Research & Scholarship Programme · BIMHSE · April 2026

Are Your Tests Measuring Up?

Uncovering the Fundamentals of Item Analysis for Classroom Assessments — Part II

PRESENTER

Prof Fraide Ganotice Jr., PhD

Director, Associate Professor · BIMHSE, HKUMed

PRESENTER

Mr John Ian Wilzon Dizon, MSc

Research Assistant · BIMHSE, HKUMed

AFTER THIS WORKSHOP, WE HOPE TO ATTAIN THESE

Learning Outcomes.

01

Analyze item analysis

Evaluate the crucial aspects that determine whether a test question is doing its job.

02

Assess item effectiveness

Judge the overall quality and performance of any individual test item.

03

Make informed decisions

Apply evidence to keep, revise, or retire items in your item bank.

THE FOUNDATION

What is item analysis?

A systematic evaluation of test items across four key dimensions — to ensure every question earns its place.

DIFFICULTY

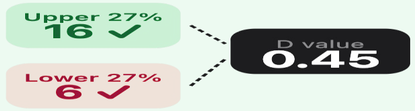
How hard is the item?

Proportion who answered correctly. Aim for P = 0.50–0.70.

DISCRIMINATION

Does it separate high scorers from low?

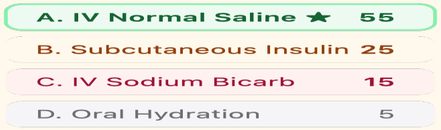
Top 27% vs. bottom 27% of performers. D ≥ 0.40 is excellent.



DISTRACTORS

Are the wrong answers working?

Good distractors attract students with gaps in knowledge.



RELIABILITY

Does it measure consistently?

Point-biserial correlation: item-to-total score relationship.

Note: *This workshop covers Classical Test Theory (CTT) — practical for classroom-level review. IRT offers additional advantages for large-scale assessments but requires larger samples and specialized software.*

REFERENCE

de Ayala, R. J. (2009). *The theory and practice of item response theory.* Guilford Press.

WHY IT MATTERS

Why Item Analysis Matters in Medical Education.

Medical decisions affect patient lives. That raises the stakes for assessment considerably.

REASON 01

Quality Assurance

Licensing exams, board certifications, clinical competency assessments — item analysis is our quality control mechanism.

REASON 02

Educational Impact

Test questions don't just measure learning — they actively shape it. Poor MCQs drive memorization over understanding.

REASON 03

Resource Optimization

Identify which items are worth keeping, which need revision, and which should be retired — protecting faculty time.

EDUCATIONAL IMPACT

MCQs shape how students study — for better or worse.

Item quality isn't just a psychometric concern. Every question you write signals to students what matters — and how they should think about it.

POOR ITEMS DRIVE…

- Rote memorization
- Misleading clinical priorities
- Reinforcement of errors

GOOD ITEMS DRIVE…

- Deep clinical reasoning
- Real decision-making
- True competency validation

KEY ASPECTS

Key Aspects of Item Analysis.

P

ITEM DIFFICULTY

How hard is the item?

P = 0 means nobody got it right. P = 1 means everyone did. Items in the middle range give us the most diagnostic information.

P = Correct ÷ Total

Target: 0.50 ≤ P ≤ 0.70

P VALUE	INTERPRETATION
< 0.50	Difficult item
0.50 – 0.70	**Medium difficulty ★ Target**
> 0.70	Easy item

WORKED EXAMPLE — ITEM DIFFICULTY · N = 60 STUDENTS

Stem: A 67-year-old man with 3-day fever, productive cough with rusty sputum, dyspnea. CXR: right lower lobe consolidation. Temp 38.9°C, SpO_2 94%. **What is the most appropriate first-line antibiotic?**

A. Azithromycin	12 students
B. Amoxicillin ★	**40 students**
C. Ceftriaxone	5 students
D. Levofloxacin	3 students

P = 40 / 60 = **0.67**

Medium Difficulty ✓ — 0.50 ≤ 0.67 ≤ 0.70

D

ITEM DISCRIMINATION

Point-biserial correlation: does the item consistently measure what matters?

We rank all students by total score and compare how the top group vs. bottom group performed on a single item. A D ≥ 0.40 is excellent.

$$D = (UG - LG) \div N$$

UG = upper group correct · LG = lower group correct · N = students per group

D VALUE	INTERPRETATION
≥ 0.40	Excellent ★
0.30 – 0.39	Good
0.20 – 0.29	Fair
≤ 0.19	Poor — review item
< 0	Negative — revise/remove

COMPUTING THE UPPER & LOWER 27% GROUPS

1. **Step 1:** Rank-order all students from highest to lowest total test score.
2. **Step 2:** Multiply total students × 0.27 to get group size. Round to nearest whole number.
3. **Step 3:** Top 27% = Upper Group · Bottom 27% = Lower Group.

Example: 80 students × 0.27 = 21.6, rounded to 22 students per group.

WORKED EXAMPLE — ITEM DISCRIMINATION · N = 80 STUDENTS

Stem: A 45-year-old woman with sudden palpitations, ECG shows narrow QRS at 180 bpm. BP 110/70 mmHg. **What is the most appropriate first-line management?**

A. Diltiazem	15
B. Synchronized cardioversion	12
C. Vagal maneuvers ★	**45**
D. Amiodarone	8

Upper group (22 students): **16 correct** · Lower group (22 students): **12 correct**

$$D = (16 - 12) \div 22 = \mathbf{0.18}$$

Poor Discrimination (D ≤ 0.19) — This item needs review

DISTRACTOR ANALYSIS · N = 100 STUDENTS · UPPER/LOWER GROUP = 27 EACH

Stem: A 52-year-old man with type 2 diabetes presents with glucose 485 mg/dL, polyuria, and confusion. **What is the most appropriate initial management?**

A. IV Normal Saline ★	**55**
B. Subcutaneous Insulin	25
C. IV Sodium Bicarbonate	15
D. Oral Hydration	5

Group	A ★	B	C	D
Upper (27)	**20**	4	2	1
Lower (27)	**8**	**12**	5	2

DISTRACTOR B — WELL-FUNCTIONING

12 lower vs. 4 upper. Plausible to students with incomplete knowledge. Keep.

DISTRACTOR C — MODERATE

5 lower vs. 2 upper. Functioning but could be more plausible.

DISTRACTOR D — NON-FUNCTIONING

Only 2 vs. 1. Not plausible enough — revise or replace.

r_{pbi}

ITEM RELIABILITY

Point-biserial correlation: does the item consistently measure what matters?

Students who perform well overall should get this item right more often. Point-biserial correlation (r_{pbi}) quantifies that relationship — and flags items that don't discriminate on meaningful knowledge.

Target: $r_{pbi} \geq 0.30$

R_{PBI}	INTERPRETATION
> 0.40	Excellent ★
0.30 – 0.39	Good
0.20 – 0.29	Fair
< 0.20	Poor — review item

BEYOND THE FOUR METRICS

Three patterns that also flag items for revision.

Mis-keyed

High performers consistently choose an option marked wrong. D will be negative or near zero.

WHAT IT LOOKS LIKE

The keyed answer attracts far fewer upper-group responses than a distractor.

Group	A ★	B	C	D
Upper	4	20	2	1
Lower	8	12	5	2

Upper group prefers B — the key is likely wrong.

Guessing

$r_{pbi} < 0.20$. Upper group spreads nearly evenly — can't distinguish knowledge from chance.

WHAT IT LOOKS LIKE

No option draws meaningfully more high-performers; the item provides no signal.

Group	A ★	B	C	D
Upper	7	8	6	6
Lower	8	12	5	2

Upper group nearly even — item doesn't cue meaningful knowledge.

Ambiguity

Upper group splits between two options. They're not confused — they're seeing a real defensible ambiguity.

WHAT IT LOOKS LIKE

Two options attract equal numbers of top performers.

Group	A ★	B	C	D
Upper	10	4	10	3
Lower	8	12	5	2

Upper group splits A vs. C — both appear defensible.

SUMMARY

Workshop at a Glance.

THE FOUNDATION

Item Analysis

Systematic evaluation across four key dimensions.

DIFFICULTY

P

P = Correct / Total

0.50 – 0.70

GROUPS

27%

Top 27% vs. Bottom 27%

DISCRIMINATION

D

D = (UG – LG) / N

≥ 0.30

DISTRACTORS

φ

Are wrong options attracting the right students?

WATCH FOR

- Mis-keyed
- Guessing
- Ambiguous

POINT-BISERIAL CORRELATION

r_{pbi}

≥ 0.20 Item-to-total score correlation

THEORETICAL FRAMEWORK

CTT

Practical for classroom use — today's focus.

IRT

For large-scale, high-stakes assessment.

OFF-PRESENTATION ACTIVITY

Practical Application

10 minutes

TAKE-HOME MESSAGES

Three ideas worth carrying.

"

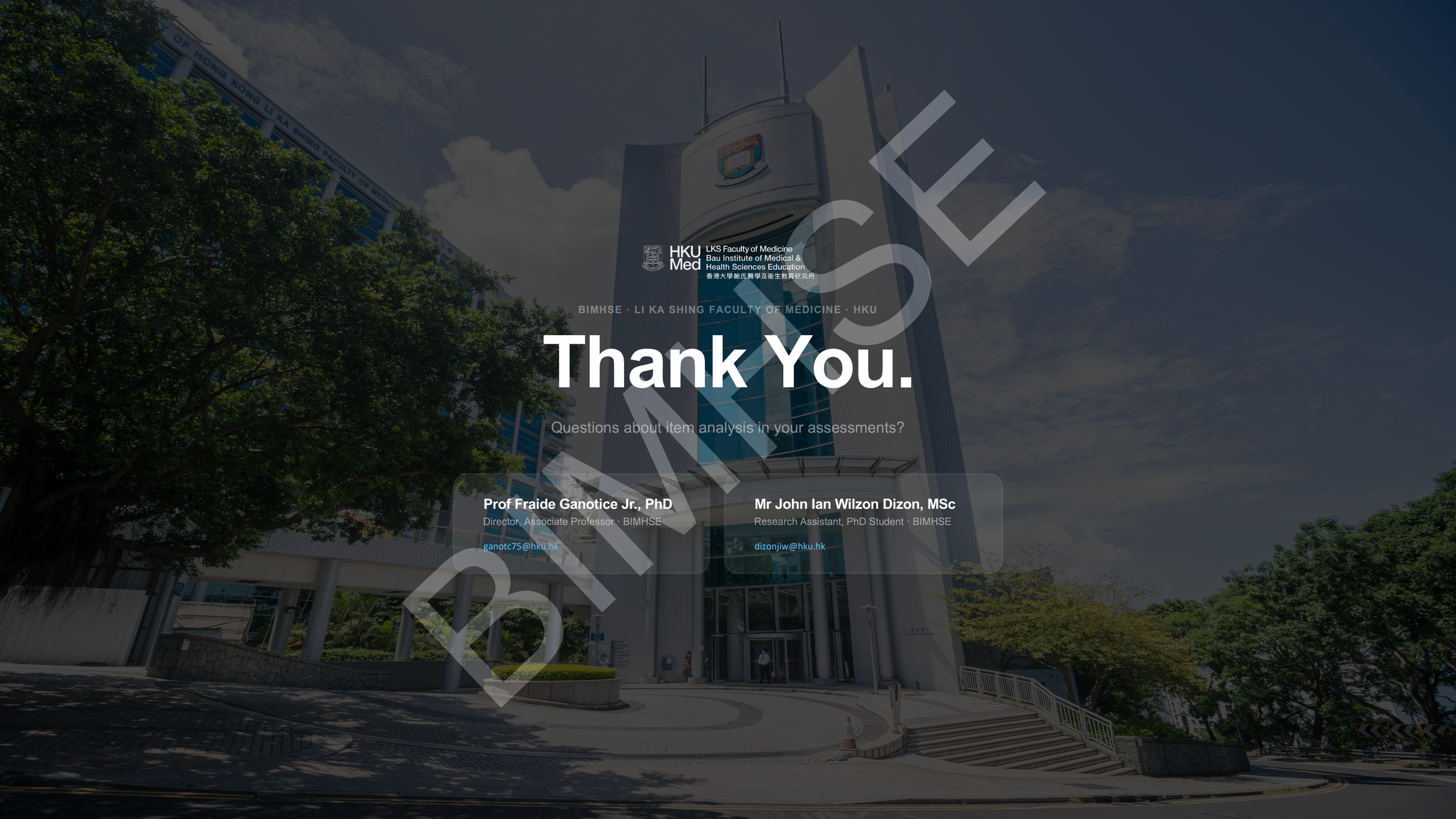
No item is perfect on the first writing; improvement is an ongoing process.

"

Good items don't happen by accident — they require careful analysis and refinement.

"

Item analysis statistics are diagnostic tools, not verdicts — always combine statistical evidence with content review before deciding to keep, revise, or drop an item.



HKU Med
LKS Faculty of Medicine
Bau Institute of Medical & Health Sciences Education
香港大學鮑氏醫學及衛生教育研究所

BIMHSE · LI KA SHING FACULTY OF MEDICINE · HKU

Thank You.

Questions about item analysis in your assessments?

Prof Fraide Ganotice Jr., PhD
Director, Associate Professor · BIMHSE
ganotc75@hku.hk

Mr John Ian Wilzon Dizon, MSc
Research Assistant, PhD Student · BIMHSE
dizonjiw@hku.hk