Advancing Best Care through Education & Research
6th Asia Pacific Primary Care Research Conference
in conjunction with Family Medicine Symposium

Using psychometrics to improve assessment in medical education

Gominda Ponnamperuma
Ruth Lim

Organised by: SingHealth DukeNUS Academic Medical Centre FAMILY MEDICINE
Supported by: AFPM PRIME

What is psychometrics?

Psychometrics is the field of study concerned with the theory and technique of educational and psychological measurement, including measurement of knowledge, abilities, attitudes, and personality traits.

I. **construction of instruments and procedures for measurement**

II. **development and refinement of theoretical approaches to measurement.**

What is psychometrics?

1. Reliability

2. Validity

1. Reliability *contd...*

- Are the exam results the same when repeated?

  Yes - 100% reliable test

But, in real life – things are not so straight forward

- How close the exam results are when repeated?

  - 80% reliable test

Reliability = reproducibility

1. Reliability *contd...*

1. Reliability *contd...*

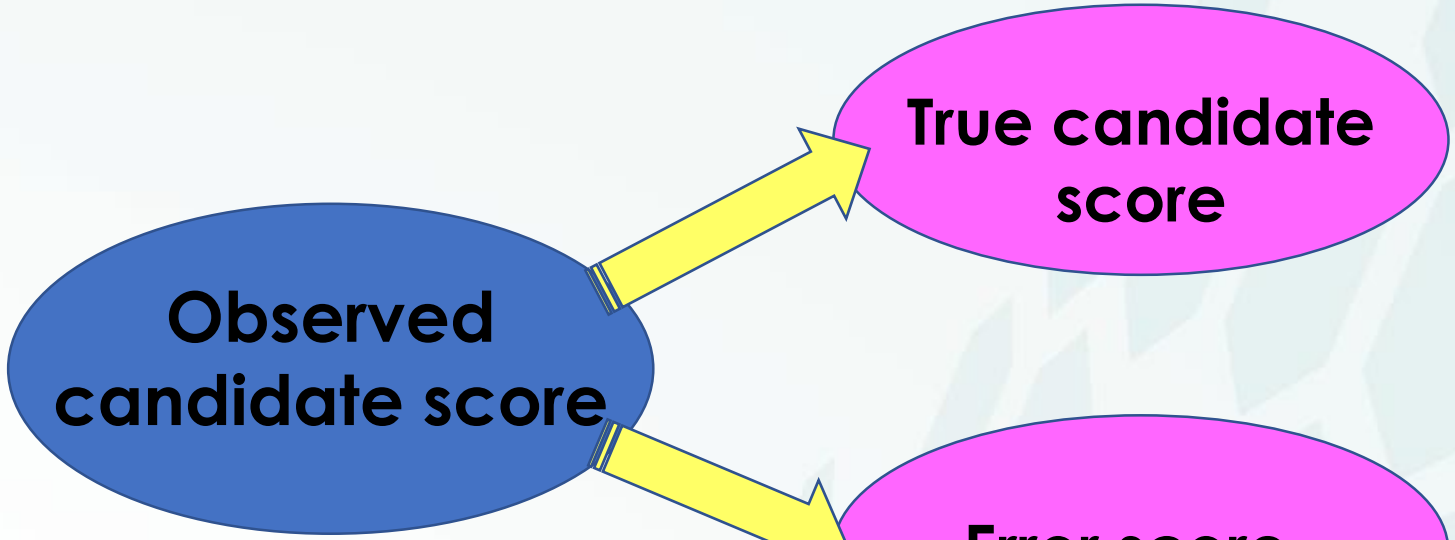
Very reliable

1. Reliability *contd...*

Not reliable

1. Reliability *contd...*

Reliability as a yardstick of measurement error



- test items
- examiners
- others; e.g. patients
- acts of God
- interaction among any

1. Reliability *contd...*

Observed score = True score + Error

$$\text{Reliability} \propto \frac{\text{Variability of true scores}}{\text{Variability of true scores} + \text{Variability of error}}$$

1. Reliability *contd...*

$$\text{Reliability} = \frac{\text{True score variance}}{\text{True score variance} + \text{Error variance}} \times \frac{n}{n-1}$$

1. Reliability *contd...*

How to interpret reliability?

Value	Interpretation
0 – 0.7	Poor
0.7 – 0.8	Good
0.8 – 0.9	Very good
0.9 – 1.0	Excellent

If-item-deleted reliability < reliability of entire test

2. Validity

Does the exam assess what it should assess?

e.g. Assessment of clinical skills

Short answer question

Write short notes on 'urethral catheterisation'.

How can we make the test material valid?

OSCE Station

Carry out 'urethral catheterisation' on the simulator (dummy). Carry out all steps thinking that you are attending to a real patient. Assume that patient consent has been obtained.

2. Validity *contd...*

Reliable, but not valid

2. Validity *contd...*

Not reliable; not valid

2. Validity *contd...*

1. Face validity

Are the stakeholders happy with the exam?

2. Content validity

Is the exam assessing the right content?

3. Criterion-oriented validity

(a) Concurrent validity

(b) Predictive validity

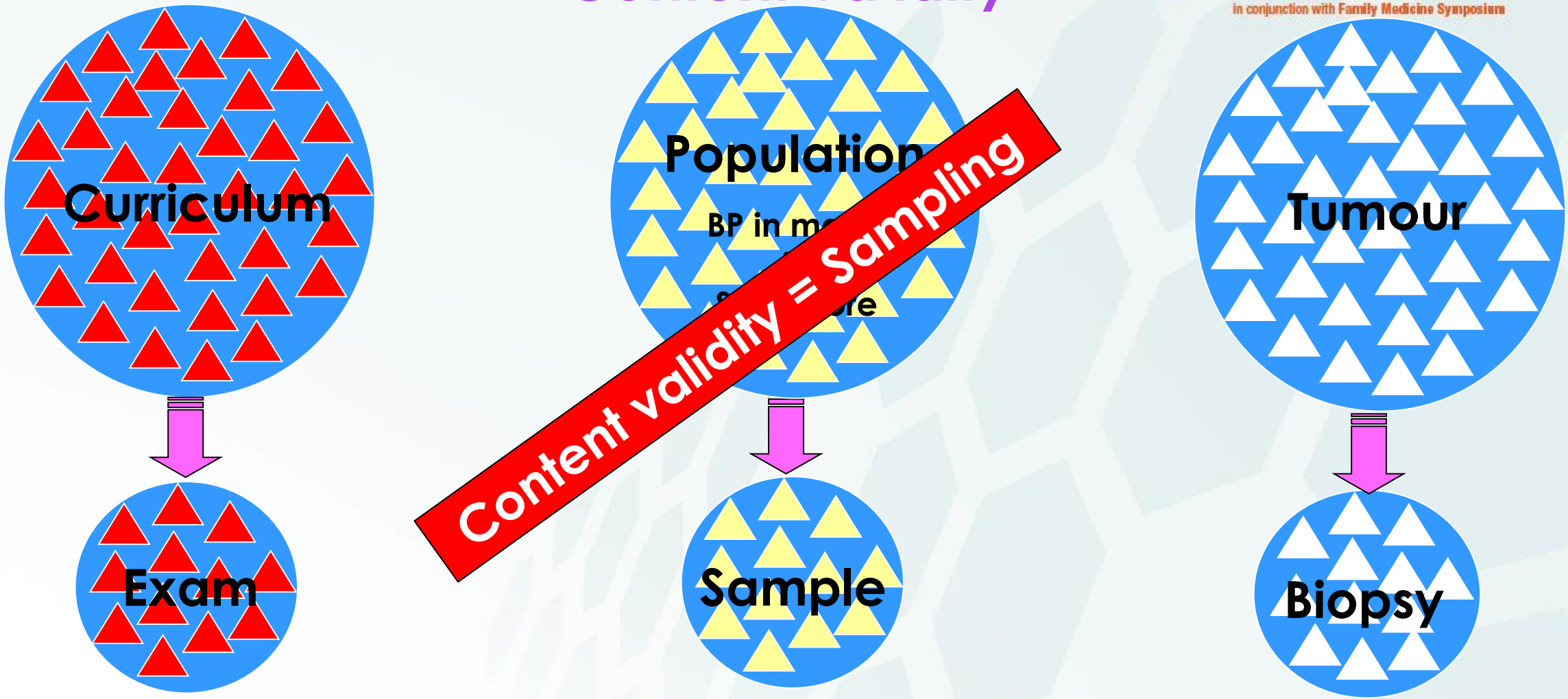
How is the exam, compared to another exam?

4. Construct validity

Does the exam assess a latent trait/characteristic?

2. Validity *contd...*

Content validity



2. Validity *contd...*

Assessment blueprint

	Good clinical care	Communication skills	Maintaining good medical practice	Working with colleagues	Teaching and training	Basic sciences	Decision making	Research	Professional development
Trauma	✓						✓		
Shoulder & elbow			✓						✓
Hand			✓		✓				
Spine				✓					
Hip						✓			
Knee		✓						✓	
Ankle & foot		✓							
Paediatric orthopaedics							✓		✓

2. Validity *contd...*

Validity indices using item analysis

Difficulty (facility) index [*p* value] = $\frac{\text{No. of candidates correctly answering}}{\text{Total no. of candidates}}$

Discrimination index = **Proportion of candidates correctly answering in the upper 1/3 of the batch** − **Proportion of candidates correctly answering in the lower 1/3 of the batch**

2. Validity *contd...*

Interpreting item analysis indices

Difficulty index

Value	Interpretation
0 – 0.25	Very easy
0.25 – 0.75	Easy
0.75 – 1.0	Very easy

Discrimination index

Value	Interpretation
Negative	Unacceptable
0 – 0.1	Poor
0.1 – 0.3	Good
> 0.3	Very good

Standard setting

Standard	≈	Measure of quality
Standard setting	≈	Determining the level of quality

Standard setting of an assessment is determining the pass mark.

Standard setting *contd...*

Standard setting methods

- **Criterion-referenced**
  - **Test-centred**
    - Angoff
    - Ebel
    - Nedelsky
    - Jaeger
    - Bookmark
  - **Examinee-centred**
    - Borderline group
    - Contrasting
- **Norm-referenced**
- **Compromise**
  - Hofstee

Conjunctive or Compensatory

Standard setting

Angoff method

Examiners (judges) discuss the characteristics of a minimally competent person (the just-passing student)

Each judge is asked to determine the probability of a minimally competent person (the just-passing student) answering each item correctly

Standard setting *contd...*

Angoff method *contd...*

Question no.	Probability of a borderline candidate passing										Total probability	Mean per examiner
	Ex1	Ex2	Ex3	Ex4	Ex5	Ex6	Ex7	Ex8	Ex9	Ex10		
Q 1	0.20	0.65	0.51	0.70	0.40	0.72	0.32	0.56	0.62	0.55	5.23	0.52
Q 2	0.15	0.58	0.45	0.75	0.38	0.75	0.35	0.54	0.65	0.53	5.13	0.51
Q 3	0.18	0.59	0.48	0.77	0.45	0.69	0.40	0.51	0.59	0.58	5.24	0.52
Q 4	0.25	0.63	0.55	0.80	0.48	0.78	0.39	0.58	0.68	0.49	5.63	0.56
Q 5	0.19	0.66	0.52	0.79	0.38	0.82	0.33	0.60	0.66	0.48	5.43	0.54
Total pass mark of all five questions (sum total of mean probabilities)												**2.65**
Mean pass mark; i.e. the pass/fail standard												**0.53**
Pass/fail standard as a percentage cut score												**53%**

Standard setting *contd...*

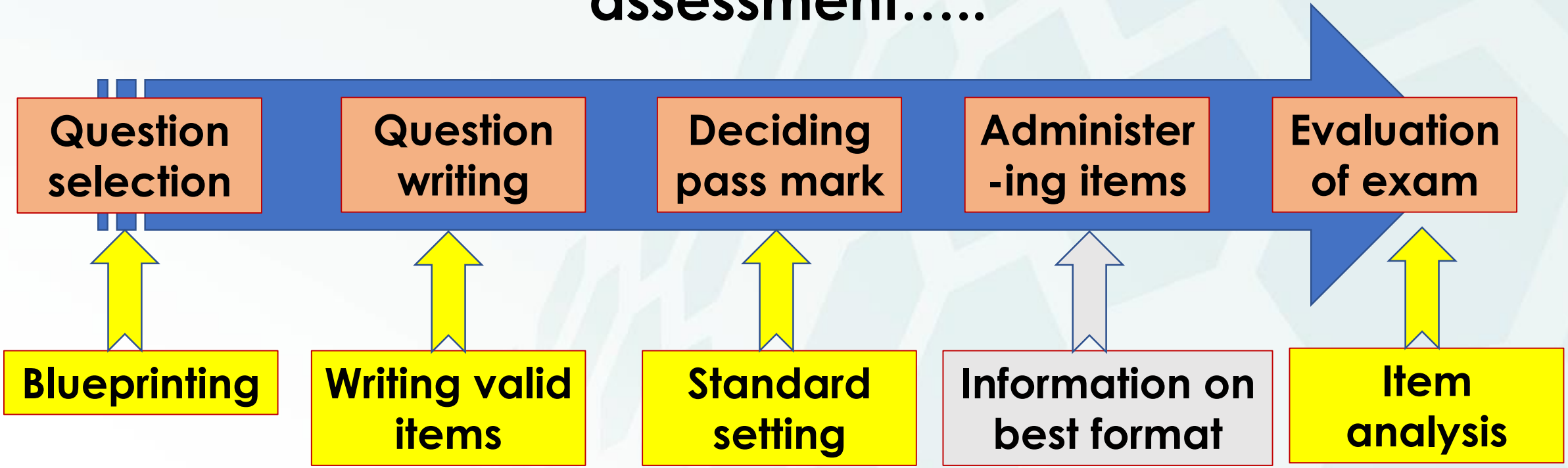
Angoff method *contd...*

- ✓ If 10 experts involved, each produces a cut score and the mean is calculated
- ✓ The modification involves the provision of actual exam data; ie, how a group of students actually performed in the test
- ✓ The experts revise their decisions if necessary based on the actual test data

Summary

What have we learnt?

How psychometrics can help us in the journey of an assessment…..



Summary

What have we learnt *contd...*?

- ✓ Reliability is reproducibility of results, but actually a measure of assessment error
- ✓ Validity is finding out what we want to measure
- ✓ Blueprinting ensures content validity – a candidate should be tested on each part of the curriculum
- ✓ Item analysis (difficulty, discrimination and alpha-if-deleted indices) inform us the quality of each assessment item; e.g. question
- ✓ Standard setting using a criterion referenced method such as modified Angoff will ensure that the passing candidate is fit-for-practice

Thank you