: Primary and Final MCQ examinations

The cut score for the Primary and Final MCQ examinations is established by the Angoff referencing method. In following best practice, a dedicated Angoff referencing group of examiners use the Angoff process to determine a cut score and make an adjustment of 1 Standard Error of Measurement (SEM) to arrive at the pass mark. Training is given to all members of the Final and Primary MCQ Angoff reference groups in the process, and to develop a collective understanding of the 'minimally competent' candidate, as defined below:

*"For the purposes of the (Primary or Final) MCQ examination, a 'minimally competent' candidate is one who has only just enough depth and breadth of knowledge stipulated within the (stages 1 and 2) curriculum to underpin their current clinical practice and equip them for the next phase of their anaesthetic training. If they pass the written examination but do not undertake further preparation and gain more understanding, they are likely to fail any subsequent SOE or OSCE on the same curriculum areas."*

In determining the 'minimally competent' candidate, members of the Angoff referencing groups are encouraged to use personal experience of anaesthetists in training sitting the examination at the particular stage of training.

After each examination, the examiner groups carry out an in-depth item analysis on items with unexpected performance statistics. Item deemed problematic, are removed from the paper before scores are finalised.

**Angoff method**

**SEM adjustment**

**Borderline standard**



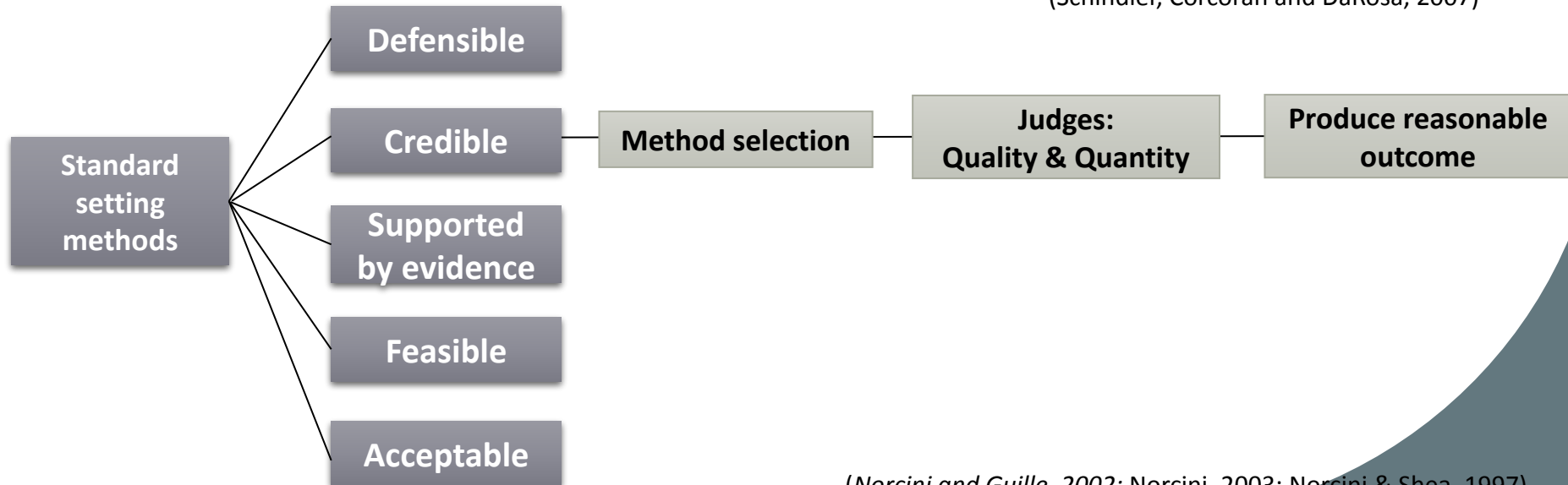
## STANDARD SETTING: Types of standard

Relative	Absolute	Compromise
Norm-referenced	Criterion-referenced	Combine both
“Top 60% will pass”	“Candidate who gets more than 60% pass”	
‘Limited seats’ - Admission	High stakes examination	

# Standard is arbitrary.

“..... even the most rigorous standard-setting method, followed meticulously, will be somewhat arbitrary however, they should be **credible**.”

(Schindler, Corcoran and DaRosa, 2007)



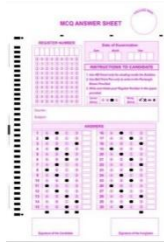
(Norcini and Guille, 2002; Norcini, 2003; Norcini & Shea, 1997)

## STANDARD SETTING: Method Selection

### Absolute

### Compromise

Test item based



**Angoff families**

Ebel

Nedelsky

Bookmark

Cohen

Test examinees based



Borderline group/  
regression  
Contrasting group

**Hofstee**

No gold standard.

But commonly used - Angoff (theory) & Borderline Regression (OSCE)

# STANDARD SETTING: Method Selection

**MCQs**

**Essays**

**Clinical exam /  
Viva**

**Portfolios**

Angoff family

Ebel

Nedelsky

Bookmark

Borderline group/  
regression

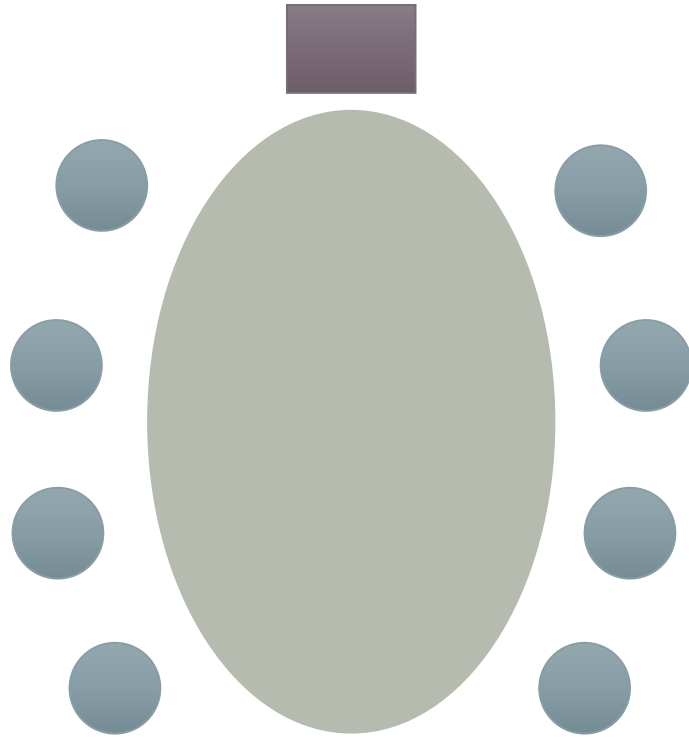
Contrasting group

Body of Work

Hofstee



## STANDARD SETTING: Judges



**Subject matter experts**

**Know target population**

**Understand task and assessment tool**

**Fair-minded**

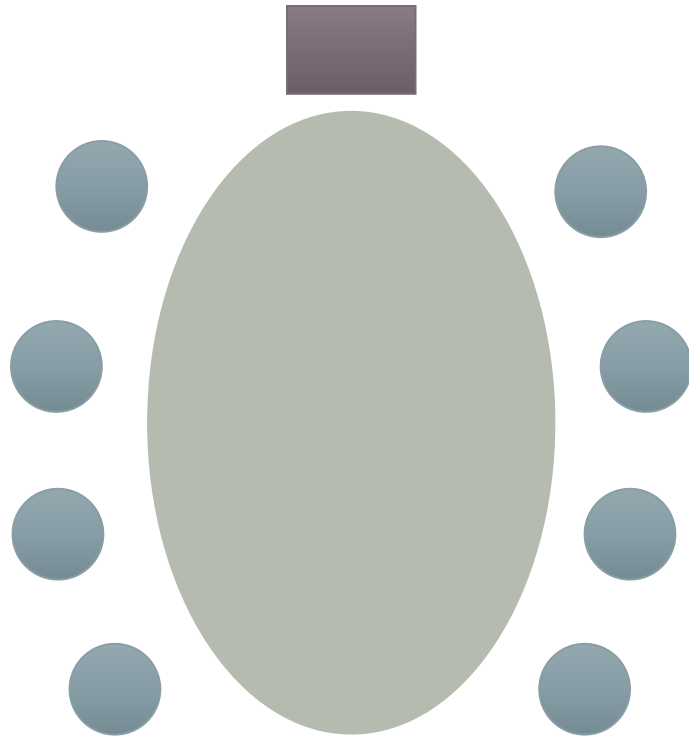
**Willing to follow directions**

**Give full attention to the process**

**Demographically diverse to avoid bias**

**6 considered minimum**

Norcini, J., & Guille, R. (2011). Combining tests and setting standards. In *International handbook of research in medical education* (pp. 811-834). Dordrecht: Springer Netherlands



SCREEN

Select the judges

Discuss

- a. Purpose of the assessment
- b. Nature of examinees
- c. Components of adequate/inadequate knowledge

Select the methods – train judges

Define borderline standard

## STANDARD SETTING: Nedelsky - DURING



Read through each question

Judges: Working individually, judges mark the wrong answers the borderline students would be able to eliminate.

**SCREEN**

Cizek, G. J. (2006). Standard setting. *Handbook of test development*, 225-258



## STANDARD SETTING: Angoff's families

### Yes/No Angoff

Panels make judgment whether a 'minimally acceptable candidate' can answer the question.

1 or 0

### Modified Angoff

Panels make judgment the probability of 'minimally acceptable candidate' can answer the question.

0, 10, 20, 30, 40, 50,  
60, 70, 80, 90, 100

### Modified Angoff

Panels make judgment what a 'minimally acceptable candidate' would score in the question.

Raw marks

## STANDARD SETTING: Modified Angoff - DURING



**Read through question 1**



**Judges: Individually, estimate mark that  
might be obtained by borderline examinee  
for question 1**

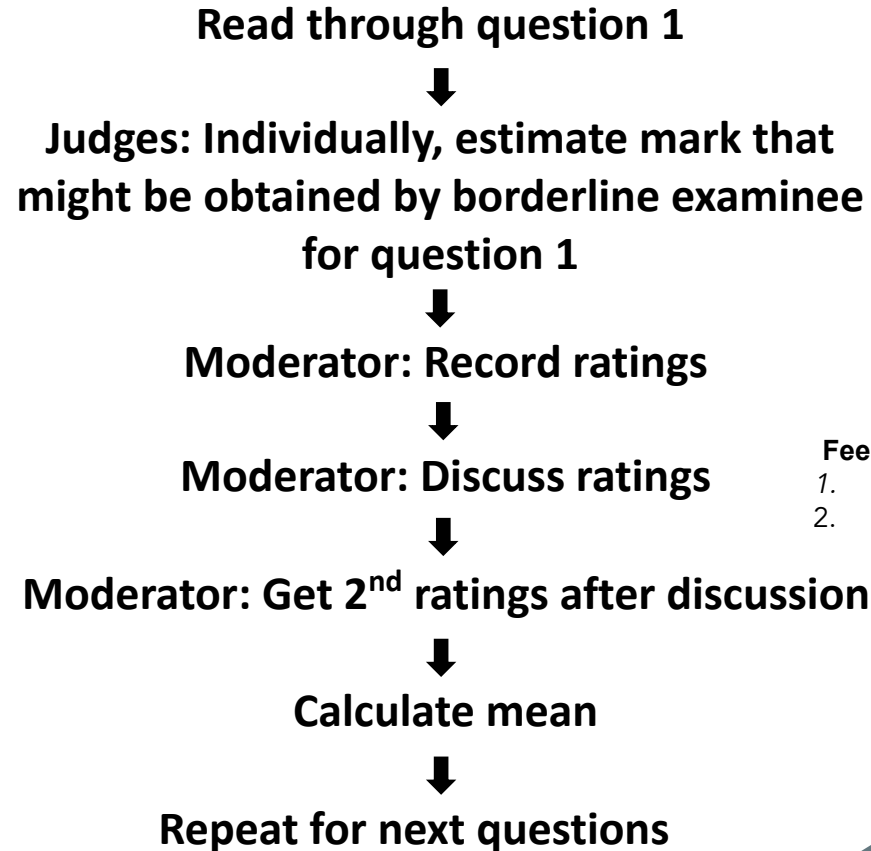
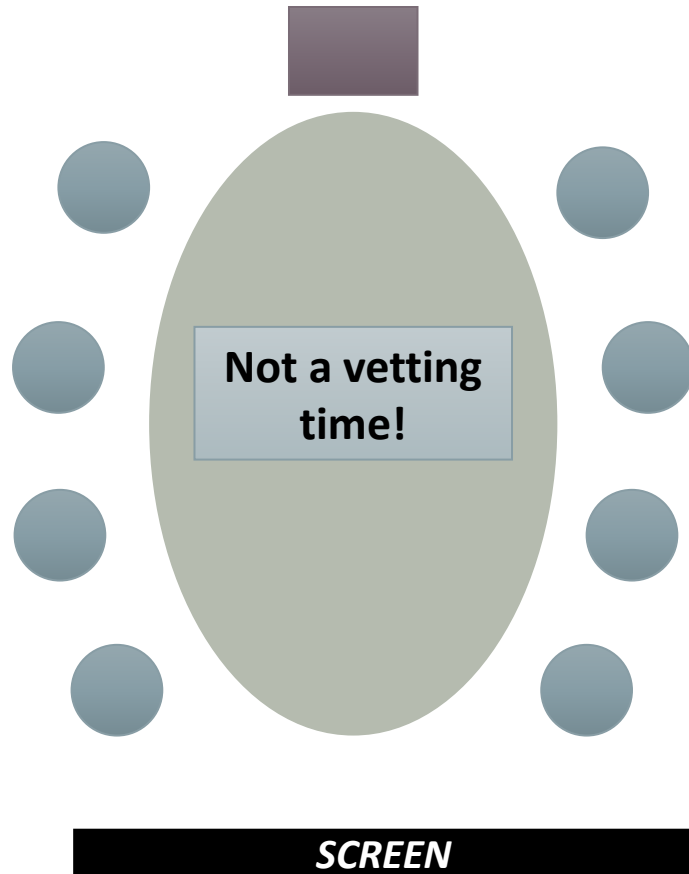


**Moderator: Record ratings**





# STANDARD SETTING: Modified Angoff - DURING



- Feedback**
1. Normative
  2. Reality data



1. A 50-year-old lady presented with greenish discharge from the left nipple. She is noted to have slit-like nipple retraction. There is no palpable mass. Mammogram showed coarse calcifications in bilateral breasts.

Which of the following is the most likely diagnosis?

A	Ductal ectasia
B	Ductal papilloma
C	Infiltration carcinoma
D	Sclerosing adenosis

**Possible rating  
0-100  
(multiple of 10)**

[illegible]

- |    |  |
|----|--|
| 1. | A 50-year-old lady presented with greenish discharge from the left nipple. She is noted to have slit-like nipple retraction. There is no palpable mass. Mammogram showed coarse calcifications in bilateral breasts.<br><br>Which of the following is the most likely diagnosis? |
|----|--|

A	Ductal ectasia
B	Ductal papilloma
C	Infiltration carcinoma
D	Sclerosing adenosis

**Possible rating  
0-100  
(multiple of 10)**

[illegible]

S8Q MEDICAL BASED #3 WITH ANSWER

3.

On examination, she was conscious, afebrile, with PR 60 beats/min, regular rhythm and BP 110/70 mmHg. Her speech was broken into separate syllables, often separated by pauses and spoken with varying volume.

Cerebellar examination: Truncal and appendicular instability.  
Repetitive involuntary oscillation of the eyes.

Funduscopy: Bilateral papilloedema.

a. List THREE (3) other expected neurological signs (3 marks)

- Dysdiadochokinesia

- Intention Tremor

- Pass pointing of dysmetria (finger-nose, heel-shin)

- Inability to walk in straight line/tandem gait (broad based gait)

- Rebound phenomenon

b. List **TWO (2)** possible sites of lesion (2 marks)

- Cerebellar vermis / Midbrain lesion/ 4th ventricular tumour  
Cerebellum (1/2) /Posterior fossa (1/2)

c. State the most likely cause for the clinical presentation (1+1 mark)

- Obstructive Hydrocephalus secondary cerebellar tumour! Increased ICP secondary to Cerebellar tumour! Space occupying lesion! Cerebellar lesion

d. List **TWO (2)** investigations to assist the diagnosis ( 1 mark)

#### CT brain imaging

**MRI brain**

Panels give mark based on how many mark can ONE borderline standard obtain from the answer scheme







**Read through question 1**



**Judges: Individually, estimate mark that might be obtained by borderline examinee for question 1**



**Moderator: Record ratings**

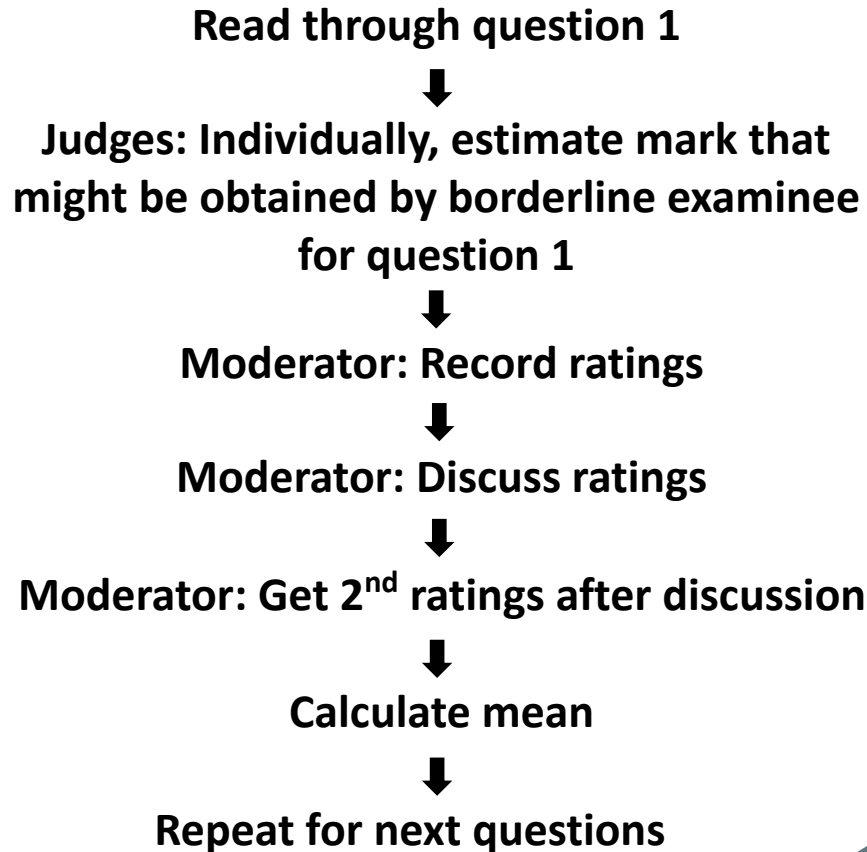
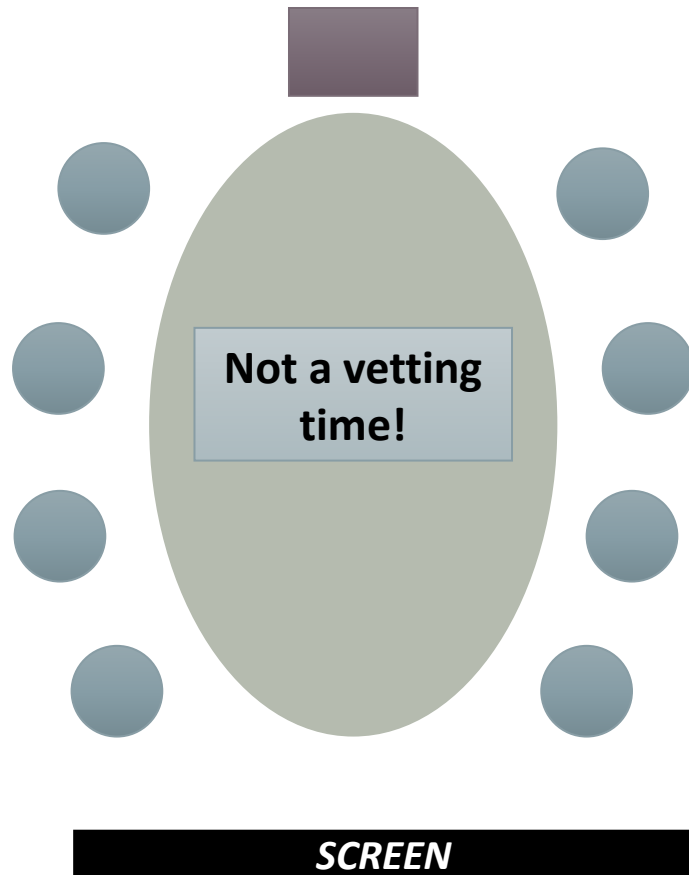


**Moderator: Discuss ratings**

1.  $SD > 10\%$  of score
2.  $SD < 10\%$  but panels want to discuss

**SCREEN**

## STANDARD SETTING: Modified Angoff - DURING



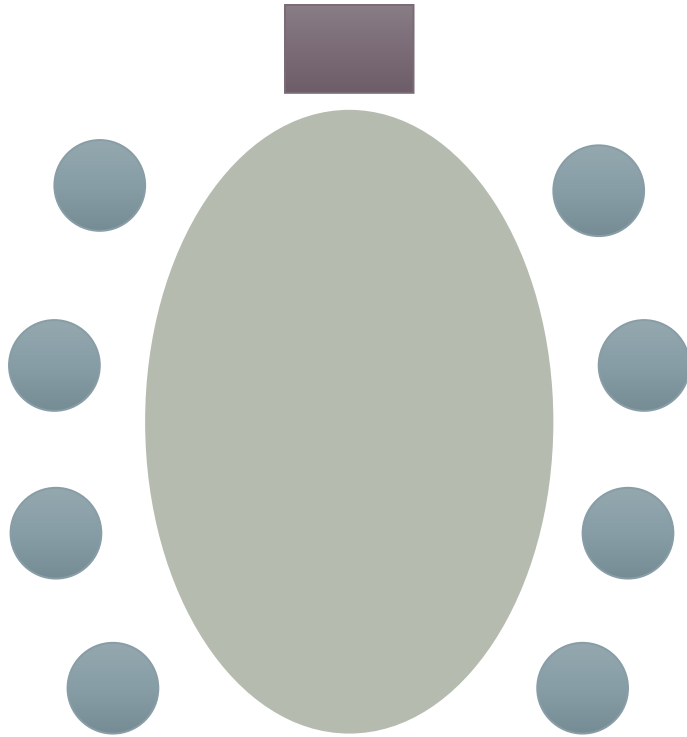
# Selecting cut off scores

	<i>Mean</i>	<i>SD</i>
<b>ROUND 1</b>	62.00	0.55
<b>ROUND 2</b>	60.67	0.51

## PASSING SCORE

1. **2<sup>nd</sup> round mean**
2. Higher mean
3. Ones with lower SD
4. Average the two means

## STANDARD SETTING: Modified Angoff - POST



***SCREEN***

- Evaluate the process**
- **Judges confidence in the process**
  - **Resulting cut off scores**

**Documentation**

Cizek, G. J. (2006). Standard setting. *Handbook of test development*, 225-258



DEFINING A BORDERLINE STANDARD

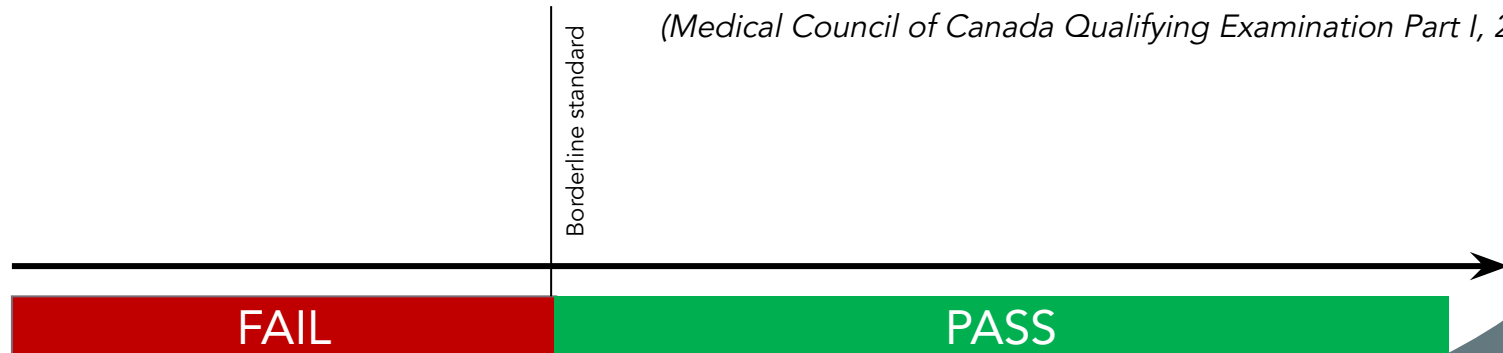
## STANDARD SETTING: Borderline standards

Most methods rely on a borderline standard to help raters arrive at a passing score.

Some also define this as *minimally competent candidate*.

The “minimally competent” candidate entering supervised practice has just enough knowledge and skills to provide safe and effective patient care, no more, no less.

*(Medical Council of Canada Qualifying Examination Part I, 2018)*



# STANDARD SETTING: Guides to define Borderline

**# 2 Knowledge** – e.g. demonstrate adequate knowledge for safe clinical judgment, decision making and management



**#4 Soft skills** - e.g. conduct themselves professionally



**#1 Setting** – e.g. graduate of the ophthalmology program



**#3 Skills** – e.g. be able to work with moderate supervision, equipped with acceptable technical ability



**#5 Errors** (considering the forgivable or unforgivable) – e.g. safe clinical judgment, decision making and management



# Australian Board Examinations (General Practice)

- Sturmberg & Hinchy, 2010

**Table 1** Characteristics of the borderline candidate

---

Judges consistently describe borderline candidates'

- Knowledge as:
    - Superficial, possibly some isolated areas of depth
    - Fact-oriented, tends to be concrete
    - Not systematically organized; scattered or disorganized
  - Problem-solving skills as:
    - Poor pattern recognition
    - Difficulty prioritizing
    - Rigid style
    - Poor awareness of own limitations
  - Communication/social skills as:
    - Often poor communication with patients/colleagues
    - Inability to fit illness into social context
-

## STANDARD SETTING: Borderline Standards (Undergraduate)

*Setting*

*Knowledge*

*Skills*

*Attitude*

*Errors*

*Forgivable, non-forgivable*

"The borderline passing graduate of MBBS IIUM should demonstrate adequate fundamental knowledge, safe clinical judgement and decision-making ability, able to work with supervision, has effective communication and upholding professionalism and ethical values incorporating Islamic values."

**FAIL**

**PASS**

"The borderline graduate of the pediatrics program should demonstrate adequate knowledge for safe clinical judgment, decision making and management, be able to work with moderate supervision, has an acceptable communication skills, and conduct themselves professionally."

(Standard Setting Workshop, Conjoint Pediatric, 18 January 2018, USM)

"A borderline student of Radiology MMED program demonstrates basic knowledge for safe clinical decision and management, be able to work under minimal supervision, be equipped with basic radiological skills, and conduct themselves professionally."

(Radiology Conjoint, 9 April 2019, Standard Setting Workshop)

## Variation of borderline standards

"The borderline graduate of the emergency medicine program should demonstrate adequate knowledge for safe clinical judgment, decision making and management, be able to work with moderate supervision, equipped with acceptable life saving skills and technical ability, and conduct themselves professionally."

(SCCEM, 10 Nov 2018, A Workshop on Standard Setting A & E Workshop, UM, Kuala Lumpur)

"The borderline graduate of the ophthalmology program should demonstrate adequate knowledge for safe clinical judgment, decision making and management, be able to work with moderate supervision, equipped with acceptable technical ability, and conduct themselves professionally."

(MUCCO, 20-22 Aug 2014, A Workshop on Examination Questions Preparation, Kuala Lumpur)

# Borderline candidates

“The graduate of USM MD program should demonstrates adequate knowledge for safe clinical decision and management, be able to work under supervision, be equipped with standard clinical skills, and conduct themselves professionally.”

Component	Fail	Borderline	Pass
Clinical decision and management		Having knowledge but superficial Able to detect life threatening and emergency condition Unable to translate theory to practical Approach not systematic  Rigid problem solving Text-book oriented	
Supervision		1 to 1 supervision Not able to perform on their own without supervision Prone to make minor mistakes Need repeated sessions of training Willing to learn and still have insights Aware of own limitation and need guidance  Require repeated close supervision	

# Borderline candidates

“The graduate of USM MD program should demonstrates adequate knowledge for safe clinical decision and management, be able to work under supervision, be equipped with standard clinical skills, and conduct themselves professionally.”

Component	Fail	Borderline	Pass
Clinical skills	Unable to recognize the majority of clinical findings	Able to recognize majority (and severity) of clinical findings but unable to formulate complex diagnosis	Able to recognize majority (and severity) of clinical findings and formulate diagnosis
Professionalism	Fail to show commitment, respect, accountability.	Just adequate commitment, responsibility, accountability and respect Able to convey minimal correct message to colleague Show some empathy	Very committed, respectful, accountable, responsible with patients, colleagues, staffs and supervisors

# Borderline candidates - Feb 2023

“The borderline graduate of the pediatrics program should demonstrate adequate knowledge for safe clinical judgment, decision making and management, be able to work with moderate supervision, has an acceptable communication skills, and conduct themselves professionally.”

Element	Pass	Borderline	Fail
Content mastery (Must know - Should know - Nice to Know)		<ul style="list-style-type: none"><li>• Know key facts or common conditions (must know such as asthma /DM/RDS/IDM)</li><li>• Know common presentation and its management - vomiting</li><li>• Difficulty in recognizing/explaining uncommon conditions</li></ul>	
Clinical judgment		<ul style="list-style-type: none"><li>• Able to detect sick patients</li><li>• Knows normal values for vital signs</li><li>• Appropriate decision in clear cut or obvious severe condition</li><li>• May not able to decide well in unclear parameters</li><li>• Unable to prioritize in complex / chronic illnesses</li></ul>	
Problem solving		<ul style="list-style-type: none"><li>• Able to list out the common DDx</li></ul>	
Management		<ul style="list-style-type: none"><li>• Knows acute/preliminary steps in emergency cases</li><li>• More supervision in long term Mx / complex problem</li></ul>	

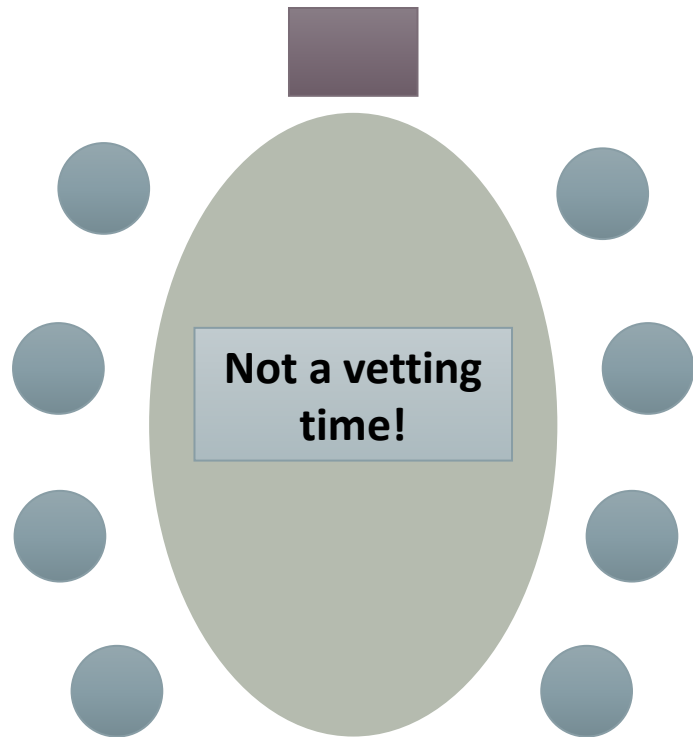
1.	A 50-year-old lady presented with greenish discharge from the left nipple. She is noted to have slit-like nipple retraction. There is no palpable mass. Mammogram showed coarse calcifications in bilateral breasts.	
	Which of the following is the most likely diagnosis?	
	A	Ductal ectasia
	B	Ductal papilloma
	C	Infiltration carcinoma
	D	Sclerosing adenosis

#### Possible rating

- Can eliminate all 3 distractors – 1
- Can eliminate all 2 distractors – 0.5
- Can eliminate all 1 distractors – 0.33
- Cannot eliminate any distractor – 0.25

	PANEL 1	PANEL 2	PANEL 3	PANEL 4	PANEL 5	PANEL 6	Mean	SD
Qs 1	0.5	0.5	1	0.25	0.25	0.5	0.50	0.27
Qs 2								
Qs 3								
Qs n								
Mean								

## STANDARD SETTING: Nedelsky - DURING



**SCREEN**

Read through each question

Judges: Working individually, judges mark the wrong answers the borderline students would be able to eliminate.

Moderator: Record ratings

Repeat for next questions

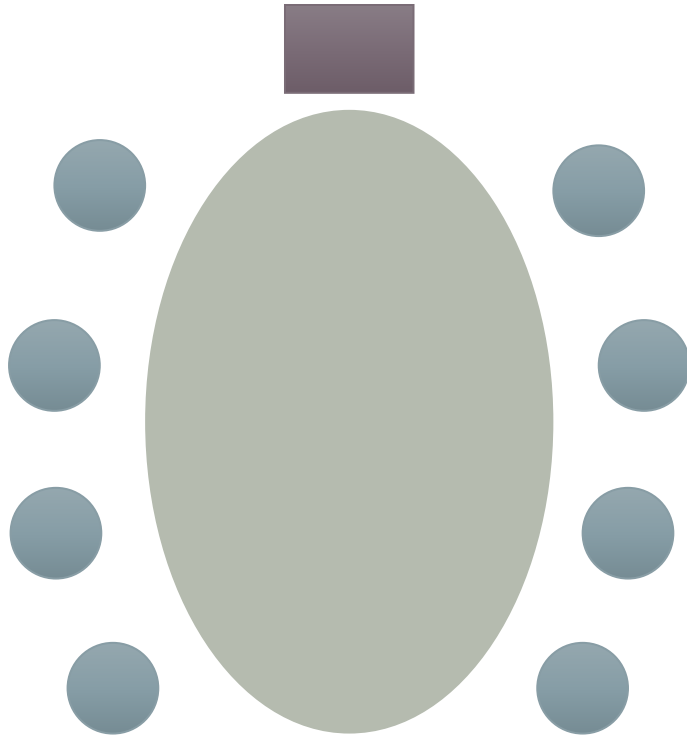
Calculate passing score (Average of the probabilities across all questions)



**Table 1.** Example of calculations for Nedelsky's method applied to a test scored without correction for guessing

Question	Answers *	Number of answers not eliminated	Expected score
1	A (B) <del>C</del> <del>D</del> <del>E</del>	2	1/2 = .50
2	<del>A</del> <del>B</del> <del>C</del> <del>D</del> (E)	1	1/1 = 1.00
3	<del>A</del> <del>B</del> C (D) <del>E</del>	2	1/2 = .50
4	A <del>B</del> C (D) <del>E</del>	3	1/3 = .33
5	(A) <del>B</del> <del>C</del> <del>D</del> <del>E</del>	1	1/1 = 1.00
6	A B (C) D E	5	1/5 = .20
7	A B C <del>D</del> (E)	4	1/4 = .25
8	(A) B <del>C</del> D E	4	1/4 = .25
9	A (B) C D E	5	1/5 = .20
10	A (B) C D E	5	1/5 = .20
Cut-off score		Sum = 4.43	
Expected total score = 4.43			

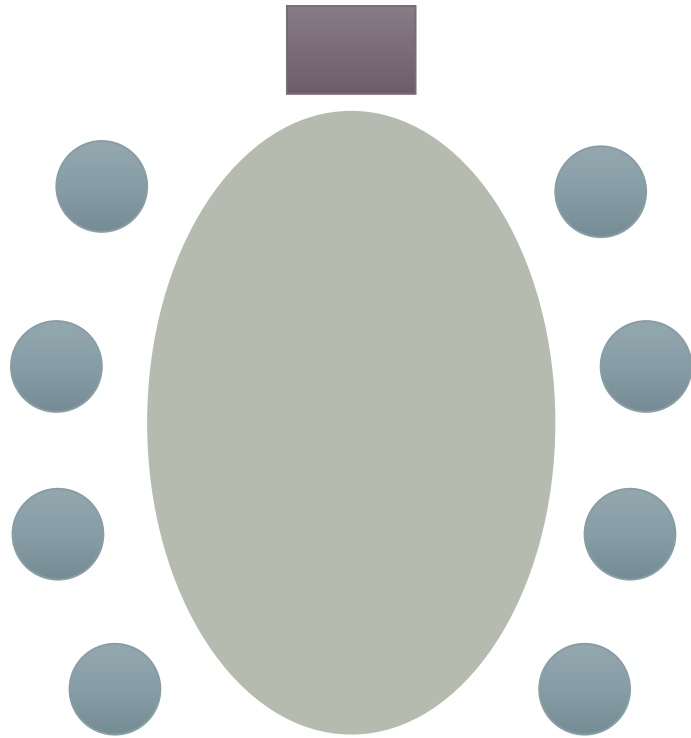
\* A circle indicates the correct answer; an X indicates an answer the borderline test-taker would eliminate.



SCREEN

- Evaluate the process
- Judges confidence in the process
  - Resulting cut off scores

Documentation



Select the judges

Discuss

- a. Purpose of the assessment
- b. Nature of examinees
- c. Components of adequate/inadequate knowledge

Select the methods – train judges

Review the test in general

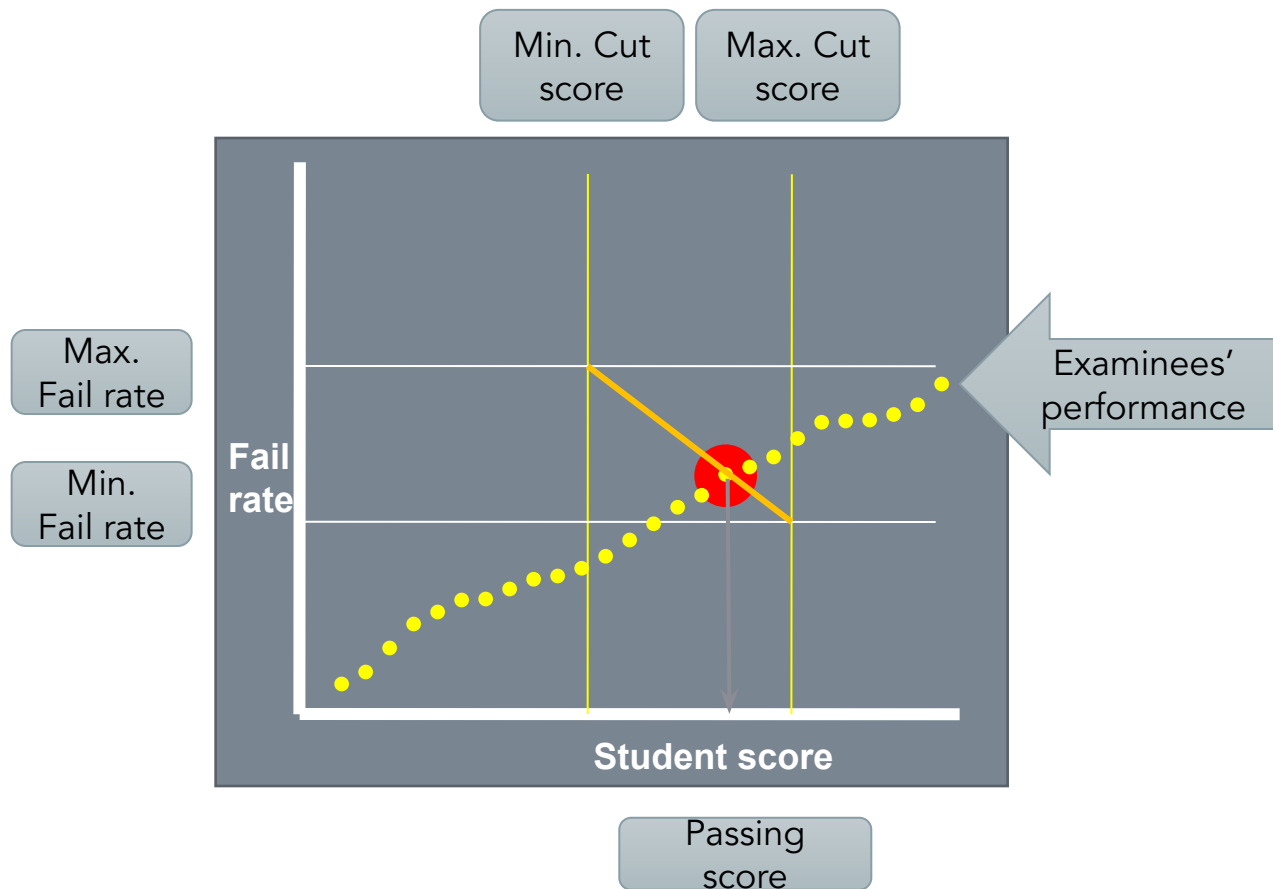
*SCREEN*



Ask the judges to answer 4 questions:

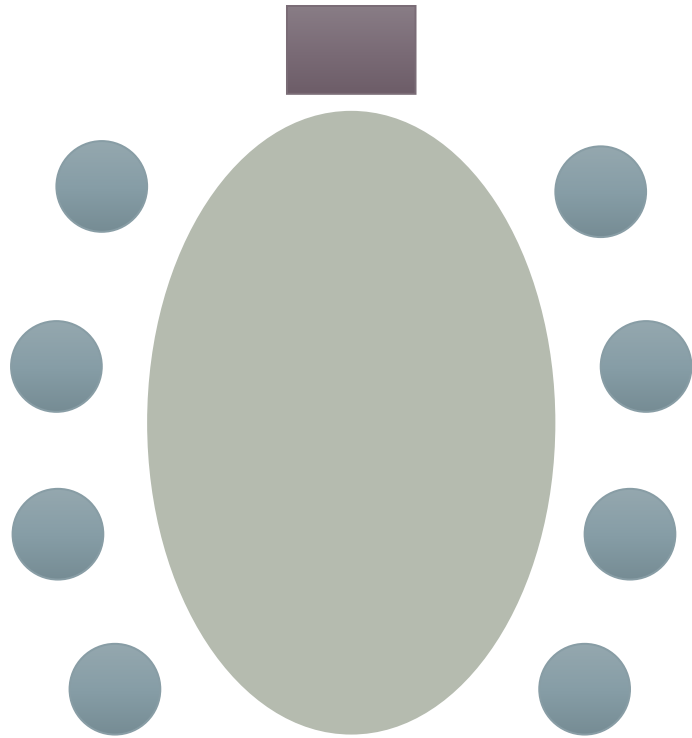
- ✓ What is the minimum acceptable cut score?
- ✓ What is the maximum acceptable cut score?
- ✓ What is the minimum acceptable fail rate?
- ✓ What is the maximum acceptable fail rate?

After the test is given, graph the distribution of scores and select the cut score.



## STANDARD SETTING: Types of standard

Relative	Absolute	Compromise
Norm-referenced	Criterion-referenced	Combine both
“Top 60% will pass”	“Candidate who gets more than 60% pass”	
‘Limited seats’ - Admission	High stakes examination	



- Evaluate the process
- Judges confidence in the process
  - Resulting cut off scores

Documentation

### Advantages

- Easy to implement
- Educators are comfortable with the decision



### 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.



# Questions number

## *SAMPLING MIXTURE?*

Relevance	Difficulty		
	Easy	Medium	Hard
Essential	4 questions 95% correct	3 questions 85% correct	1 question 80% correct
Important	3 questions 90% correct	3 questions 75% correct	2 questions 60% correct
Acceptable	1 question 80% correct	2 questions 55% correct	2 questions 35% correct
Questionable	1 question 50% correct	0 questions	2 questions 20% correct

EBEL METHOD – Based on item relevance and difficulty  
(but less used as compared to Angoff's)

# Hands-on links

Group 1 (BMS)	<a href="https://tinyurl.com/komgroup1">https://tinyurl.com/komgroup1</a>
Group 2 (Medical based)	<a href="https://tinyurl.com/komgroup2">https://tinyurl.com/komgroup2</a>
Group 3 (Medical based)	<a href="https://tinyurl.com/komgroup3">https://tinyurl.com/komgroup3</a>
Group 4 (Surgical based)	<a href="https://tinyurl.com/komgroup4">https://tinyurl.com/komgroup4</a>
Group 5 (Surgical based)	<a href="https://tinyurl.com/komgroup5">https://tinyurl.com/komgroup5</a>

## STANDARD SETTING: Post exercise

Compensatory stations: Calculate together  
Rounding?

48      48.4      49

False positive  
False negative  
Cost

1. Calculate cut off  
score

2. Feedback from  
examiners

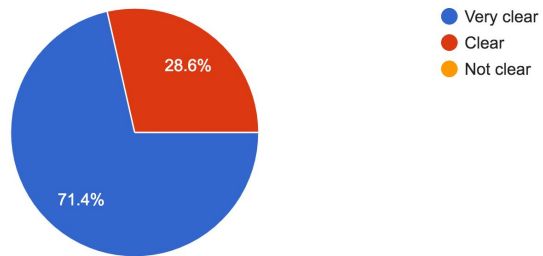
Feedback on

1. Clarity of task
2. Time needed
3. Confidence in cut off score



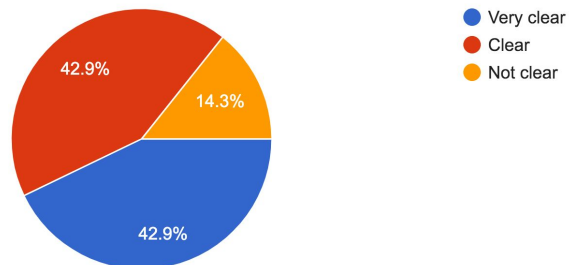
How clear is the rating task to be performed?

7 responses



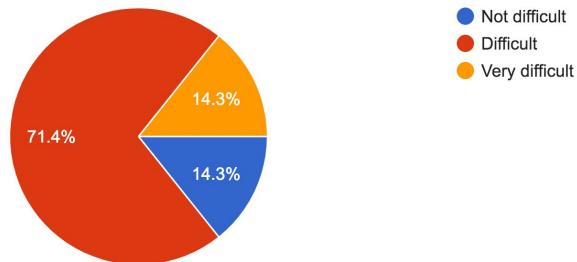
How clear are the characteristics of a borderline examinee?

7 responses



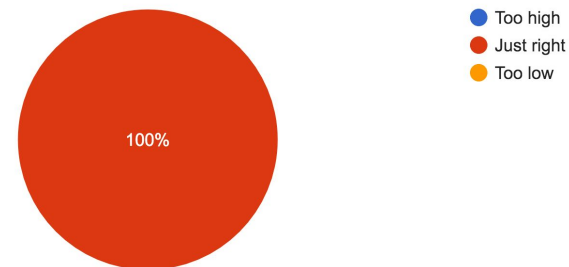
How difficult was it to provide ratings?

7 responses



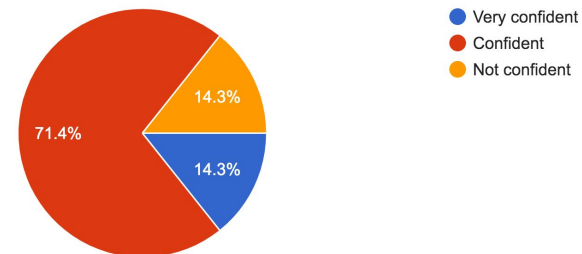
Do you think the final passing scores are appropriate for the examinees?

7 responses



How confident are you in the appropriateness of the cut scores?

7 responses



## STANDARD SETTING: Post exercise

Compensatory stations: Calculate together  
Rounding?

48      48.4      49

False positive  
False negative  
Cost

1. Calculate cut off score

2. Feedback from examiners

Feedback on

1. Clarity of task
2. Time needed
3. Confidence in cut off score



Cut off score is policy making decision  
Decision makers may

1. Accept
2. Suggest to add or minus SEM
3. Suggest second standard setting
4. Stick to 50%

4. Presentation to decision makers

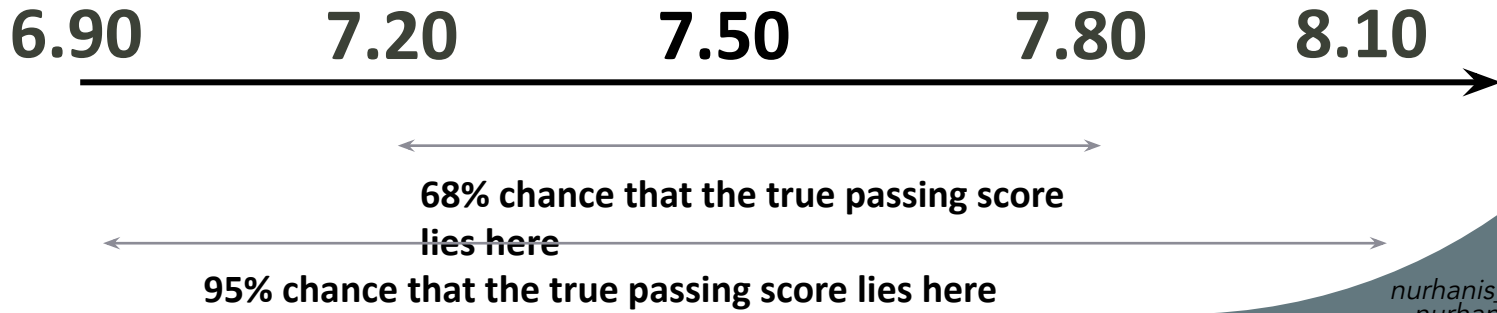
3. Standardize cut off score

Converting the cut off score to 50%  
More palatable to score users

## TRUE CUT OFF SCORE

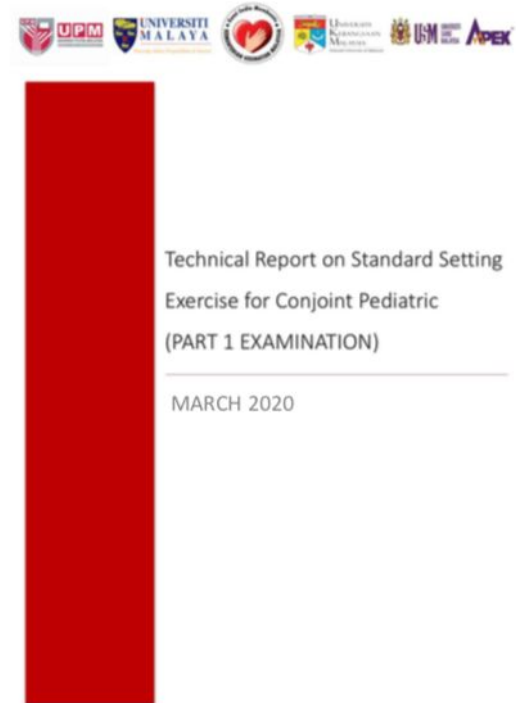
- In assessment, the main variance influencing the test score should be the ability of the candidates in the subject tested.
- But there is a possibility that the obtained score is actually less or more than the candidates should have actually obtained - due to error (*Observed score = True score + Error*)
- To estimate the true score, we can calculate Standard Error of Measurement (SEM)  
= Standard Deviation  $\times \sqrt{(1-\text{reliability})}$

$$SEM = 0.30$$



1. Compare with historical standards or external measure
2. Consider stakeholder opinion
3. Reasonable with competence markers

(Norcini, 2003)



Assessment utility =  
Validity X Reliability X Educational Impact X Acceptability X Cost

Low stake assessment

Validity X Reliability X Educational Impact X Acceptability X Cost

High stake assessment

Validity X Reliability X Educational Impact X Acceptability X Cost



# Thank you for your time

Nurhanis Syazni Roslan (nurhanis\_syazni@usm.my)

Department of Medical Education, USM  
Examination Unit, USM

