ศูนย์ความเป็นเลิศด้านการศึกษาวิทยาศาสตร์สุขภาพ
คณะแพทยศาสตร์ศิริราชพยาบาล มหาวิทยาลัยมหิดล

Mahidol University
Faculty of Medicine
Siriraj Hospital

SHEE
Siriraj Health science Education Excellence center

Assessment workshop for clinical teachers

วัดผลนักศึกษาอย่างไร ให้ถูกต้อง เที่ยงตรง และเป็นธรรม

หัวข้อที่น่าสนใจ

Part 1 : หลักการพื้นฐานของการวัดผล (28 ต.ค.)

- Basic principles of assessment
- Standard setting
- Item analysis
- Grading

Part 2 : การพัฒนาข้อสอบ (29 - 30 ต.ค.)

- Multiple-choice questions
- Constructed response item
- Long case examination
- OSCE
- Portfolio
- Clinical performance ratings
- Workplace-based assessment

เอกสารประกอบการอบรม

28 - 30 ต.ค. 63
ณ ห้องบรรยาย 3A01
อาคารศรีสวรินทิรา ชั้น 3A

Contact Us

ศูนย์ความเป็นเลิศด้านการศึกษาวิทยาศาสตร์สุขภาพ
โทร. 024199978 / 024196637 E-mail : sishee@mahidol.edu

shee.si.mahidol.ac.th
mahidol.shee

สารบัญ

กำหนดการอบรมเชิงปฏิบัติ เรื่อง Assessment workshop for clinical teachers

ระหว่างวันที่ 28 – 30 ตุลาคม พ.ศ.2563

ณ ห้องบรรยาย 3A01 อาคารศรีสวรินทิรา ชั้น 3A คณะแพทยศาสตร์ศิริราชพยาบาล

Part 1: หลักการพื้นฐานของการวัดผล		
วันพุธที่ 28 ตุลาคม พ.ศ.2563		**วิทยากรหลัก**
08.00 – 08.30 น.	ลงทะเบียน	
08.30 – 10.15 น.	Basic principles of assessment	รศ. ดร.นพ.เชิดศักดิ์ ไอรมณีรัตน์
10.15 – 10.30 น.	พักรับประทานอาหารว่าง	
10.30 – 12.00 น.	Standard setting	รศ. พญ.กษณา รักษมณี
12.00 – 13.00 น.	*พักรับประทานอาหารกลางวัน*	
13.00 – 14.30 น.	Item analysis (MCQ MEQ OSCE)	รศ. ดร.นพ.เชิดศักดิ์ ไอรมณีรัตน์
14.30 – 14.45 น	พักรับประทานอาหารว่าง	
14.45 – 15.45 น.	Grading	รศ. ดร.นพ.เชิดศักดิ์ ไอรมณีรัตน์
15.45 – 16.00 น.	Summary	รศ. ดร.นพ.เชิดศักดิ์ ไอรมณีรัตน์

SHEE Siriraj Health science Education Excellence center

ศูนย์ความเป็นเลิศด้านการศึกษาวิทยาศาสตร์สุขภาพ (ศศว) | Siriraj Health science Education Excellence center (SHEE)
สำนักงาน: ตึกอดุลยเดชวิกรม ชั้น 6 (ห้อง 656) คณะแพทยศาสตร์ศิริราชพยาบาล เลขที่ 2 แขวงศิริราช เขตบางกอกน้อย กรุงเทพฯ 10700
โทรศัพท์. 0 2419 9978, 0 2419 6637 **โทรสาร**. 0 2412 3901 **E-mail** : shee.mahidol@gmail.com

กำหนดการอบรมเชิงปฏิบัติ เรื่อง Assessment workshop for clinical teachers
ระหว่างวันที่ 28 – 30 ตุลาคม พ.ศ.2563
ณ ห้องบรรยาย 3A01 อาคารศรีสวรินทิรา ชั้น 3A คณะแพทยศาสตร์ศิริราชพยาบาล

Part 2 : การพัฒนาข้อสอบ			
วันพฤหัสบดีที่ 29 ตุลาคม พ.ศ.2563		**วิทยากรหลัก**	**วิทยากรร่วม**
08.30 – 10.15 น.	Multiple-choice questions item development	รศ. ดร.นพ.เชิดศักดิ์ ไอรมณีรัตน์	
10.15 – 10.30 น.	พักรับประทานอาหารว่าง		
10.30 – 12.00 น.	Multiple-choice questions item review	รศ. ดร.นพ.เชิดศักดิ์ ไอรมณีรัตน์	รศ. พญ.กษณา รักษมณี ผศ. นพ.สุประพัฒน์ สนใจพาณิชย์ ผศ. นพ.ทศ หาญรุ่งโรจน์ อ. ดร. นพ.ยอดยิ่ง แดงประไพ อ.นพ.พงษ์เทพ พิศาลธุรกิจ
12.00 – 13.00 น.	*พักรับประทานอาหารกลางวัน*		
13.00 – 14.00 น.	Constructed response item development	ผศ. นพ.สุประพัฒน์ สนใจพาณิชย์	
14.00 – 15.00 น.	OSCE item development	รศ. ดร.นพ.เชิดศักดิ์ ไอรมณีรัตน์	
15.00 – 16.00 น.	Group exercise	ผศ. นพ.สุประพัฒน์ สนใจพาณิชย์	รศ. ดร.นพ.เชิดศักดิ์ ไอรมณีรัตน์ รศ. พญ.กษณา รักษมณี ผศ. นพ.ทศ หาญรุ่งโรจน์ อ.นพ.ภูมิ์ ตริตระการ ผศ. ดร.ทัศนียา รัตนฤาทัย นพรัตน์แจ่มจำรัส
วันศุกร์ที่ 30 ตุลาคม พ.ศ.2563		**วิทยากรหลัก**	**วิทยากรร่วม**
08.00 – 08.30 น.	ลงทะเบียน		
08.30 – 09.15 น.	Long case examination	ผศ. นพ.สุประพัฒน์ สนใจพาณิชย์	
09.15 – 12.00 น.	MEQ & OSCE item review	ผศ. นพ.สุประพัฒน์ สนใจพาณิชย์ ผศ. นพ.ทศ หาญรุ่งโรจน์	รศ. ดร.นพ.เชิดศักดิ์ ไอรมณีรัตน์ รศ. พญ.กษณา รักษมณี ผศ.ดร. วรวรรณ วาณิชย์เจริญชัย
12.00 – 13.00 น.	*พักรับประทานอาหารกลางวัน*		
13.00 – 13.45 น.	Portfolio	รศ. พญ.กษณา รักษมณี	
13.45 – 14.45 น.	Clinical performance ratings	รศ. ดร.นพ.เชิดศักดิ์ ไอรมณีรัตน์	
15.00 – 15.45 น.	Workplace-based assessment	อ.นพ.ภูมิ์ ตริตระการ	รศ. พญ.กษณา รักษมณี
15.45 – 16.00 น.	Summary	รศ. ดร.นพ.เชิดศักดิ์ ไอรมณีรัตน์	

หมายเหตุ: กำหนดการอาจมีการเปลี่ยนแปลงตามความเหมาะสม

ศูนย์ความเป็นเลิศด้านการศึกษาวิทยาศาสตร์สุขภาพ (ศศว) | Siriraj Health science Education Excellence center (SHEE)
สำนักงาน: ตึกอดุลยเดชวิกรม ชั้น 6 (ห้อง 656) คณะแพทยศาสตร์ศิริราชพยาบาล เลขที่ 2 แขวงศิริราช เขตบางกอกน้อย กรุงเทพฯ 10700
โทรศัพท์. 0 2419 9978, 0 2419 6637 **โทรสาร**. 0 2412 3901 **E-mail** : shee.mahidol@gmail.com

รายชื่อผู้ร่วมอบรม

โครงการอบรมเชิงปฏิบัติ เรื่อง Assessment workshop for clinical teachers
ระหว่างวันที่ 28 – 30 ตุลาคม พ.ศ. 2563

กลุ่มที่ 1					
ลำดับ	คำนำหน้า	ชื่อ	สกุล	สังกัด	หน่วยงาน/ภาควิชา
1	ผศ. นพ.	สุภเลิศ	ประคุณหังสิต	คณะแพทยศาสตร์ศิริราชพยาบาล	ภาควิชาจักษุวิทยา
2	อ.นพ.ดร.	นพศักดิ์	ผาสุขกิจวัฒนา	คณะแพทยศาสตร์ศิริราชพยาบาล	ภาควิชาจักษุวิทยา
3	ผศ.พญ.	สุธาสินี	บุญโสภณ	คณะแพทยศาสตร์ศิริราชพยาบาล	ภาควิชาจักษุวิทยา
4	อ. พญ.	เพียงพร	ศักดิ์ศิริวุฒิโฒ	คณะแพทยศาสตร์ศิริราชพยาบาล	ภาควิชาจักษุวิทยา
5	รศ.พญ.	ชารีนันท์	จิรภาไพศาล	คณะแพทยศาสตร์ศิริราชพยาบาล	ภาควิชาจักษุวิทยา

กลุ่มที่ 2					
ลำดับ	คำนำหน้า	ชื่อ	สกุล	สังกัด	หน่วยงาน/ภาควิชา
1	พญ.	พัฒนรินทร์	จุฬาลักษณ์ศิริบุญ	คณะแพทยศาสตร์โรงพยาบาลรามาธิบดี มหาวิทยาลัยมหิดล	ภาควิชาวิสัญญีวิทยา
2	นพ.	พรเทพ	รัชฎาภรณ์กุล	โรงพยาบาลมหาวิทยาลัยนเรศวร	ภาควิชาศัลยศาสตร์
3	นพ.	ราวิน	วงษ์สถาปนาเลิศ	โรงพยาบาลเจริญกรุงประชารักษ์	กลุ่มงานศัลยกรรม
4	พญ.	กุณฑลี	สุกาญจนาเศรษฐ์	โรงพยาบาลมหาวิทยาลัยเทคโนโลยีสุรนารี	Orthopedic
5	ดร.	จริยา	บุญเอี่ยม	วิทยาลัยวิทยาศาสตร์การแพทย์เจ้าฟ้าจุฬาภรณ์ ราชวิทยาลัยจุฬาภรณ์	กายภาพบำบัด

กลุ่มที่ 3					
ลำดับ	คำนำหน้า	ชื่อ	สกุล	สังกัด	หน่วยงาน/ภาควิชา
1	พญ.	ณิชพร	สงวนดีกุล	โรงพยาบาลพระนั่งเกล้า	สูตินรีเวชกรรม
2	พญ.	ธนิษฐา	ศิริธัญญาลักษณ์	โรงพยาบาลสิรินธร	สูตินรีเวชกรรม
3	พญ.	กฤตยา	ภิรมย์	โรงพยาบาลสิรินธร	สูตินรีเวชกรรม
4	นางสาว	วรรณิกา	แสงสุริย์	โรงพยาบาลตากสิน	สูติศาสตร์-นรีเวช

กลุ่มที่ 4					
ลำดับ	คำนำหน้า	ชื่อ	สกุล	สังกัด	หน่วยงาน/ภาควิชา
1	นพ.	ศุภวุฒิ	สุขสันติเลิศ	โรงพยาบาลเจริญกรุงประชารักษ์	กุมารเวชกรรม
2	พญ.	วาศินีย์	นรเศรษฐีกุล	โรงพยาบาลเจริญกรุงประชารักษ์	กุมารเวชกรรม
3	พญ.	เมธินี	โพธิวราพรรณ	โรงพยาบาลเจริญกรุงประชารักษ์	กุมารเวชกรรม
4	พญ.	ปัทมา	เชาว์โพธิ์ทอง	คณะแพทยศาสตร์ศิริราชพยาบาล	ภาควิชาสูติศาสตร์-นรีเวชวิทยา
5	รศ.พญ.	สุชาดา	อินทวิวัฒน์	คณะแพทยศาสตร์ศิริราชพยาบาล	ภาควิชาสูติศาสตร์-นรีเวชวิทยา

กลุ่มที่ 5					
ลำดับ	คำนำหน้า	ชื่อ	สกุล	สังกัด	หน่วยงาน/ภาควิชา
1	นางสาว	ศศิมา	เอี่ยมพันธุ์	คณะแพทยศาสตร์ศิริราชพยาบาล	ภาควิชาตจวิทยา
2	พญ.	ปภาพิต	ตู้จินดา	คณะแพทยศาสตร์ศิริราชพยาบาล	ภาควิชาตจวิทยา
3	อ.พญ.มล	กัญญ์ทอง	ทองใหญ่	คณะแพทยศาสตร์ศิริราชพยาบาล	ภาควิชาโสต นาสิก ลาริงซ์วิทยา
4	พญ.	วรรณวรางค์	ตีรสมิทธ์	คณะแพทยศาสตร์ศิริราชพยาบาล	ภาควิชารังสีวิทยา

กลุ่มที่ 6					
ลำดับ	คำนำหน้า	ชื่อ	สกุล	สังกัด	หน่วยงาน/ภาควิชา
1	นางสาว	ปณิธิดา	แซ่เฮง	คณะแพทยศาสตร์ศิริราชพยาบาล	สถานการแพทย์แผนไทยประยุกต์
2	นางสาว	ดรุณี	รัตนวงศาเมธากุล	คณะแพทยศาสตร์ศิริราชพยาบาล	สถานการแพทย์แผนไทยประยุกต์
3	นาย	ศุภกิจ	สุวรรณไตรย์	คณะแพทยศาสตร์ศิริราชพยาบาล	สถานการแพทย์แผนไทยประยุกต์
4	นาง	เอื้อมพร	สุวรรณไตรย์	คณะแพทยศาสตร์ศิริราชพยาบาล	สถานการแพทย์แผนไทยประยุกต์
5	นาง	พนิตอนงค์	คำแก้ว	คณะแพทยศาสตร์ศิริราชพยาบาล	สถานการแพทย์แผนไทยประยุกต์

เอกสารประกอบการอบรม

28 October 2020

Part 1 : หลักการพื้นฐานของการวัดผล

รศ.ดร. นพ.เชิดศักดิ์ ไอรมณีรัตน์

หัวข้อ : Basic principles of assessment

Basic Principles of Assessment

นพ. เชิดศักดิ์ ไอรมณีรัตน์

ภาควิชาศัลยศาสตร์ คณะแพทยศาสตร์ศิริราชพยาบาล

มหาวิทยาลัยมหิดล

Assessment

- The process of documenting, usually in measurable terms, knowledge, skills, attitudes and beliefs.

Assessment drives instruction.

"Purposeful assessment drives instruction and affects learning."

Wisconsin's guiding principles for teaching and learning

Objectives

เมื่อสิ้นสุดการอบรมแล้ว อาจารย์ผู้เข้ารับการอบรมสามารถ

- ระบุถึงปัจจัยสำคัญในการวางแผนการประเมินผล
- นำข้อแนะนำของการจัดประเมินผลไปปรับใช้ในการจัดสอบต่างๆ ในสถาบันของตนเองเพื่อให้เกิดการประเมินผลที่มีประสิทธิภาพ
- อธิบายถึงลำดับขั้นของการประเมินผลทั้งสี่ลำดับและจัดหาเครื่องมือเพื่อใช้สำหรับการประเมินผลในลำดับขั้นต่างๆได้อย่างเหมาะสม

Outline

- Assessment and instruction
- Basic considerations in planning an assessment
- Guidelines for effective assessment
- Choosing assessment methods

A Research Study

- 124 university students age 18 – 24 years
- Subject: English reading comprehension
- 2 x 3 groups
- Two learning approaches
  - Group A: Study, Study
  - Group B: Study, Test
- Three testing times: 5 min, 2 days, 1 week

Roediger HL, Karpicke JD. Test-enhanced learning: Taking memory tests improves long-term retention. Psychological Science 2006, 17(3): 249-55.

A Research Study

- 180 university students age 18 – 24 years
- Subject: English reading comprehension
- 3 x 2 groups
- Three learning approaches
  - Group A: Study, Study, Study, Study
  - Group B: Study, Study, Study, Test
  - Group C: Study, Test, Test, Test
- Two testing times: 5 min, 1 week

Roediger HL, Karpicke JD. Test-enhanced learning: Taking memory tests improves long-term retention. Psychological Science 2006, 17(3): 249-55.

The Benefit of Testing

- Repeated testing is an effective learning strategy to promote long term memory.
- Self-test should be done early.

Testing Effect or Test-enhanced learning

Karpicke JD, Butler AC, Roediger HL. Metacognitive strategies in student learning: Do students practise retrieval when they study on their own? Memory 2009, 17(4): 471-9.
Roediger HL, Karpicke JD. Test-enhanced learning: Taking memory tests improves long-term retention. Psychological Science 2006, 17(3): 249-55

Assessment and Instructional Process

- Placement
  - Aims at determining the readiness of students for the planned instruction
- Formative
  - Aims at providing feedback to students and teachers concerning learning successes and failures
- Summative
  - Aims at determining the extent to which instructional goals have been achieved; used primarily for assigning grades

Medical Council of Thailand Core Competencies (2012)

- พฤตินิสัย เจตคติ คุณธรรม และจริยธรรมแห่งวิชาชีพ Professional habits, attitudes, moral, and ethics
- ทักษะในการสื่อสารและสร้างสัมพันธภาพ Communication and interpersonal skills
- ความรู้พื้นฐาน Medical knowledge
- การบริบาลผู้ป่วย Patient care
- การสร้างเสริมสุขภาพและระบบสุขภาพ Health promotion and health care system
- การพัฒนาความสามารถทางวิชาชีพอย่างต่อเนื่อง Continuous professional development

Activity

- ให้อาจารย์แต่ละกลุ่ม ช่วยกันระดมสมอง หาวิธีการประเมินการเรียนรู้ในวัตถุประสงค์ต่อไปนี้

1. พฤตินิสัย คุณธรรม จริยธรรม
2. ทักษะในการสื่อสาร
3. ทักษะการแปลผลการตรวจค้นเพิ่มเติม
4. ทักษะการทำหัตถการเพื่อตรวจรักษาโรค
5. การสร้างเสริมสุขภาพ
6. ทักษะการพัฒนาความสามารถทางวิชาชีพ

(เวลา 5 นาที)

Criteria for Good Assessment

- Validity
- Reliability (Reproducibility)
- Equivalence
- Feasibility
- Educational Effect
- Catalytic Effect
- Acceptability

Norcini J, et al. Criteria for good assessment: Consensus statement and recommendations from the Ottawa 2010 conference. Med Teach 2011: 33 (3) 206-14.

1. Validity

- The extent to which an assessment instrument measures what it intends to measure
- The degree to which evidence and theory support the interpretations of test scores entailed by the proposed uses of tests

Validity Threats

- **Construct Underrepresentation**

  The degree to which a test fails to capture important aspects of the construct. The test does not adequately sample some parts of the content

- **Construct-Irrelevant Variance**

  The degree to which test scores are affected by processes that are extraneous to its intended construct

Examples

- Vocabulary or sentence structures are too difficult
- Inadequate time
- Teachers' bias in rating/scoring
- Students' access to test item pool

2. Reliability

- Consistency of test scores
  - If we test the students/residents again, will they get the same scores?

Classical Test Theory

$$T = O + e$$

T = True score

O = Observe score

e = Error

Error

- Systematic error
- Random error

Random Error

- Impact scores in an unpredictable manner
- Causes
  - Fluctuation in memory
  - Variations in motivation
  - Variations in concentration
  - Carelessness
  - Luck in guessing

Reliability of Test Scores

- Reliability coefficient / Reliability index
- Indicate the consistency of test scores from one measurement to another
- Range: 0 – 1
- High values: highly consistent test scores

Reliability of Written Tests

- Test-retest method
- Equivalent-forms method
- Test-retest with equivalent forms
- Internal consistency

Internal Consistency Reliability

- Split-half method

$$Reliability = \frac{2r}{1+r}$$

r = Reliability for half test

- Kuder-Richarson Formula 20 (KR-20)
  An average of all split-half coefficients when the test is split in all possible ways

KR-20

$$KR20 = \left(\frac{n}{n-1}\right)\left(1 - \frac{\sum pq}{Var}\right)$$

n = number of items
Var = Variance of the whole test
p = Proportion of people passing the item
q = Proportion of people failing the item

How Much is Enough?

- Depends on test scores uses
  - High-stakes exam: 0.9 or higher
  - Medium-stakes exam: 0.80 – 0.89
  - Low-stakes exam: 0.70 – 0.79

Improving Reliability

- Increase the number of test items
- Adjust item difficulty to obtain larger spread of test scores
- Adjust testing conditions to eliminate interruptions, noise, and other disrupting factors
- Eliminate subjectivity in scoring

Spearman-Brown Formula

$$r_k = \frac{kr_1}{1 + (k-1)r_1}$$

- r_k = Reliability of a test “k” times long
- r_1 = Reliability of the original test
- k = factor by which test length is changed

Example

- Original test = 10 items, KR-20 = 0.67
- What is the reliability if the test is lengthen to 20 items
- K = 2
- r = 2(0.67)/[1+(2-1)(0.67)] = 0.80

3. Equivalence

- การจัดสอบหัวข้อเดียวกันกับนักศึกษาระดับชั้นเรียนเดียวกัน ที่จัดสอบกันต่างเวลา ได้คะแนนที่เทียบเคียงกันได้

4. Feasibility

ความเป็นไปได้ของการจัดสอบ

The assessment is practical, realistic, and sensible, given appropriate contexts:

- Time
- Money
- Expertise
- Administration

5. Educational Effect

- การประเมินผลนั้นกระตุ้นให้ผู้เรียนมีการเรียนรู้ในเรื่องที่ควรเรียนรู้ … educational benefit

6. Catalytic Effect

- การประเมินผลก่อให้เกิดการนำผลของการสอบไปใช้ให้ feedback เพื่อสร้าง หรือส่งเสริม หรือสนับสนุนการเรียนรู้ของนักศึกษา

7. Acceptability

- ผู้เกี่ยวข้อง (stakeholders) ทั้งหมดเชื่อถือผลการประเมิน

Practical guidelines

- Eight basic guidelines for effective assessment
- Gronlund NE. Assessment of student achievement, 7th ed. Boston, MA: Pearson education; 2003.

Guidelines for Effective Assessment (1)

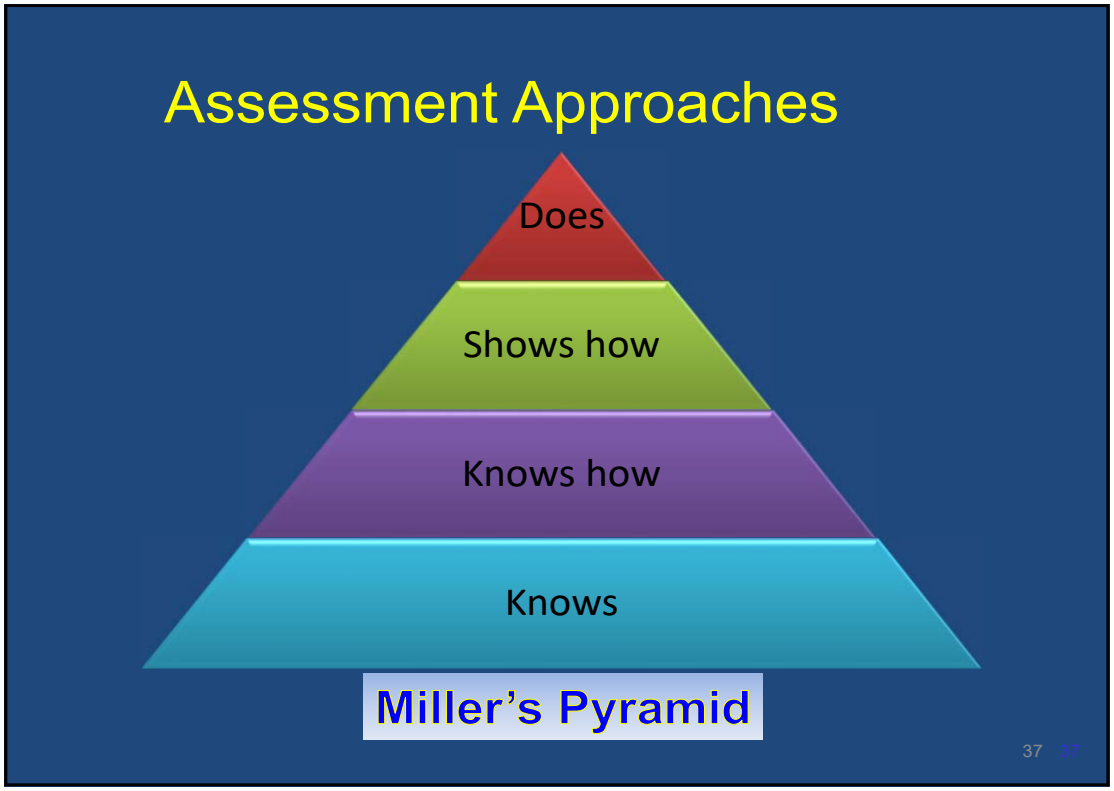
1. Effective assessment requires a clear conception of all intended learning outcomes.
2. Effective assessment requires that a variety of assessment procedures be used.
3. Effective assessment requires that the instructional relevance of the procedures be considered.

Guidelines for Effective Assessment (2)

4. Effective assessment requires an adequate sample of student performance.
5. Effective assessment requires that the procedures be fair to everyone.
6. Effective assessment requires the specifications of criteria for judging successful performance.

Guidelines for Effective Assessment (3)

7. Effective assessment requires feedback to students that emphasizes strengths of performance and weaknesses to be corrected.
8. Effective assessment must be supported by a comprehensive grading and reporting system



หลักในการเลือกวิธีประเมินผล

1. คำนึงถึงผลลัพธ์ที่ต้องการวัดว่าเป็นความรู้ ความสามารถในระดับใดของ Miller's pyramid
2. คำนึงถึงหลักในการประเมินผลที่ดี (criteria for good assessment)
3. เลือกวิธีการที่บรรลุวัตถุประสงค์ได้ด้วยความประหยัดทรัพยากร

Summary

- Assessment and instruction
- Basic considerations in planning an assessment
- Guidelines for effective assessment
- Choosing assessment methods

Iramaneerat C. Validity threats [Thai]. Medical Education Pamphlet 2006; 2(9): 1.

สิ่งไม่พึงประสงค์ในการสอบ

เชิดศักดิ์ ไอรมณีรัตน์

ในบทความนี้ผมจะขอกล่าวถึงสิ่งอันไม่พึงประสงค์ในการสอบ (Validity threats) ที่เราต้องคำนึงถึงในการจัดสอบ ดังที่ได้กล่าวในบทความก่อนหน้านี้แล้วว่า Validity นั้นคือการประเมินคุณค่าของการแปลผลและการนำผลสอบไปใช้ ดังนั้น สิ่งอันไม่พึงประสงค์ในการสอบ หรือ validity threats ก็คือสิ่งใดก็ตามที่เข้ามารบกวนการแปลผลสอบ สิ่งรบกวนเหล่านี้แยกได้เป็น 2 ปัจจัยหลัก คือ construct underrepresentation และ construct-irrelevant variance

Construct underrepresentation หมายถึงการประเมินผลที่ไม่ครอบคลุมสิ่งที่ต้องการวัดอย่างเพียงพอ ทำให้ผลการสอบไม่สามารถบ่งบอกถึงความสามารถของนักเรียนผู้สอบในเรื่องที่ต้องการวัดผลอย่างครบถ้วน ตัวอย่างเช่นในการสอบ OSCE เพื่อวัดความสามารถของแพทย์ประจำบ้านในการให้คำแนะนำปรึกษาแก่ผู้ป่วย หากเกณฑ์การให้คะแนนมีเพียงหัวข้อที่เกี่ยวกับการพูดกับผู้ป่วย แต่ไม่มีหัวข้อที่เกี่ยวกับการใช้ อวัจนภาษา เช่น การใช้ท่าทาง น้ำเสียง การรับฟังปัญหา เป็นต้น ก็จัดว่า ทำการประเมินไม่ครอบคลุมเนื้อหา ผลการประเมินก็นำไปใช้บอกได้เพียงว่าแพทย์ประจำบ้านให้ข้อมูลผู้ป่วยครบถ้วน แต่ไม่สามารถบอกได้ว่าแพทย์ประจำบ้านทำการสื่อสารกับผู้ป่วยได้ดีในทุกด้าน ในการสอบข้อเขียนสำหรับวัดความรู้ของนักเรียน หากใช้ข้อสอบที่สั้นเกินไป มีจำนวนข้อสอบไม่กี่ข้อ ก็จะมีปัญหาที่ไม่สามารถวัดความรู้ของนักเรียนได้ครอบคลุมเนื้อหาที่ต้องการวัดผล

Construct-irrelevant variance หมายถึง ปัจจัยอื่นที่นอกเหนือไปจากความรู้ความสามารถของนักเรียนที่สามารถส่งผลต่อคะแนนสอบของนักเรียนได้ ปัจจัยที่อาจรบกวนคะแนนสอบ multiple-choice examination ได้แก่

- ข้อสอบที่ไม่มีคุณภาพ โจทย์คำถามกำกวม มีตัวเลือกที่ถูกมากกว่า 1 ตัวเลือก ทำให้นักเรียนที่มีความรู้ตอบผิด หรือโจทย์คำถามบอกใบ้ให้นักเรียนตอบถูกโดยไม่ต้องใช้ความรู้ ข้อสอบเก่าที่รั่วไหลออกจากคลังข้อสอบทำให้นักเรียนที่รู้ข้อสอบมาก่อนสามารถตอบได้โดยไม่ต้องคิด
- นักเรียนที่ทุจริตในการสอบ ลอกข้อสอบของเพื่อน หรือใช้วิธีการอื่นๆในการได้มาซึ่งคำตอบโดยที่ไม่ได้ใช้ความรู้ในเรื่องที่ทำการสอบ
- อาจารย์ที่บอกข้อสอบให้นักเรียนในการสอน ทำให้นักเรียนที่ท่องคำตอบเข้าไปสอบ ทำข้อสอบได้โดยไม่ต้องคิด

สำหรับการสอบในรูปแบบอื่นที่ต้องใช้กรรมการให้คะแนน เช่น OSCE การสอบข้อสอบบรรยาย หรือการสอบปากเปล่า นั้นจะมีปัจจัยที่เกี่ยวเนื่องกับกรรมการผู้ให้คะแนนเข้ามารบกวนการแปลผลคะแนนสอบได้ด้วย เช่น

- ความไม่เสมอภาคของอาจารย์ในเกณฑ์การให้คะแนน นักเรียนที่สอบกับอาจารย์ที่กดคะแนน เสียเปรียบนักเรียนที่สอบกับอาจารย์ที่ใจดี และปล่อยคะแนน
- ความไม่สม่ำเสมอของอาจารย์ในการให้คะแนน อาจารย์บางท่านมีแนวโน้มจะให้คะแนนต่ำลงในกลุ่มนักเรียนที่สอบตอนท้าย เนื่องด้วยความเหนื่อยล้า ในขณะที่อาจารย์บางท่านมีแนวโน้มจะให้คะแนนสูงขึ้นในตอนท้ายของการสอบ เนื่องจากได้เห็นความสามารถของนักเรียนจำนวนหนึ่งแล้วพบว่าเกณฑ์ที่ตั้งเป้าไว้นั้นสูงเกินความสามารถของนักเรียนส่วนใหญ่จึงปรับเกณฑ์การให้คะแนนให้ง่ายลง ทำให้นักเรียนในกลุ่มหลังได้คะแนนง่ายขึ้น
- การจำกัดช่วงของคะแนน ที่พบบ่อยคืออาจารย์บางท่านนิยมเดินสายกลาง ไม่ว่านักเรียนจะทำดีมากหรือน้อยเพียงใด ก็มักจะให้คะแนนอยู่ในเกณฑ์ปานกลาง ไม่กล้าให้คะแนน 0 ในรายที่ทำไม่ดี แต่ก็ไม่กล้าให้คะแนนเต็มในนักเรียนที่ทำได้ดี

ปัจจัยต่างๆ เหล่านี้ เป็นสิ่งที่ผู้จัดสอบต้องคำนึงถึงเสมอในการจัดสอบและตั้งมาตรการเพื่อควบคุมและกำจัดปัจจัยรบกวนเหล่านี้จากการสอบ เพื่อให้ได้ผลการสอบที่มีความเที่ยงตรง เป็นธรรม และสามารถใช้บอกความรู้ ความสามารถของนักเรียนได้ตามที่ต้องการ

Iramaneerat C. Reliability: Part I [Thai]. Medical Education Pamphlet 2006; 2(10): 4.

Iramaneerat C. Reliability: Part II [Thai]. Medical Education Pamphlet 2006; 2(11): 4.

ความแม่นยำของคะแนนสอบ (Reliability)

เชิดศักดิ์ ไอรมณีรัตน์

ในบทความนี้ผมจะขอกล่าวถึงการประเมินความแม่นยำของคะแนนสอบ (Reliability) การตรวจสอบความแม่นยำของคะแนนสอบเป็นการตอบคำถามว่า หากทำการสอบซ้ำนักเรียนจะได้คะแนนเท่าเดิมหรือไม่ ในการสอบทั่วไปมักรายงานความแม่นยำของคะแนนสอบด้วยค่า reliability coefficient ซึ่งมีค่าได้ตั้งแต่ 0 ถึง 1 โดยค่ายิ่งสูงบ่งบอกว่าผลสอบมีความน่าเชื่อถือมาก ค่า reliability coefficient = 0 บอกถึงคะแนนสอบที่ขาดความแม่นยำโดยสิ้นเชิง เทียบได้กับการให้คะแนนนักเรียนโดยการสุ่มตัวเลขให้ ส่วนค่า reliability coefficient = 1 บอกถึงคะแนนสอบที่มีความแม่นยำมาก หากให้นักเรียนสอบซ้ำก็จะได้คะแนนเท่าเดิม เพื่อขยายความเข้าใจผมจะขอกล่าวถึงคุณลักษณะที่สำคัญของ reliability ได้แก่

1. Reliability เป็นคุณสมบัติของคะแนนสอบ ไม่ใช่ตัวข้อสอบ ข้อสอบชุดหนึ่งทำการสอบกับนักเรียนกลุ่มหนึ่งพบว่ามีความแม่นยำสูง แต่เมื่อเอาข้อสอบชุดเดียวกันไปทำการสอบนักเรียนอีกกลุ่มหนึ่ง อาจมีความแม่นยำต่ำได้

2. Reliability มีด้วยกันหลายชนิด และค่า reliability coefficient ที่ได้จากการประเมินความแม่นยำแต่ละชนิดก็แปลผลแตกต่างกัน ดังได้กล่าวแล้วว่า การประเมินความแม่นยำของคะแนนสอบ เป็นการตรวจสอบว่าหากทำการสอบซ้ำจะได้คะแนนเท่าเดิมหรือไม่ ประเด็นสำคัญคือเราจะทำการสอบซ้ำอย่างไร จะสอบซ้ำด้วยข้อสอบชุดเดิม หรือ ข้อสอบชุดใหม่ที่ออกแบบให้เปรียบเทียบได้กับข้อสอบชุดเดิม, สอบซ้ำ ณ เวลาเดียวกัน หรือใกล้เคียงกัน หรือเวลาห่างกันเป็นสัปดาห์, สอบซ้ำโดยใช้กรรมการให้คะแนนคนเดิม หรือสอบซ้ำโดยเปลี่ยนกรรมการให้คะแนน จะเห็นได้ว่า วิธีการสอบซ้ำต่างกันก็บอกความแม่นยำของคะแนนในสถานการณ์ต่างกัน (ความแม่นยำเมื่อเปลี่ยนชุดข้อสอบ หรือความแม่นยำเมื่อเปลี่ยนเวลา หรือ ความแม่นยำเมื่อเปลี่ยนกรรมการให้คะแนน) ดังนั้นการแปลผลของค่า reliability coefficient ต้องทำควบคู่ไปความเข้าใจว่าค่าดังกล่าวบอกถึงความแม่นยำชนิดใด โดยทั่วไปในการวัดความแม่นยำของคะแนนสอบ multiple-choice examination จากการสอบครั้งเดียว มักเป็นการประเมิน internal consistency reliability ซึ่งบ่งบอกว่าข้อสอบทุกข้อที่ใช้ในการสอบนักเรียนกลุ่มหนึ่งๆทำการวัดความรู้ในเรื่องเดียวกันหรือไม่

3. Reliability เป็นปัจจัยที่สำคัญเพียงปัจจัยหนึ่งในการประเมินคุณค่าของผลสอบ ผลสอบที่ไม่มีความแม่นยำนั้นเป็นผลสอบที่มีคุณค่าต่ำไม่สามารถให้ข้อมูลที่เป็นประโยชน์เกี่ยวกับนักเรียนผู้สอบได้ แต่ผลสอบที่มีความแม่นยำสูงนั้นก็ไม่จำเป็นว่าจะเป็นผลสอบที่เราสามารถนำไปใช้ประโยชน์ได้เสมอไป จำเป็นต้องพิจารณาปัจจัยร่วมอื่นๆ อีกหลายอย่าง เช่น หากมีนักเรียนทุจริตในการสอบ คะแนนสอบที่ได้ก็อาจมีค่า reliability coefficient สูง แต่ผลสอบนั้นก็เป็นผลสอบที่บิดเบือน ไม่สามารถบอกได้ว่านักเรียนที่ได้คะแนนสูงเป็นนักเรียนที่มีความรู้ หรือเป็นนักเรียนที่ไม่มีความรู้แต่ลอกข้อสอบเพื่อน

ประเด็นที่ได้รับความสนใจกันมากคือ ค่า reliability coefficient ต้องสูงแค่ไหนจึงจะเพียงพอที่จะนำผลสอบไปใช้ได้ โดยทั่วไปนั้นจำเป็นต้องพิจารณาควบคู่ไปกับการนำผลสอบไปใช้ หากผลสอบนั้นนำไปใช้ในการตัดสินใจที่สำคัญ เมื่อตัดสินใจไปแล้วผลเป็นที่สุดไม่สามารถเปลี่ยนแปลงได้ และส่งผลยาวนาน โดยเฉพาะการตัดสินใจที่ส่งผลต่อตัวบุคคล มักต้องการคะแนนสอบที่มีค่า reliability coefficient สูงมาก ในทางกลับกัน หากผลสอบนั้นใช้ในการตัดสินใจที่ไม่ค่อยสำคัญ มีผลระยะสั้น และการตัดสินผลอาจเปลี่ยนแปลงได้หลังจากการสอบนี้โดยพิจารณาจากการสอบอื่นที่จะจัดตามมาภายหลัง โดยเฉพาะการตัดสินใจที่มีผลต่อนักเรียนเป็นกลุ่ม ไม่ส่งผลต่อตัวบุคคล มักไม่ต้องการค่า reliability coefficient ที่สูงมาก โดยทั่วไปสำหรับการสอบย่อยๆ ใน

ชั้นเรียน ควรให้ค่า reliability coefficient สูงกว่า 0.7 สำหรับการสอบลงกองของนักศึกษาแพทย์ การสอบปลายภาค หรือการสอบใหญ่ต่างๆ ในโรงเรียนแพทย์ ควรให้ค่า reliability coefficient สูงกว่า 0.8 สำหรับการสอบที่มีความสำคัญมาก เช่น การสอบคัดเลือกเข้าเรียนมหาวิทยาลัย การสอบใบอนุญาตประกอบวิชาชีพเวชกรรม การสอบวุฒิบัตรผู้เชี่ยวชาญเฉพาะทาง มักต้องให้ reliability coefficient สูงกว่า 0.9

อีกประเด็นหนึ่งที่มีความสำคัญคือ มีปัจจัยใดบ้างที่ส่งผลต่อค่า reliability coefficient สิ่งเหล่านี้มีความสำคัญมากเมื่อเราต้องการอธิบายว่าเหตุใดคะแนนสอบที่ได้จึงไม่แม่นยำ และเราต้องทำอย่างไรจึงจะทำให้คะแนนสอบมีความแม่นยำมากขึ้น โดยทั่วไปปัจจัยที่สำคัญที่ส่งผลต่อความแม่นยำของคะแนนสอบมีด้วยกัน 4 ปัจจัย คือ

1. จำนวนข้อสอบ ถ้าทำการสอบด้วยข้อสอบที่สั้น ประกอบด้วยคำถามไม่กี่ข้อ คะแนนสอบที่ได้มักไม่แม่นยำ วิธีเพิ่มความแม่นยำของคะแนนสอบที่ง่ายที่สุดคือการเพิ่มจำนวนข้อสอบ
2. การกระจายตัวของคะแนนสอบ ถ้าคะแนนสอบมีความแตกต่างกันมาก มีทั้งนักเรียนที่ทำคะแนนได้สูง และนักเรียนที่ทำคะแนนได้ต่ำ คะแนนสอบมักมีความแม่นยำสูง ในทางตรงข้ามหากนักเรียนทำคะแนนใกล้เคียงกัน คะแนนเกาะกลุ่มกันมาก คะแนนสอบมักมีความแม่นยำต่ำ วิธีการเพิ่มความแม่นยำของคะแนนสอบโดยการเพิ่มการกระจายตัวของคะแนนของนักเรียนทำได้โดยใช้ข้อสอบที่มีความยากมากขึ้น
3. ปัจจัยรบกวนการสอบของนักเรียน หากทำการจัดสอบไม่ดี มีสิ่งมารบกวนนักเรียนในขณะที่ทำการสอบ (เช่น มีเสียงดังรบกวน ห้องสอบร้อนอบอ้าวจนนักเรียนไม่มีสมาธิ) คะแนนสอบมักมีความแม่นยำต่ำ ดังนั้นผู้คุมสอบต้องจัดสถานที่สอบให้ดี เพื่อให้นักเรียนมีสมาธิในการทำข้อสอบ ซึ่งจะนำไปสู่คะแนนสอบมีความแม่นยำสูง
4. ลักษณะการให้คะแนนของข้อสอบ ข้อสอบที่ไม่ต้องใช้กรรมการตรวจ เช่น multiple-choice examination มักให้คะแนนที่มีความแม่นยำสูง ในทางตรงข้ามข้อสอบที่ต้องใช้กรรมการให้คะแนน เช่น ข้อสอบบรรยาย ข้อสอบ OSCE คะแนนที่ได้มักมีความแม่นยำไม่สูงนักเนื่องจากมีปัจจัยที่นอกเหนือไปจากความสามารถของนักเรียน (เช่น ความเหนื่อยล้าของกรรมการ ความไม่สม่ำเสมอของการใช้เกณฑ์ให้คะแนน หรือ อารมณ์ของกรรมการตรวจข้อสอบ) เข้ามาส่งผลต่อคะแนนสอบ

รศ. พญ.กษณา รักษมณี

หัวข้อ : Standard setting

Iramaneerat C. Passing standard: Part I [Thai]. Medical Education Pamphlet 2006; 2(1): 3.

วิธีการตั้งเกณฑ์สอบผ่าน (passing standard) (ตอนที่ 1)

เชิดศักดิ์ ไอรมณีรัตน์

เกณฑ์สอบผ่าน (passing standard) คือคะแนนสอบที่น้อยที่สุดที่คณาจารย์ยินยอมให้นักเรียนสามารถสอบผ่าน นักเรียนที่สอบได้คะแนนน้อยกว่าเกณฑ์สอบผ่านจะถูกตัดสินว่าสอบตก การตั้งเกณฑ์สอบผ่านจัดเป็นขั้นตอนที่มีความสำคัญมากในการจัดสอบ แต่กลับไม่ได้รับความสนใจเท่าที่ควรในการวัดผลทางแพทยศาสตรศึกษาจำนวนมาก ในบทความนี้ผมขอเสนอเกร็ดความรู้เกี่ยวกับวิธีการตั้งเกณฑ์สอบผ่าน ผมหวังว่าอาจารย์ผู้อ่านจะสามารถนำเกร็ดความรู้นี้ไปใช้พัฒนาคุณภาพของการตั้งเกณฑ์สอบผ่านได้ไม่มากก็น้อยครับ

เกณฑ์สอบผ่านในทางแพทยศาสตรศึกษาจัดว่ามีความสำคัญมากเนื่องจากเกณฑ์สอบผ่านเป็นการแสดงออกถึงมาตรฐานของวิชาชีพที่อาจารย์ยอมรับ เกณฑ์สอบผ่านที่ดีต้องได้รับการตั้งขึ้นโดยใช้ดุลยพินิจของคณาจารย์ผู้เชี่ยวชาญในสาขาวิชานั้นๆเพื่อรักษามาตรฐานการประกอบวิชาชีพเพื่อให้สังคมได้รับบริการทางการแพทย์ที่มีคุณภาพ ในขณะเดียวกันกับให้ความเป็นธรรมกับนักเรียนผู้สอบ เนื่องจากเกณฑ์สอบผ่านเป็นการแสดงออกถึง "ความยอมรับได้" ในดุลยพินิจของคณาจารย์ผู้เชี่ยวชาญ จึงไม่มีวิธีการทางวิทยาศาสตร์ใดที่จะตัดสินว่าเกณฑ์ที่ตั้งขึ้นนั้นถูกหรือผิด สิ่งที่สำคัญที่สุดในการตั้งเกณฑ์สอบผ่านหาใช่ "ตัวเลข" คะแนนที่จะใช้ตัดสินได้ตก หากแต่เป็น "กระบวนการ" ให้ได้มาซึ่งเกณฑ์ดังกล่าว เกณฑ์สอบผ่านที่ตั้งขึ้นโดยใช้อาจารย์ 1 ท่านเลือกตัวเลข 1 ตัวเลขขึ้นมาโดยไม่ได้พิจารณาถึงข้อสอบหรือนักเรียนผู้สอบ เป็นวิธีการตั้งเกณฑ์ที่ล่อแหลมต่อการถูกวิจารณ์ (และประท้วง) โดยผู้ที่ไม่พอใจในผลสอบ วิธีการตั้งเกณฑ์สอบผ่านที่ดีนั้นต้องมีหลักการและเหตุผลประกอบ และผ่านดุลยพินิจของคณาจารย์ จำนวนของอาจารย์ผู้เชี่ยวชาญที่ต้องใช้ในการตั้งเกณฑ์นั้นขึ้นกับความสำคัญของการสอบนั้นๆ ในการสอบที่มีความสำคัญสูงเช่นการสอบวุฒิบัตรแพทย์ผู้เชี่ยวชาญ แนะนำให้ใช้คณาจารย์อย่างน้อย 6 – 8 ท่าน ในการตั้งเกณฑ์ แต่หากเป็นการสอบเล็กๆ เช่น การทดสอบหลังการสอนกลุ่มย่อย อาจใช้อาจารย์เพียง 1 ท่านก็ได้

การตั้งเกณฑ์สอบผ่านมี 2 ชนิดคือ การตัดสินแบบอิงเกณฑ์ (criterion-referenced standard, absolute standard) และการตัดสินแบบอิงกลุ่ม (norm-referenced standard, relative standard) การตัดสินแบบอิงเกณฑ์ เป็นการตั้งว่า <u>คะแนน</u>เท่าไร จึงจัดว่าผ่านการสอบ ในทางตรงข้าม การตัดสินแบบอิงกลุ่ม เป็นการตั้งว่า จะให้ <u>นักเรียน</u>จำนวนเท่าไร ผ่านการสอบ การตัดสินแบบอิงเกณฑ์นั้นเหมาะกับการสอบเพื่อวัดว่าผู้สอบมีความรู้ความสามารถในด้านใดด้านหนึ่งเพียงพอหรือไม่ ส่วนการสอบแบบอิงกลุ่มนั้นเหมาะสำหรับการสอบแข่งขันเพื่อเข้าศึกษาต่อ หรือ ทำงาน ในสถาบันที่มีตำแหน่งที่จะรับได้จำกัด เช่น การสอบเข้าโรงเรียนแพทย์ หรือ การสอบคัดเลือกแพทย์ประจำบ้าน การสอบส่วนใหญ่ในทางแพทยศาสตรศึกษานั้นเหมาะกับการตัดสินแบบอิงเกณฑ์ หากผู้สอบทุกคนมีความสามารถเพียงพอก็ไม่จำเป็นต้องมีผู้สอบตก การใช้การตัดสินแบบอิงกลุ่มเพื่อวัดความรู้ความสามารถในสถานการณ์อื่นนอกจากการสอบคัดเลือกนั้นเป็นการส่งเสริมให้นักเรียนเกิดความแข่งขันกัน (แทนที่จะช่วยกันเรียน) โดยไม่จำเป็น

เนื่องจากการสอบทางแพทยศาสตรศึกษาแทบทั้งหมดเหมาะกับการตั้งเกณฑ์สอบผ่านแบบอิงเกณฑ์ ผมจะขอขยายความวิธีการตั้งเกณฑ์สอบผ่านแบบอิงเกณฑ์ที่สำคัญและใช้บ่อย 2 วิธีใหญ่ๆ คือ 1. การตั้งเกณฑ์โดยพิจารณาข้อสอบ และ 2. การตั้งเกณฑ์โดยพิจารณาจากผู้สอบ ในบทความตอนต่อไปครับ

Iramaneerat C. Passing standard: Part II [Thai]. Medical Education Pamphlet 2006; 2(2): 2.

วิธีการตั้งเกณฑ์สอบผ่าน (passing standard) (ตอนที่ 2)

เชิดศักดิ์ ไอรมณีรัตน์

ในบทความนี้ผมจะขอแนะนำวิธีการตั้งเกณฑ์สอบผ่านโดยพิจารณาตัวข้อสอบที่ใช้สอบ วิธีการตั้งเกณฑ์ผ่านแบบนี้เหมาะสำหรับการสอบ multiple-choice questions ซึ่งอาจารย์ผู้ตั้งเกณฑ์ผ่านสามารถประเมินความน่าจะเป็นของการตอบข้อสอบแต่ละข้อถูกได้ การตั้งเกณฑ์ผ่านแบบนี้ประกอบด้วย 3 ขั้นตอนหลักคือ

1. ระบุลักษณะของนักเรียน"คาบเส้น" (borderline examinees): นักเรียนในกลุ่มคาบเส้นนี้คือนักเรียนที่มีความรู้ความสามารถอยู่ระหว่าง "ยอมรับได้" กับ "ยอมรับไม่ได้" นักเรียนกลุ่มนี้มีความรู้ไม่มากพอที่อาจารย์จะตัดสินให้สอบผ่านได้อย่างสบายใจ แต่ก็มีความรู้ไม่น้อยจนอาจารย์จะตัดสินให้สอบตกได้โดยไม่มีข้อสงสัย คณะกรรมการตั้งเกณฑ์สอบผ่านต้องระบุลักษณะของนักเรียนในกลุ่มคาบเส้นนี้อย่างชัดเจนว่า ในเนื้อหาวิชาที่ทำการสอบ นักเรียนกลุ่มนี้ควรมีความรู้ในเรื่องใด และไม่มีความรู้ในเรื่องใด ขั้นตอนนี้อาจทำได้ง่ายขึ้นหากอาจารย์แต่ละท่านนึกภาพของนักเรียนจริงที่อาจารย์เคยรู้จักที่สมควรถูกจัดให้อยู่ในกลุ่มนักเรียนคาบเส้น แล้วบรรยายลักษณะของนักเรียนคนนั้นๆ ว่าทำอะไรได้ และทำอะไรไม่ได้ รู้เรื่องอะไรบ้าง ไม่รู้เรื่องอะไรบ้าง
2. ให้กรรมการแต่ละท่านพิจารณาข้อสอบแต่ละข้อ และตัดสินว่านักเรียนคาบเส้นน่าจะมีโอกาสตอบข้อสอบถูกมากน้อยเพียงใด ขั้นตอนนี้สามารถทำได้หลายวิธีด้วยกัน ผมขอยกตัวอย่างวิธีที่เป็นที่แพร่หลายมาก 2 วิธีด้วยกัน คือ
   2.1. Angoff's method: ให้อาจารย์ระบุว่าหากนักเรียนคาบเส้น 100 คนทำข้อสอบข้อนั้น จะมีนักเรียนกี่คนที่ตอบข้อสอบข้อนั้นถูก (หรือความน่าจะเป็นที่นักเรียนคาบเส้นตอบข้อสอบข้อนั้นถูก)
   2.2. Ebel's method: ให้อาจารย์สร้างตารางแยกประเภทข้อสอบตามความสำคัญของเนื้อหาและตามความยากง่ายของข้อสอบและระบุว่าในข้อสอบแต่ละกลุ่ม หากนักเรียนคาบเส้น 100 คนทำข้อสอบจะมีนักเรียนกี่คนที่ตอบถูก หลังจากนั้นให้อาจารย์พิจารณาข้อสอบแต่ละข้อแล้วจัดประเภทเข้าในกลุ่ม ตัวอย่างเช่น

ความยากง่าย / ความสำคัญ	ง่าย	ปานกลาง	ยาก
สำคัญมาก	95%	85%	80%
สำคัญพอควร	90%	75%	60%
สำคัญน้อย	80%	55%	35%
สำคัญน้อยมาก	50%	30%	20%

3. ทำการคิดเกณฑ์สอบผ่านสำหรับข้อสอบนั้น

   3.1. Angoff's method เกณฑ์ผ่านคือผลรวมของความน่าจะเป็นของการตอบข้อสอบแต่ละข้อถูก

Item	1	2	3	4	5	Passing score
Probability	0.95	0.85	0.30	0.40	0.70	3.20

   3.2. Ebel's method เกณฑ์ผ่านคือผลรวมของ (จำนวนข้อสอบในแต่ละกลุ่ม x ความน่าจะเป็นของการตอบข้อสอบถูกสำหรับข้อสอบในกลุ่มนั้น)จากข้อสอบทั้ง 12 กลุ่ม

ความยากง่าย / ความสำคัญ	ง่าย (24 ข้อ)	ปานกลาง (15 ข้อ)	ยาก (11 ข้อ)
สำคัญมาก (15 ข้อ)	95% x 5	85% x 5	80% x 5
สำคัญพอควร (20 ข้อ)	90% x 10	75% x 7	60% x 3
สำคัญน้อย (10 ข้อ)	80% x 5	55% x 3	35% x 2
สำคัญน้อยมาก (5 ข้อ)	50% x 4	30% x 0	20% x 1
Passing score	37.6		

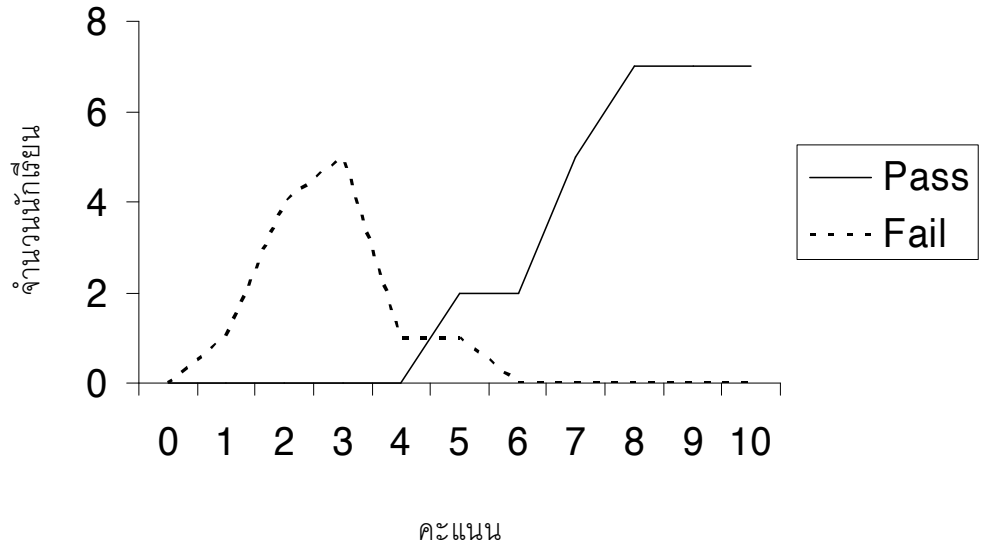
Iramaneerat C. Passing standard: Part III [Thai]. Medical Education Pamphlet 2006; 2(3): 1.

วิธีการตั้งเกณฑ์สอบผ่าน (passing standard) (ตอนที่ 3)

เชิดศักดิ์ ไอรมณีรัตน์

ในบทความนี้ผมจะขอแนะนำวิธีการตั้งเกณฑ์สอบผ่านโดยพิจารณานักเรียนผู้สอบ วิธีการตั้งเกณฑ์ผ่านแบบนี้เหมาะสำหรับการสอบวัดทักษะ การสอบสัมภาษณ์ หรือการประเมินการปฏิบัติงาน ซึ่งมักตัดสินการสอบผ่านโดยดูจากความสามารถของผู้สอบโดยรวมได้ง่ายกว่าดูจากคะแนนที่ได้ในหัวข้อประเมินแต่ละข้อ วิธีการตั้งเกณฑ์ผ่านลักษณะนี้ที่ใช้บ่อยมีด้วยกัน 2 วิธีคือ

1. Borderline-group method: การตั้งเกณฑ์ผ่านวิธีนี้เริ่มจากให้คณะกรรมการสอบประชุมตกลงกันก่อนถึงลักษณะของผู้สอบที่อยู่ในกลุ่มคาบเส้น (ผู้สอบที่มีความรู้ไม่มากพอที่อาจารย์จะให้สอบผ่านได้อย่างสบายใจ แต่ก็มีความรู้ไม่น้อยจนอาจารย์สามารถตัดสินให้สอบตกได้โดยไม่มีข้อสงสัย) หลังจากนั้นอาจารย์พิจารณาความสามารถโดยรวมของผู้สอบแต่ละคน (โดยไม่ทราบคะแนนที่ผู้สอบคนนั้นได้รับ) แล้วระบุว่าผู้สอบคนใดจัดว่ามีความสามารถอยู่ในเกณฑ์ "คาบเส้น" เมื่อระบุว่าผู้สอบคนใดบ้างจัดว่ามีความสามารถคาบเส้นแล้วให้ตั้งเกณฑ์สอบผ่านที่คะแนน median ของผู้สอบกลุ่มนี้ (ไม่แนะนำให้ใช้ค่าเฉลี่ย (mean) เนื่องจากเกณฑ์ผ่านจะเบี่ยงเบนได้มากหากมีคะแนนที่สูงหรือต่ำมากเข้ามาร่วมในการคำนวณ)
2. Contrasting groups method: การตั้งเกณฑ์ผ่านวิธีนี้เริ่มจากการระบุลักษณะของผู้สอบที่ควรสอบผ่าน และ ผู้ที่ควรสอบตก หลังจากนั้นให้อาจารย์พิจารณาความสามารถของผู้สอบทีละคน (โดยไม่ทราบคะแนนที่ผู้สอบคนนั้นได้รับ) แล้วระบุว่าผู้สอบคนนั้นควรอยู่ในกลุ่ม "สอบผ่าน" หรือ "สอบตก" หลังจากนั้นให้ทำการวาดกราฟแสดงความสัมพันธ์ระหว่างจำนวนนักเรียนที่ถูกจัดให้สอบผ่าน และ สอบตก กับคะแนนที่นักเรียนได้รับ ดังตัวอย่างข้างล่าง



เกณฑ์ผ่านคือคะแนน ณ จุดที่ false positive และ false negative passing เท่ากัน (ในกรณีตัวอย่างนี้คือ 5 คะแนน) (คณะกรรมการตั้งเกณฑ์ผ่านอาจปรับเกณฑ์ผ่านได้เพื่อปรับอัตรา false positive และ false negative passing ได้ตามวัตถุประสงค์ของการสอบ)

รศ. ดร.นพ.เชิดศักดิ์ ไอรมณีรัตน์

หัวข้อ : Item analysis (MCQ MEQ OSCE)

MCQ Item Analysis

Cherdsak Iramaneerat
Department of Surgery
Faculty of Medicine Siriraj Hospital
Mahidol University

MCQ item analysis

Item Analysis

- A group of statistical analyses having two characteristics:
  - —The data consist of actual responses of test takers to individual test items
  - —The primary purpose is to gain information about the items (rather than about test takers)

Livingston SA. Item analysis. In: Downing SM, Haladyna TM. Handbook of test development. Mahwah, NJ: LEA, 2006, p. 421-444.

MCQ item analysis

Objectives

- เมื่อสิ้นสุดการอบรมแล้ว อาจารย์ผู้เข้าอบรมสามารถ
  - อธิบายผลการวิเคราะห์ข้อสอบ MCQ ที่ใช้บ่อยทางแพทยศาสตรศึกษาได้อย่างถูกต้อง
  - นำผลการวิเคราะห์ข้อสอบไปเป็นแนวทางในการพัฒนาคุณภาพของข้อสอบ MCQ ในภาควิชาของตนได้
  - บอกถึงข้อควรระวัง และข้อจำกัดในการวิเคราะห์ผลการสอบ MCQ
  - ประยุกต์ใช้หลักการวิเคราะห์ข้อสอบปรนัยในการวิเคราะห์ข้อสอบประเภทอื่นได้อย่างเหมาะสม

MCQ item analysis

Outline

- MCQ item analysis
  - Item statistics
  - Test statistics
  - Applications
  - Limitations
- Test score analysis in other types of assessment

MCQ item analysis

MCQ Item Analysis

- Item statistics
  - Item difficulty
  - Item discrimination
  - Distractor functionality
- Test statistics
  - Internal consistency reliability
  - Standard deviation and mean
  - Average difficulty
  - Average discrimination

MCQ item analysis

Item Statistics

Looking at individual test items

MCQ item analysis

Item Difficulty

- Proportion of examinees answering an item correctly (p)

$$p = \frac{C}{C + I}$$

C = number of examinees with a correct answer

I = number of examinees with incorrect answers

- Ideal: 0.45 – 0.75
- Good: 0.76 – 0.91
- Acceptable: 0.25 – 0.44
- Problematic: < 0.24 or > 0.91

MCQ item analysis

Item Discrimination

- The ability of an item to discriminate high scorers from low scorers
- Point-biserial correlation (r)

$$r = \frac{M_p - M_q}{SD}\sqrt{pq}$$

Mp	=	Mean score of examinees with a correct answer
Mq	=	Mean score of examinees with incorrect answers
SD	=	Standard deviation of test scores
p	=	Proportion of examinees with a correct answer
q	=	Proportion of examinees with incorrect answers

MCQ item analysis

Point-Biserial Correlation

- The correlation between an item score with the total score
  - **Range: −1.0 – 1.0**
  - **Point–biserial of an item should be positive**
    - Ideal: 0.20 or higher
    - Acceptable: 0.1 – 0.19
    - Problematic: < 0

MCQ item analysis

Distractor Functionality

A functioning distractor is an incorrect option that:

1. Is chosen by at least 5 percent of examinees
2. Has a negative point–biserial correlation with the total score

MCQ item analysis

Example 1

Number 148		Correct answer = 2				
P-VALUE = 0.65	PT BISERIAL =0.1					Total number of examinees
DISTRACTOR	1	2	3	4	5	
N OF PEOPLE	4	158	17	58	5	242
MEAN SCORE	77.25	84.81	81.35	83.86	76.6	
P-VALUE	0.02	0.65	0.07	0.24	0.02	
PT BISERIAL	-0.09	0.1	-0.07	-0.01	-0.11	

MCQ item analysis

Example 2

Number 145		Correct answer = 3				
P-VALUE = 0.79	PT BISERIAL =0.34					Total number of examinees
DISTRACTOR	1	2	3	4	5	
N OF PEOPLE	7	27	190	9	9	242
MEAN SCORE	77	78.11	85.81	78.22	75.89	
P-VALUE	0.03	0.11	0.79	0.04	0.04	
PT BISERIAL	-0.12	-0.21	0.34	-0.11	-0.16	

MCQ item analysis

Example 3

Number 124	Correct answer = 2				
P-VALUE = 0.14	PT BISERIAL =0.14				Total number of examinees
DISTRACTOR	1	2	3	4	5
N OF PEOPLE	8	33	22	133	46	242
MEAN SCORE	87	87.52	78.05	84.3	83.17	
P-VALUE	0.03	0.14	0.09	0.55	0.19	
PT BISERIAL	0.05	0.14	-0.19	0.03	-0.04	

MCQ item analysis

Example 4

Number 112	Correct answer = 3				
P-VALUE = 0.73	PT BISERIAL = -0.05				Total number of examinees
DISTRACTOR	1	2	3	4	5
N OF PEOPLE	0	1	177	1	63	242
MEAN SCORE	0	84	83.74	83	84.92	
PVALUE	0	0	0.73	0	0.26	
PT BISERIAL	0	0	-0.05	-0.01	0.05	

MCQ item analysis

Siriraj Hospital's IA report

No. : 1		p Value : 0.64				r_{pbi} : 0.23			
A		B		C		* D		E	
r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%
0.02	6.98	-0.18	5.08	-0.17	8.57	0.23	63.81	-0.07	15.56

MCQ item analysis

Item Analysis and Option Analysis

Faculty of Medicine Siriraj Hospital
Mahidol University

No. : 1		p Value : 0.64				r_{pbi} : 0.23			
A		B		C		* D		E	
r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%
0.02	6.98	-0.18	5.08	-0.17	8.57	0.23	63.81	-0.07	15.56

No. : 2		p Value : 0.34		r_{pbi} : 0.19					
A		B		C		D		* E	
r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%
0.01	4.76	-0.02	25.40	-0.19	10.79	-0.06	24.76	0.19	33.97

No. : 3		p Value : 0.56		r_{pbi} : 0.35					
A		B		* C		D		E	
r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%
-0.03	8.89	-0.26	23.17	0.35	55.87	-0.05	3.17	-0.16	8.89

No. : 4		p Value : 0.50		r_{pbi} : 0.33					
A		* B		C		D		E	
r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%
-0.15	1.90	0.33	50.48	-0.15	4.13	-0.18	10.48	-0.13	33.02

No. : 5		p Value : 0.24		r_{pbi} : 0.06					
A		B		C		* D		E	
r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%
-0.06	3.49	-0.08	53.02	0.05	12.06	0.06	23.81	0.02	7.62

No. : 6		p Value : 0.53		r_{pbi} : 0.20					
A		B		* C		D		E	
r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%
-0.16	23.17	-0.11	3.81	0.20	53.33	-0.02	5.40	-0.02	14.29

MCQ item analysis

Test Statistics

Looking at the whole test

MCQ item analysis

Reliability

- Consistency of test scores
  - If we test the students again, will they get the same scores?
  - Range: 0 – 1
  - High values: highly consistent test scores

KR-20

$$KR20 = \left(\frac{n}{n-1}\right)\left(1 - \frac{\sum pq}{Var}\right)$$

n = number of items

Var = Variance of the whole test

p = Proportion of people passing the item

q = Proportion of people failing the item

How Much is Enough?

- Depends on test scores uses
  - High-stakes exam: 0.9 or higher
  - Medium-stakes exam: 0.80 – 0.89
  - Low-stakes exam: 0.70 – 0.79

Mean and Standard Deviation

- Effective instruction => All students can do the test well.
  - —High mean scores
  - —Low standard deviation
- High standard deviation: Wide range of students' scores
  - —Some students can solve the problems in the tests, while some students cannot do.
- Too difficult test => Most students fail to get correct answers.
  - —Low mean scores
  - —Low standard deviation

MCQ item analysis

Average Difficulty

- Average of p values of all items on the test
- Small group of students:
  - —Difficult to interpret
  - —Depends on the ability distribution of students
- Large group of students:
  - —Assume a fair sampling of students
  - —Indicates the average difficulty of the whole test

MCQ item analysis

Average Discrimination

- Average point-biserial correlation of the whole test
- Indicates how good the items on the test can differentiate high scorers from low scorers.
- High values generally indicate a good test.
- Effective instruction: All students can do well on the test.
  - —A low value does not necessarily indicate bad items.

MCQ item analysis

Applications

1. Posttest score adjustment
2. Item revision
3. Item pool management
4. Improvement of instruction

MCQ item analysis

Limitations

1. Sample dependency
2. Reliability is the property of test scores, not test items.
3. Numbers are there to serve us, not the other way around.

MCQ item analysis

Constructed Response Items

- Dichotomous => Polytomous scores
- Item analysis
  - Item difficulty
    - Proportion => Percentage
  - Item discrimination
    - Point-biserial correlation => Pearson / Spearman correlation

Constructed Response Items

- Test statistics
  - Reliability
    - KR20 => Cronbach's Alpha
  - Average difficulty: Average percentage
  - Average discrimination: Average correlation

Cronbach's Alpha

- Consistency of test scores: If we test the students again, will they get the same scores?

$$\alpha = \frac{n}{n-1}\left(1 - \frac{\sum \sigma_{x_i}^2}{\sigma_x^2}\right)$$

n	=	number of testlets
σ_x^2	=	score variance of total scores
$\sigma_{x_i}^2$	=	score variance of the i^{th} testlet

MCQ item analysis

OSCE

- Classical test theory
  - —Inter-rater agreement
    - Percentage of agreement between the two
    - Correlation between the two
    - Intraclass correlation
- Item response theory
  - —Multi-faceted assessment

Facets Model

$$P_{mnijk}(X \mid \theta) = \frac{e^{\Sigma(B_n - C_j - D_i - F_m - E_{ik})}}{\Sigma e^{\Sigma(B_n - C_j - D_i - F_m - E_{ik})}}$$

P = Probability of student n being rated by SP (rater) j on skill i of case m with rating category k

B_n = Clinical skills competence of a student n

C_j = Severity level of a SP (rater) j

D_i = Difficulty level of a skill i

F_m = Difficulty level of a case m

E_{ik} = Difficulty of rating k relative to $(k\text{-}1)$ for skill i

Summary

- MCQ item analysis
  - —Item statistics
  - —Test statistics
  - —Applications
  - —Limitations
- Test score analysis in other types of assessment
  - —Constructed response items
  - —OSCE

MCQ item analysis

การวิเคราะห์ข้อสอบปรนัย

อาจารย์ นายแพทย์เชิดศักดิ์ ไอรมณีรัตน์
ภาควิชาศัลยศาสตร์, คณะแพทยศาสตร์ศิริราชพยาบาล, มหาวิทยาลัยมหิดล, กรุงเทพมหานคร ๑๐๗๐๐.

การวิเคราะห์ข้อสอบปรนัย (Item analysis) เป็นการใช้วิธีการทางสถิติเพื่อวิเคราะห์คำตอบที่ผู้สอบตอบข้อสอบปรนัยในการสอบครั้งหนึ่ง เพื่อประเมินว่าข้อสอบที่นำมาใช้ในการสอบครั้งนั้นมีคุณสมบัติอย่างไร ทำงานได้ตามที่ต้องการหรือไม่ มีระดับความยากง่ายของข้อสอบเหมาะสมหรือไม่ มีข้อบกพร่องหรือไม่ และควรได้รับการปรับปรุงแก้ไขอย่างไร[๑] การวิเคราะห์ข้อสอบเป็นศาสตร์ที่ได้รับการพัฒนาอย่างต่อเนื่องมาเป็นเวลานาน มีเทคนิคและวิธีการต่าง ๆ มากมายที่ผู้วิเคราะห์สามารถใช้เพื่อบอกคุณสมบัติของข้อสอบแต่ละข้อ ตั้งแต่วิธีการง่าย ๆ ไปจนถึงวิธีการที่มีความซับซ้อนมาก โดยแต่ละเทคนิคการวิเคราะห์ก็มีจุดประสงค์แตกต่างกันไป ตั้งแต่การบอกระดับความยากง่าย การบอกถึงความสามารถในการแยกผู้สอบที่เก่งออกจากผู้สอบที่ไม่เก่ง ไปจนถึงเทคนิคขั้นสูงที่สามารถบอกได้ว่าข้อสอบมีความลำเอียงต่อผู้สอบเพศใดเพศหนึ่ง หรือผู้สอบจากสถาบันใดสถาบันหนึ่งเป็นพิเศษหรือไม่ มีการเดาข้อสอบมากน้อยเพียงใด ผู้สอบรู้ข้อสอบมาก่อนเข้าสอบหรือไม่ หรือมีความน่าจะเป็นมากน้อยเพียงใดที่ผู้สอบลอกคำตอบ ในบทความนี้ผู้นิพนธ์ไม่ได้ตั้งเป้าประสงค์ที่จะรวบรวมและอภิปรายเทคนิคการวิเคราะห์ข้อสอบทุกวิธีที่มีใช้อยู่ในปัจจุบัน แต่ต้องการเพียงนำเสนอความรู้พื้นฐานที่เกี่ยวกับการวิเคราะห์ข้อสอบและอธิบายถึงวิธีการวิเคราะห์ข้อสอบที่นิยมใช้กันในทางแพทยศาสตรศึกษา โดยเฉพาะในประเทศไทย โดยประสงค์ให้อาจารย์ผู้อ่านสามารถนำเอาความรู้ที่ได้จากบทความนี้ไปใช้แปลผลการวิเคราะห์ข้อสอบที่ตนเกี่ยวข้อง และดำเนินการปรับปรุงคุณภาพของข้อสอบได้อย่างเหมาะสม

ความรู้พื้นฐานเกี่ยวกับข้อสอบปรนัย

ก่อนที่จะกล่าวถึงรายละเอียดในการวิเคราะห์ข้อสอบ ผู้นิพนธ์ก็จะขอทบทวนความรู้พื้นฐานเกี่ยวกับข้อสอบปรนัยก่อน โดยทั่วไปข้อสอบปรนัยแต่ละข้อมีส่วนประกอบสำคัญ ๒ ส่วนด้วยกันคือ

๑. โจทย์ (stem) เป็นข้อมูลของโรค หรือภาวะ หรือผู้ป่วยตามด้วยคำถาม หรือเว้นช่องว่างสำหรับเติมคำหรือข้อความที่เหมาะสมลงไป

๒. ตัวเลือก (options) คือคำ หรือข้อความที่ผู้ออกข้อสอบนำเสนอตามหลังจากโจทย์เพื่อให้ผู้สอบเลือกไปใช้ตอบคำถาม หรือเติมลงในช่องว่างในโจทย์

๒.๑ ตัวเลือกที่ถูกต้อง (correct option) เป็นคำตอบที่ถูกต้องมีเพียงตัวเลือกเดียวต่อข้อสอบข้อหนึ่ง

๒.๒ ตัวลวง (distractors) เป็นคำตอบที่ผิด มีไว้ลวงให้ผู้สอบที่ไม่มีความรู้ หรือมีความเข้าใจไม่ถูกต้องในเนื้อหาที่นำมาออกข้อสอบเลือกตอบ ข้อสอบที่ใช้ในคณะแพทยศาสตร์ศิริราชพยาบาล และที่ใช้ทั่วไปในการสอบของนักศึกษาแพทย์ และแพทย์ประจำบ้านในประเทศไทย นิยมจัดให้มีตัวลวง ๔ ตัวต่อข้อสอบ ๑ ข้อ

ทฤษฎีที่ใช้ในการวิเคราะห์ข้อสอบ

ทฤษฎีที่ใช้ในการวิเคราะห์ข้อสอบในปัจจุบันนั้นมี ๒ ทฤษฎีด้วยกัน ได้แก่ทฤษฎีการสอบแบบดั้งเดิม

(classical test theory) และทฤษฎีการตอบสนองต่อข้อสอบ (item response theory) ทฤษฎีการสอบแบบดั้งเดิมนั้นเป็นทฤษฎีที่ได้ถูกพัฒนาขึ้นตั้งแต่ตอนต้นของศตวรรษที่ ๒๐ โดยมีการรวบรวมเป็นตำราในครั้งแรกตั้งแต่ปี ค.ศ. ๑๙๒๑ โดย William Brown และ Godfrey H Thomson[๒] หลังจากนั้นทฤษฎีนี้ก็ได้รับการใช้อย่างแพร่หลายในการวิเคราะห์ข้อสอบและได้รับการพัฒนาอย่างต่อเนื่อง ทฤษฎีการสอบแบบดั้งเดิมนี้วางรากฐานอยู่บนสมมติฐานว่าคะแนนสอบที่ได้มานั้นประกอบไปด้วยคะแนนที่แท้จริง (true score) กับความผิดพลาดจากการวัด (error) ซึ่งสมมติฐานดังกล่าวต่อมาพบว่ามีข้อจำกัดหลายประการด้วยกัน ในราว ค.ศ. ๑๙๗๐ จึงได้มีความพยายามพัฒนาทฤษฎีที่ใช้ในการวิเคราะห์ข้อสอบแบบใหม่ขึ้นซึ่งใช้หลักการของความน่าจะเป็นมาวิเคราะห์ข้อสอบ ทำให้สามารถแยกผลการวิเคราะห์ข้อสอบแต่ละข้อเป็นอิสระจากข้อสอบข้ออื่นในการสอบเดียวกัน ทฤษฎีใหม่นี้เรียกว่าทฤษฎีการตอบสนองต่อข้อสอบ (item response theory) ทฤษฎีใหม่นี้มีข้อได้เปรียบกว่าทฤษฎีเดิมหลายประการด้วยกัน ได้แก่ ความสามารถในการปรับตัวเข้ากับสถานการณ์ต่าง ๆ (flexibility) ความมีประสิทธิภาพในการใช้ข้อมูล (efficiency) และความสามารถในการวิเคราะห์ถึงคุณภาพของข้อสอบ และผู้สอบโดยละเอียด (in-depth analysis)[๓] จึงเป็นเหตุให้ทฤษฎีการตอบสนองต่อข้อสอบนี้ได้รับความนิยมอย่างกว้างขวางตั้งแต่ในค.ศ. ๑๙๘๐ ในปัจจุบันการสอบต่าง ๆ ได้ถูกวิเคราะห์ด้วยทฤษฎีการตอบสนองต่อข้อสอบนี้มากขึ้นเรื่อย ๆ

เนื่องจากการวิเคราะห์ข้อสอบในวงการแพทยศาสตรศึกษาในประเทศไทยทั้งหมดในปัจจุบันยังใช้เทคนิคต่าง ๆ ตามทฤษฎีการสอบแบบดั้งเดิมอยู่ ดังนั้นผู้นิพนธ์จะขอกล่าวถึงเทคนิคการวิเคราะห์ข้อสอบตามทฤษฎีการสอบแบบดั้งเดิมเท่านั้น เพราะจะเป็นสิ่งที่อาจารย์แพทย์ทุกท่านจะได้พบและใช้งานเป็นประจำ

การวิเคราะห์ข้อสอบตามทฤษฎีการสอบแบบดั้งเดิม

การวิเคราะห์ข้อสอบตามทฤษฎีการสอบแบบดั้งเดิมนี้ประกอบไปด้วย ๒ ส่วนใหญ่ ๆ คือ (๑) การวิเคราะห์ข้อสอบรายข้อ (item analysis) และ (๒) การวิเคราะห์ข้อสอบโดยรวม (test analysis)

๑. การวิเคราะห์ข้อสอบรายข้อ (item analysis)

การวิเคราะห์ข้อสอบแต่ละข้อให้อาจารย์พิจารณา ๓ ปัจจัย คือ

๑.๑ ความยากง่ายของข้อสอบ (item difficulty, p)

ความยากง่ายของข้อสอบวัดโดยใช้ค่า p ซึ่งย่อมาจาก proportion of examinees answering items correctly (สัดส่วนของผู้สอบที่ตอบข้อสอบข้อนั้นถูก) ซึ่งหาได้จากการนำจำนวนผู้สอบที่ตอบข้อสอบข้อนั้นถูกต้องหารด้วยจำนวนผู้สอบที่ตอบข้อสอบข้อนั้นทั้งหมด หากข้อสอบข้อนั้นเป็นข้อสอบที่ง่ายผู้สอบทุกคนตอบถูก ค่า p ก็จะเป็น ๑ หากไม่มีผู้สอบคนใดตอบถูกเลยข้อสอบข้อนั้นก็จะมีค่า p เป็น ๐ หากมีคนตอบถูก ๗๐% ข้อสอบข้อนั้นก็มีค่า p เท่ากับ ๐.๗ ข้อสอบที่ดีมากจะมีค่า p อยู่ในช่วง ๐.๔๕ – ๐.๗๕, ข้อสอบที่ดีจะมีค่า p อยู่ในช่วง ๐.๗๖ – ๐.๙๑, ข้อสอบที่พอใช้ได้มีค่า p อยู่ในช่วง ๐.๒๕ – ๐.๔๔, ข้อสอบที่มีค่า p ต่ำกว่า ๐.๒๕ เป็นข้อสอบที่ยากเกินไป และข้อสอบที่มีค่า p สูงกว่า ๐.๙๑ เป็นข้อสอบที่ง่ายเกินไป[๔-๖]

๑.๒ ความสามารถในการจำแนกผู้สอบตามระดับความสามารถ (item discrimination, r)

ความสามารถในการจำแนกผู้สอบ หมายถึงความสามารถของข้อสอบข้อหนึ่ง ๆ ในการแยกผู้สอบที่ทำคะแนนได้ดี ออกจากผู้สอบที่ทำคะแนนได้ไม่ดี ข้อสอบที่มีความสามารถในการแยกแยะได้ดีนั้นผู้สอบที่ตอบข้อสอบข้อนั้นถูกมักจะได้คะแนนสูง และผู้สอบที่ตอบข้อสอบข้อนั้นผิดมักจะได้คะแนนต่ำ ดัชนีที่ใช้วัดความสามารถในการจำแนกผู้สอบที่ใช้กันมากที่สุดในปัจจุบันคือค่า point-biserial correlation ซึ่งนิยมใช้อักษรย่อเป็น r_{pbis}[๔] ซึ่งสามารถคำนวณได้จากสูตรต่อไปนี้[๗]

$$r = \frac{M_p - M_q}{SD}\sqrt{pq}$$

เมื่อ Mp = คะแนนรวมเฉลี่ยของผู้สอบที่ตอบข้อสอบถูก

Mq = คะแนนรวมเฉลี่ยของผู้สอบที่ตอบข้อสอบผิด

SD = ค่าเบี่ยงเบนมาตรฐาน (standard deviation) ของคะแนนสอบ

p = สัดส่วนของผู้สอบที่ตอบข้อสอบถูกต่อผู้สอบทั้งหมด

q = สัดส่วนของผู้สอบที่ตอบข้อสอบผิดต่อผู้สอบทั้งหมด

ค่า point-biserial correlation ที่คำนวณได้นี้มีค่าอยู่ในช่วง -๑ ถึง ๑ โดยค่าที่ติดลบหมายถึง ข้อสอบข้อนั้นผู้ที่ตอบถูกมักสอบได้คะแนนรวมต่ำ แต่ผู้ที่ตอบผิดมักสอบได้คะแนนรวมสูง ในทางตรงข้าม หากค่า point-biserial ยิ่งสูง แสดงถึงข้อสอบที่มีความสามารถในการแยกแยะดี ผู้ที่ตอบข้อสอบข้อนั้นถูกมักทำคะแนนรวมได้สูง ข้อสอบที่ดีควรมีค่า point-biserial สูงกว่า ๐.๒๐, ข้อสอบที่พอใช้ได้ควรมีค่า point-biserial อยู่ในช่วง ๐.๑ – ๐.๑๙, ข้อสอบที่มีค่า point-biserial ต่ำกว่า ๐.๑ เป็นข้อสอบที่ไม่สู้ดีนัก โดยเฉพาะอย่างยิ่งข้อสอบที่มีค่า point-biserial ต่ำกว่า ๐ ไม่ควรนำมาคิดคะแนน[๔-๖] (โดยทั่วไปแล้วข้อสอบที่มีค่า point-biserial ติดลบ ให้สงสัยว่าจะเฉลยผิด)

๑.๓ ประสิทธิภาพของตัวลวง (distractor functionality)

ตัวลวงที่มีประสิทธิภาพนั้นมีคุณสมบัติ ๒ ประการคือ[๕]

(๑) มีผู้สอบเลือกตัวลวงนั้นไม่ต่ำกว่าร้อยละ ๕ ของจำนวนผู้สอบทั้งหมด

(๒) มีค่า point-biserial correlation ของตัวลวงนั้นเป็นลบ กล่าวคือตัวลวงที่ดีจะลวงให้ผู้สอบที่มีความรู้ไม่ดี (มีคะแนนต่ำ) มาเลือก แต่ไม่ลวงให้ผู้สอบที่มีความรู้ดี (มีคะแนนสูง) มาเลือก หากตัวลวงใดมีค่า point-biserial correlation เป็นบวก ให้ทบทวนข้อสอบข้อนั้นดูว่าอาจจะเฉลยผิดหรือมีคำตอบที่ถูกต้องมากกว่า ๑ ตัวเลือก

ตัวลวงใดที่มีผู้สอบเลือกน้อย หรือลวงให้ผู้ที่มีความรู้ดีมาเลือกจัดเป็นตัวลวงที่ไม่ดี สมควรพิจารณาตัดทิ้งหรือปรับเปลี่ยน

๒. การวิเคราะห์ข้อสอบโดยรวม (test analysis)

การวิเคราะห์ข้อสอบโดยรวมเป็นการพิจารณาว่าเมื่อข้อสอบทั้งชุดทำงานร่วมกันแล้วผลสอบที่ได้ออกมาเป็นอย่างไร มีระดับความยากง่ายเป็นอย่างไร มีการกระจายตัวของคะแนนเป็นอย่างไร มีความน่าเชื่อถือของคะแนนสอบมากน้อยเพียงใด ดัชนีต่าง ๆ ที่ต้องพิจารณาได้แก่

๒.๑ ความเที่ยงตรงของคะแนนสอบ (internal consistency reliability)

การประเมินความเที่ยงตรงของคะแนนสอบเป็นการตรวจสอบว่าคะแนนที่ได้ออกมานั้นมีความน่าเชื่อถือเพียงใด เป็นการตอบคำถามว่าหากนำผู้สอบมาสอบใหม่ในสภาวการณ์เดิม ด้วยข้อสอบที่มีระดับความยากง่ายเท่าเดิม และผู้สอบมีความรู้เท่าเดิมไม่ได้ไปศึกษาหาความรู้เพิ่มเติม จะได้คะแนนสอบเท่าเดิมหรือไม่[๙-๑๑]

ดัชนีชี้วัดความเที่ยงตรงของคะแนนสอบที่นิยมใช้ในการรายงานผลสอบด้วยข้อสอบปรนัยคือค่าสัมประสิทธิ์ อัลฟ่า (Coefficient Alpha) ซึ่งสามารถคำนวณได้จากสูตร[๑๑]

$$\alpha = \frac{n}{n\text{-}1}\left(1\text{-}\frac{\sum \sigma^2_{\chi_i}}{\sigma^2_{\chi}}\right)$$

เมื่อ α = สัมประสิทธิ์ อัลฟ่า (Coefficient Alpha)

n = จำนวนชุดย่อยของข้อสอบที่ทำการแบ่งออกเพื่อหาความเที่ยง

σ^2_{χ} = การกระจายตัว (variance) ของคะแนนรวม

$\sigma^2_{\chi_i}$ = การกระจายตัว (variance) ของคะแนนข้อสอบย่อยชุดที่ i

ค่าสัมประสิทธิ์อัลฟ่านี้มีค่าอยู่ในช่วง ๐ – ๑ ค่าต่ำแสดงว่าคะแนนที่ได้มีความเชื่อถือได้น้อย ไม่แตกต่างไปจากการเดาสุ่ม ค่าสูงแสดงว่าคะแนนที่ได้นั้นมีความน่าเชื่อถือมาก หากทำการทดสอบซ้ำคะแนนที่ได้ก็จะใกล้เคียงเดิม โดยทั่วไประดับของความเที่ยงตรง

ของคะแนนสอบที่ยอมรับได้นั้นขึ้นกับว่าต้องการนำเอาคะแนนสอบไปใช้ทำอะไร หากการตัดสินผลสอบนั้นมีความสำคัญมาก (high-stakes examination) เช่น การตัดสินผลสอบขอรับใบประกอบวิชาชีพเวชกรรม หรือประกาศนียบัตรแพทย์ผู้เชี่ยวชาญเฉพาะสาขา มักต้องการคะแนนสอบที่มีค่าสัมประสิทธิ์ อัลฟ่า ไม่ต่ำกว่า ๐.๙ หากการตัดสินผลสอบนั้นมีความสำคัญปานกลาง (medium-stakes examination) เช่นการสอบลงกองการสอบเลื่อนชั้นเรียน มักต้องการคะแนนสอบที่มีค่าสัมประสิทธิ์ อัลฟ่า อยู่ในช่วง ๐.๘ – ๐.๘๙ หากการตัดสินผลสอบนั้นมีความสำคัญน้อย (low-stakes examination) เช่นการสอบย่อยในชั้นเรียน การสอบแบบ formative assessment มักต้องการคะแนนสอบที่มีค่าสัมประสิทธิ์ อัลฟ่า อยู่ในช่วง ๐.๗ - ๐.๗๙ [๑๒]

ประเด็นสำคัญที่ต้องพิจารณาคือเมื่อได้คะแนนสอบที่มีค่าสัมประสิทธิ์ อัลฟ่าต่ำ จะต้องดำเนินการอย่างไรเพื่อพัฒนาให้การสอบครั้งต่อไปไม่ประสบปัญหาเรื่องความไม่น่าเชื่อถือของคะแนนสอบอีก ปัจจัยหลักที่จะช่วยเพิ่มความเที่ยงตรงของคะแนนสอบปรนัยมี ๓ ปัจจัยด้วยกัน[๙] คือ

(๑) เพิ่มจำนวนข้อสอบให้มากขึ้น ยิ่งมีข้อสอบมากข้อคะแนนที่ได้ก็จะมีความเที่ยงตรงเพิ่มมากขึ้น

(๒) ปรับให้ข้อสอบมีการคละกันของข้อสอบที่ยากและง่ายอย่างเหมาะสม เพื่อปรับให้คะแนนมีการกระจายตัวมากขึ้น หากข้อสอบทั้งชุดประกอบไปด้วยข้อสอบที่ง่ายหมด ผู้สอบเกือบทั้งหมดได้คะแนนสูงมาก จะทำให้มีความแตกต่างของคะแนนน้อย โอกาสที่จะแยกแยะผู้สอบที่มีความรู้ดีออกจากผู้ที่มีความรู้ปานกลาง หรือไม่สู้ดีได้อย่างมั่นใจก็เป็นไปได้น้อย ดังนั้นหากอาจารย์ปรับให้มีการคละกันของข้อสอบยากและง่ายอย่างเหมาะสม ก็จะทำให้ผู้สอบมีระดับคะแนนแตกต่างกันมาก ค่าสัมประสิทธิ์อัลฟ่าก็จะสูงขึ้นด้วย

(๓) ปรับสภาวะแวดล้อมของการสอบให้เหมาะสม กำจัดสิ่งรบกวนสมาธิของผู้สอบให้มากที่สุด เช่น เสียงรบกวน แสงไฟที่ไม่เพียงพอ หรือไฟที่ติด ๆ ดับ ๆ เป็นต้น

๒.๒ การกระจายตัวของคะแนน และคะแนนเฉลี่ย (standard deviation and mean score)

การตรวจดูลักษณะพื้นฐานของคะแนนสอบนี้จะช่วยบอกได้คร่าว ๆ ว่าการเรียนการสอนมีประสิทธิภาพเพียงใด หากอาจารย์สอนได้ดี นักเรียนทั้งชั้นเรียนเข้าใจเนื้อหาดี คะแนนสอบที่ได้ออกมาก็ควรจะไม่กระจายตัวมากนัก (คะแนนเกาะกลุ่มกัน) และคะแนนเฉลี่ยก็ควรจะค่อนข้างสูงเมื่อเทียบกับนักเรียนรุ่นอื่น ๆ หากคะแนนสอบของนักเรียนมีการกระจายตัวมากผิดปกติ แสดงว่าอาจมีปัญหาบางประการในการเรียนการสอนทำให้นักเรียนบางคนมีความรู้ความเข้าใจดี แต่มีนักเรียนบางกลุ่มที่ไม่ค่อยรู้เรื่อง[๔]

๒.๓ ค่าความยากง่ายเฉลี่ยของข้อสอบ (average difficulty)

จากการวิเคราะห์ข้อสอบรายข้อ เราได้ค่าความยากง่ายของข้อสอบแต่ละข้อ (p) เมื่อนำค่า p ของข้อสอบทุกข้อมาหาค่าเฉลี่ย เราก็จะได้ค่าความยากง่ายของข้อสอบทั้งชุด ค่าที่ได้มานี้ใช้เป็นดัชนีชี้วัดว่าข้อสอบทั้งชุดโดยรวมแล้วมีระดับความยากง่ายเป็นอย่างไร หากผู้สอบเป็นนักศึกษากลุ่มใหญ่พอที่เราจะตั้งสมมติฐานว่าระดับความสามารถมีการกระจายตัวอย่างเหมาะสมและไม่ต่างจากระดับความสามารถเฉลี่ยของกลุ่มผู้สอบปีก่อน ๆ เราก็สามารถนำค่าความยากง่ายของข้อสอบทั้งชุดนี้มาเทียบได้ว่าข้อสอบที่นำมาใช้ในปีนี้ยาก หรือง่ายกว่าข้อสอบปีก่อน ๆ ซึ่งอาจารย์อาจนำข้อมูลนี้มาใช้พิจารณาปรับเกณฑ์การตัดเกรดด้วยว่าต้องมีการปรับระดับคะแนนที่ได้เกรดต่าง ๆ หรือไม่ อย่างไร

๒.๔ ค่าความสามารถในการแยกแยะผู้สอบเฉลี่ย (average discrimination)

การนำค่า point-biserial correlation ของข้อสอบทั้งชุดมาหาค่าเฉลี่ย เป็นการบอกคร่าว ๆ ว่าโดยรวมแล้วข้อสอบชุดนี้มีความสามารถในการแยกแยะผู้สอบตามระดับความสามารถเพียงใด ยิ่งได้ค่าสูงก็ยิ่งดี แต่มีข้อควรระวังในการแปลผลในกรณีที่การเรียนการสอนเป็นไปได้ดี และผู้สอบทั้งหมด หรือเกือบทั้งหมดทำคะแนนได้สูง ค่า point-biserial correlation เฉลี่ยของข้อสอบทั้งชุดจะไม่สูงแต่ไม่ได้แปลว่าข้อสอบที่ใช้มีคุณภาพไม่ดี[๔]

การนำผลการวิเคราะห์ข้อสอบไปใช้

ผลการวิเคราะห์ข้อสอบด้วยดัชนีชี้วัดต่าง ๆ ดังกล่าวข้างต้นสามารถนำไปใช้ประโยชน์ได้หลายประการ เช่น

๑. ใช้เป็นประโยชน์ในการปรับแก้คะแนนสอบ

จากผลการวิเคราะห์ข้อสอบจะช่วยชี้แนะให้เราทราบว่าข้อสอบข้อใดน่าจะเฉลยผิด ข้อสอบข้อใดน่าจะมีคำตอบที่ถูกมากกว่า ๑ ตัวเลือก ข้อสอบข้อใดน่าจะมีปัญหาเช่น มีความคลุมเครือในคำถาม หรือตัวเลือกมีความซ้ำซ้อนกัน หรือเนื้อหาของข้อสอบอยู่นอกเหนือไปจากสิ่งที่สอนนักเรียน เป็นต้น ข้อสอบที่มีปัญหาเหล่านี้ต้องได้รับการประเมินโดยคณะกรรมการตรวจข้อสอบซึ่งประกอบไปด้วยอาจารย์ผู้มีความรู้ความชำนาญในเนื้อหาวิชาที่ทำการสอบว่าจะดำเนินการอย่างไรกับการคิดคะแนน หากปัญหาที่พบมีความรุนแรงไม่มากจนทำให้การตัดสินใจเลือกคำตอบที่ถูกต้องเปลี่ยนไป คณะกรรมการอาจพิจารณาคิดคะแนนของข้อสอบข้อนั้นตามปกติ หากข้อสอบเฉลยผิดคณะกรรมการสามารถพิจารณาแก้คำตอบแล้วทำการตรวจให้คะแนนข้อสอบข้อนั้นใหม่ หากข้อสอบข้อใดมีคำตอบที่เหมาะสม ๒ ข้อ คณะกรรมการอาจพิจารณาให้ผู้สอบที่ตอบข้อใดข้อหนึ่งใน ๒ ข้อดังกล่าวได้คะแนนในข้อนั้น หากข้อสอบนั้นมีความคลุมเครือมากจนไม่สามารถตัดสินใจเลือกคำตอบที่เหมาะสมได้ คณะกรรมการสามารถตัดข้อสอบข้อนั้นออกจากการคิดคะแนน และปรับคะแนนเกณฑ์ผ่านลดลงตามความเหมาะสม

๒. ใช้เป็นประโยชน์ในการปรับปรุงคุณภาพข้อสอบ

ภายหลังจากการรายงานคะแนนสอบเป็นที่เรียบร้อยแล้ว คณะกรรมการสอบสามารถนำผลการวิเคราะห์ข้อสอบแต่ละข้อมาพิจารณาโดยละเอียดเพื่อดูว่าข้อสอบข้อใดสมควรได้รับการปรับปรุงแก้ไข ข้อสอบที่พบว่ายากเกินไปอาจเกิดจากโจทย์คำถามมีความคลุมเครือ ต้องทำการปรับแก้ให้โจทย์ชัดเจนขึ้น หรือเพิ่มเติมข้อมูลบางประการเข้าไปเพื่อให้การวินิจฉัยชัดเจนขึ้น ข้อสอบที่พบว่าง่ายเกินไปอาจพิจารณาปรับให้ยากขึ้นโดยการแก้โจทย์หรือตัวเลือก ข้อสอบที่มีค่า point-biserial ต่ำมักเกิดจากโจทย์ที่คลุมเครือ สร้างความสับสนให้ผู้สอบ สมควรได้รับการปรับโจทย์คำถามใหม่

นอกจากนี้อาจารย์ยังต้องพิจารณาถึงการทำงานของตัวเลือกด้วย ปัญหาที่พบบ่อยมากในการวิเคราะห์ข้อสอบปรนัยคือมีตัวลวงจำนวนมากที่ไม่ทำงาน (มีผู้สอบเลือกน้อยมาก หรือลวงเฉพาะผู้ที่มีความรู้ดีให้มาเลือก) จากการศึกษาวิจัยข้อสอบปรนัยจำนวนมากพบว่าข้อสอบส่วนใหญ่มักมีตัวเลือกที่ทำงานจริงเพียง ๓ ตัวเลือกเท่านั้น[๕] ตัวเลือกที่เหลือเป็นตัวเลือกที่ไม่มีประโยชน์ พิมพ์ลงมาในข้อสอบก็เป็นการเปลืองเนื้อที่หน้ากระดาษ และเสียเวลาอ่านโดยใช่เหตุ อาจารย์ควรพิจารณาตัดตัวลวงที่ไม่ทำงานออกเสียหรือเปลี่ยนเป็นตัวลวงอื่นที่น่าจะมีประสิทธิภาพมากขึ้น

๓. ใช้เป็นประโยชน์ในการบริหารคลังข้อสอบ

ข้อสอบแต่ละข้อนั้นได้มาด้วยความยากลำบาก อาจารย์แต่ละท่านต้องใช้เวลาและความคิดอย่างมากเพื่อพัฒนาข้อสอบที่ดีขึ้นมาใช้ ดังนั้นเมื่อนำข้อสอบมาใช้แล้วผลการวิเคราะห์ข้อสอบแสดงว่าข้อสอบข้อใดเป็นข้อสอบที่ดี มีระดับความยากง่ายเหมาะสม มีความสามารถในการจำแนกผู้สอบที่ดีก็ควรจะพิจารณาเลือกเก็บข้อสอบดังกล่าวไว้ในคลังข้อสอบเพื่อที่จะได้นำกลับมาใช้ใหม่ในอนาคต ในการเก็บข้อสอบเข้าในคลังข้อสอบก็ต้องมีการแนบข้อมูลเกี่ยวกับประวัติการใช้งานและผลการวิเคราะห์ข้อสอบในแต่ละครั้งไว้คู่กันด้วย เพื่อที่จะได้เป็นประโยชน์ในการเลือกข้อสอบมาใช้งาน หากอาจารย์ต้องการข้อสอบที่มีระดับความยากง่าย หรือความสามารถในการจำแนกผู้สอบมากน้อยเพียงใดจะได้ดึงเอาข้อสอบที่มีคุณลักษณะตามต้องการออกมาใช้ได้ตามต้องการ

๔. ใช้เป็นประโยชน์ในการพัฒนาคุณภาพการสอน

การพิจารณาผลการวิเคราะห์ข้อสอบโดยละเอียดในหัวข้อที่อาจารย์ท่านใดท่านหนึ่งรับผิดชอบ

ในการสอนนักเรียนหรือแพทย์ประจำบ้านอยู่นั้นจะทำให้ได้ข้อมูลที่เป็นประโยชน์ในการพัฒนาการเรียนการสอนได้ กล่าวคืออาจารย์สามารถตรวจสอบดูได้ว่านักเรียนหรือแพทย์ประจำบ้านมีความเข้าใจที่ถูกต้องในเรื่องดังกล่าวหรือไม่ ประเด็นใดที่มีผู้เข้าใจผิดอยู่มากก็สมควรที่อาจารย์จะทำการเน้นย้ำในบรรดานักเรียนหรือแพทย์ประจำบ้านในการสอนครั้งต่อ ๆ ไปเพื่อแก้ไขความเข้าใจผิดดังกล่าว ประเด็นใดที่นักเรียนหรือแพทย์ประจำบ้านมีความเข้าใจดีมากอยู่แล้ว อาจารย์อาจไม่ต้องใช้เวลามากนักในการสอนเรื่องดังกล่าว แต่เอาเวลามาใช้สอนในเรื่องที่นักเรียนหรือแพทย์ประจำบ้านยังไม่ค่อยเข้าใจให้มากขึ้นได้

ข้อจำกัดของการวิเคราะห์ข้อสอบ

ถึงแม้ว่าการวิเคราะห์ข้อสอบด้วยวิธีการที่ได้อธิบายมาข้างต้นจะให้ข้อมูลที่เป็นประโยชน์หลายอย่างด้วยกัน แต่เนื่องจากวิธีการวิเคราะห์เหล่านี้เป็นเทคนิคที่วางรากฐานอยู่บนทฤษฎีการสอบแบบดั้งเดิม (classical test theory) ซึ่งมีข้อจำกัดหลายประการด้วยกัน ในการนำค่าต่าง ๆ ที่ได้จากการวิเคราะห์ข้อสอบไปใช้นั้น อาจารย์ควรคำนึงถึงข้อจำกัดของผลการวิเคราะห์ด้วย ในที่นี้จะกล่าวถึงเฉพาะข้อจำกัดในการแปลผลการวิเคราะห์ขั้นพื้นฐานเท่านั้นเนื่องจากเป็นการแปลผลที่ใช้กันทั่วไปในวงการแพทยศาสตรศึกษา ข้อจำกัดในการนำผลการวิเคราะห์ไปประยุกต์ในงานวิจัยทางจิตวิทยาการศึกษายังมีอีกหลายประการที่ผู้นิพนธ์ขอไม่นำมากล่าวในที่นี้ เนื่องจากมีความซับซ้อนและไม่มีที่ใช้ในวงการแพทยศาสตรศึกษาในประเทศไทยในปัจจุบัน

พื้นฐานสำคัญที่เป็นข้อจำกัดของผลการวิเคราะห์ข้อสอบด้วยทฤษฎีการสอบแบบดั้งเดิมคือค่าต่าง ๆ ที่ได้มาจากการวิเคราะห์นั้นขึ้นอยู่กับกลุ่มตัวอย่างที่ใช้ในการเก็บข้อมูล[๑๓,๑๔] หากได้ข้อมูลมาจากกลุ่มตัวอย่างที่มีขนาดใหญ่พอและมีการกระจายตัวของระดับความสามารถของผู้สอบที่เหมาะสม ค่าต่าง ๆ ที่ได้ (p, r, coefficient alpha) จะค่อนข้างเที่ยงตรง ปัญหาที่สำคัญในการวิเคราะห์ข้อสอบในโรงเรียนแพทย์คือการสอบจำนวนมากจัดในนักศึกษากลุ่มเล็ก และนักศึกษาแต่ละกลุ่มก็มีการกระจายตัวของระดับความสามารถแตกต่างกัน นักศึกษาบางกลุ่มมีความสามารถสูงกว่านักศึกษากลุ่มอื่น ดังนั้นผลการวิเคราะห์ข้อสอบไม่ว่าจะเป็นค่า p, r, coefficient alpha, mean, หรือ standard deviation อาจจะเปลี่ยนแปลงไปในแต่ละกลุ่มของนักศึกษา ดังนั้นการนำผลการวิเคราะห์ข้อสอบไปใช้ในทางปฏิบัติจึงมีข้อควรระวังดังต่อไปนี้

การพิจารณาว่าข้อสอบยากหรือง่ายโดยใช้ค่า p นั้นเป็นค่าที่ไม่คงที่ ขึ้นอยู่กับกลุ่มผู้สอบ หากนำข้อสอบข้อหนึ่งไปไปใช้กับนักเรียนกลุ่มที่มีความรู้ดี นักเรียนส่วนใหญ่จะทำข้อสอบได้ถูกต้องทำให้ค่า p สูง แต่เมื่อนำข้อสอบข้อเดิมไปใช้กับนักเรียกลุ่มที่ความรู้ไม่ดีนัก สัดส่วนของนักเรียนที่ทำข้อสอบข้อเดียวกันได้ถูกต้องจะลดลงทำให้ค่า p ลดลง นอกจากนี้ในข้อสอบที่เน้นการท่องจำที่เคยใช้แล้ว เมื่อนำกลับมาใช้ใหม่ในนักเรียนกลุ่มใหม่ อาจมีนักเรียนจำนวนหนึ่งที่สามารถตอบข้อสอบถูกได้เนื่องจากรู้ข้อสอบมาก่อนก็จะทำให้ค่า p สูงขึ้นกว่าเดิมได้

การพิจารณาว่าข้อสอบมีความสามารถในการแยกแยะผู้สอบได้ดีเพียงใดโดยใช้ค่า r ก็ประสบปัญหาในลักษณะเดียวกัน กล่าวคือค่า r นั้นขึ้นกับกลุ่มตัวอย่างของผู้สอบ หากกลุ่มผู้สอบมีระดับความรู้ที่ใกล้เคียงกัน มีคะแนนค่อนข้างเกาะกลุ่มกัน เมื่อคิดค่า r ก็จะได้ต่ำ แต่หากใช้ข้อสอบข้อเดิมในกลุ่มผู้สอบที่มาจากหลายสถาบัน มีความแตกต่างกันของระดับความรู้อย่างมาก ก็จะได้ค่า r สูง

ค่าสัมประสิทธิ์อัลฟ่า เป็นค่าที่มีความเฉพาะเจาะจงกับการสอบของนักเรียนกลุ่มใดกลุ่มหนึ่งเท่านั้น หาใช่เป็นคุณสมบัติติดตัวข้อสอบแต่ละข้อไม่ หากข้อสอบชุดหนึ่งทำการสอบกับนักเรียนกลุ่มหนึ่งแล้วพบว่าคะแนนสอบที่ได้มานั้นมีค่าสัมประสิทธิ์อัลฟ่าสูงในระดับที่ต้องการก็ไม่ได้เป็นตัวรับประกันว่าหากนำข้อสอบชุดเดิมนั้นไปทำการสอบกับนักเรียนกลุ่มอื่นจะได้ค่าสัมประสิทธิ์อัลฟ่าที่สูงเช่นเดียวกัน นอกจากนี้ค่าสัมประสิทธิ์อัลฟ่าที่สูงไม่ได้เป็นตัวบอกถึงคุณภาพของข้อสอบรายข้อแต่อย่างใด

ค่าสัมประสิทธิ์อัลฟ่าที่สูงช่วยบอกแค่เพียงว่า

คะแนนสอบในข้อสอบข้อหนึ่งมีความผันแปรไปในทิศทางเดียวกันกับคะแนนสอบในข้อสอบข้ออื่นในการสอบชุดเดียวกัน นั่นคือในข้อสอบชุดที่มีค่าสัมประสิทธิ์อัลฟ่าสูงก็อาจประกอบไปด้วยข้อสอบที่ดี และข้อสอบที่ไม่ดีรวมกันอยู่ ต้องไปตรวจสอบดัชนีชี้วัดคุณภาพของข้อสอบตัวอื่น ๆ ในแต่ละข้ออีกครั้ง

ข้อควรจำในการวิเคราะห์ข้อสอบที่ผู้นิพนธ์ขอย้ำในตอนท้ายของบทความนี้ก็คือค่าดัชนีชี้วัดคุณภาพต่าง ๆ ของข้อสอบที่กล่าวมาทั้งหมดนี้เป็นเพียงตัวช่วยให้อาจารย์เข้าใจข้อสอบดีขึ้นและช่วยแนะแนวทางในการพัฒนาปรับปรุงข้อสอบให้ดีขึ้น ดัชนีเหล่านี้ไม่ใช่คำตัดสินหรือตัวชี้ชะตาของข้อสอบ ไม่มีดัชนีใดที่ได้จากการวิเคราะห์ข้อสอบจะมาทดแทนดุลยพินิจของอาจารย์ไปได้ ดัชนีคุณภาพของข้อสอบไม่ว่าจะคำนวณมาด้วยวิธีการที่ถูกต้องแล้วก็ตามก็เป็นเพียงตัวเลขที่สามารถเกิดความผิดพลาดในการแปลผลได้ดังเช่นการแปลผลการวิเคราะห์ทางสถิติต่าง ๆ บทบาทของอาจารย์ในการวิเคราะห์ข้อสอบคงไม่ใช่การยึดถือตัวเลขดัชนีต่าง ๆ เป็นกฎตายตัว หากแต่ใช้ดัชนีเหล่านี้ช่วยเป็นแนวทางในการพิจารณาข้อสอบ หากดัชนีตัวใดแนะว่าข้อสอบอาจมีปัญหา อาจารย์ก็นำข้อสอบนั้นมาพิจารณากันโดยคณะกรรมการข้อสอบ หากหลังจากการพิจารณาโดยถี่ถ้วนแล้วอาจารย์คิดว่าข้อสอบข้อนั้นเหมาะสมแล้ว ไม่ควรทำการปรับแก้เนื้อหา อาจารย์ก็ยืนยันไปว่าไม่แก้ไข อาจารย์คงไม่ตัดสินการรักษาผู้ป่วยโดยใช้ผลเลือดตัวใดตัวหนึ่งเป็นเกณฑ์โดยไม่พิจารณาอาการและอาการแสดงของผู้ป่วยร่วมด้วย ฉันใดก็ฉันนั้น อาจารย์ไม่ควรตัดสินชะตากรรมของข้อสอบโดยใช้เพียงค่า p หรือ r โดยไม่พิจารณาความเหมาะสมของเนื้อหาโจทย์และตัวเลือกต่าง ๆ ในข้อสอบข้อนั้น

เอกสารอ้างอิง

๑. Livingston SA. Item analysis. In: Downing SM, Haladyna TM, eds. Handbook of test development. Mahwah, NJ: Lawrence Erlbaum Associates; 2006:421-41.

๒. Brown W, Thomson GH. The essentials of mental measurement, 2nd ed. Cambridge, England: University Press; 1921.

๓. Yen WM, Fitzpatrick AR. Item response theory. In: Brennan RL, ed. Educational measurement, 4th ed. Westport, CT: Praeger Publishers; 2006:111-53.

๔. Haladyna TM. Writing test items to evaluate higher order thinking. Boston, MA: Allyn and Bacon; 1997.

๕. Haladyna TM. Writing multiple choice items. Chicago, IL: CAT Inc.; 2003.

๖. Haladyna TM. Developing and validating multiple-choice test items, 3rd ed. Mahwah, NJ: Lawrence Erlbaum Associates; 2004.

๗. Aleamoni LM, Spencer RE. A comparison of biserial discrimination, point biserial discrimination, and difficulty indices in item analysis data. Educ Psychol Meas 1969;29:353–8.

๘. Haladyna TM, Downing SM. How many options is enough for a multiple-choice test item? Educ Psychol Meas 1993;53:999–1010.

๙. Gronlund NE. Assessment of student achievement, 7th ed. Boston: Allyn & Bacon, 2003.

๑๐. Linn RL, Miller MD. Measurement and assessment in teaching, 9th ed. Upper Saddle River, NJ: Prentice Hall, 2004.

๑๑. Haertel EH. Reliability. In: Brennan RL, editor. Educational measurement, 4th ed. Westport, CT: Praeger Publishers; 2006:65-110.

๑๒. Downing SM. Reliability: On the reproducibility of assessment data. Med Educ 2004;38:1006–12.

๑๓. Embretson SE, Reise SP. Item response theory for psychologists. Mahwah, NJ: Lawrence Erlbaum Associates, 2000.

๑๔. Smith EV. Evidence for the reliability of measures and validity of measure interpretation: A Rasch measurement perspective. In: Smith EV. Smith RM, eds. Introduction to Rasch measurement: Theory, models, and applications. Maple Grove, MN: JAM Press, 2004:93–112

โปรแกรมวิเคราะห์ข้อสอบ

รุ่น 2.0

การสอบ : SIID 521 (Basic Sciences)

วันที่ : 22 ธันวาคม 2555

จำนวนข้อสอบ = 120

จำนวนผู้เข้าสอบ = 244

Difficulty Index --> p-value (proportion of students answer item correctly)

$$\text{p-Value} = \frac{\text{number of students answer correctly}}{\text{total number of students answer that item}}$$

Discrimination Index --> D or r-value --> Point-biserial correlation coefficient (r^{pbi})

==========

SCORE STATISTICS

Mean = **68.152** S.D. = **11.915**

Mode = **65** (freq = **14**)

Max = **94** Min = **28**

DIFFICULTY INDEX (p value)

Average (p-bar) = **0.566** Max p = **0.990** Min p = **0.010**

DISCRIMINATION INDEX (D or r value)

Average (D-bar) = **0.244** Max D = **0.680** Min D = **-0.180**

RELIABILITY COEFFICIENT (rtt) = **0.847**
(Kuder-Richardson formula 20)

STANDARD ERROR OF MEASUREMENT (SEM) = **4.655**
(S.D. x SQR(1-rtt))

Item Analysis and Option Analysis

Faculty of Medicine Siriraj Hospital
Mahidol University

No. : 1 p Value : 0.55 r_{pbi} : 0.37

A		B		* C		D		E	
r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%
-0.24	21.31	-0.10	13.52	0.37	54.92	-0.16	6.15	-0.07	4.10

No. : 2 p Value : 0.74 r_{pbi} : 0.00

A		B		C		* D		E	
r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%
0.02	5.33	0.07	11.48	-0.02	1.23	0.00	74.18	-0.09	7.79

No. : 3 p Value : 0.84 r_{pbi} : 0.25

A		* B		C		D		E	
r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%
-0.22	14.34	0.25	84.43	0.01	0.41	0.00	0.00	-0.12	0.41

No. : 4 p Value : 0.68 r_{pbi} : 0.43

A		B		* C		D		E	
r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%
-0.26	8.20	-0.09	8.20	0.43	68.03	-0.06	1.64	-0.29	13.93

No. : 5 p Value : 0.92 r_{pbi} : 0.26

A		B		* C		D		E	
r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%
-0.16	4.10	-0.07	0.41	0.26	91.80	-0.16	2.87	-0.08	0.82

No. : 6 p Value : 0.75 r_{pbi} : 0.30

* A		B		C		D		E	
r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%
0.30	74.59	-0.03	13.93	-0.22	2.87	-0.24	3.69	-0.17	4.92

No. : 7 p Value : 0.99 r_{pbi} : 0.06

A		B		C		D		* E	
r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%
-0.06	0.82	0.00	0.00	0.00	0.00	0.00	0.00	0.06	99.18

No. : 8 p Value : 0.70 r_{pbi} : 0.53

* A		B		C		D		E	
r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%
0.53	70.49	-0.13	1.23	-0.21	5.74	-0.38	17.21	-0.17	5.33

No. : 9 p Value : 0.63 r_{pbi} : 0.19

A		B		C		D		* E	
r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%
0.00	0.41	0.00	0.00	0.01	2.05	-0.19	34.43	0.19	63.11

No. : 10 p Value : 0.90 r_{pbi} : 0.25

* A		B		C		D		E	
r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%
0.25	90.16	-0.09	0.41	-0.22	9.02	-0.08	0.41	0.00	0.00

No. : 11 p Value : 0.54 r_{pbi} : 0.48

A		B		C		* D		E	
r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%
-0.44	31.97	-0.09	4.51	-0.05	8.61	0.48	53.69	-0.06	1.23

No. : 12 p Value : 0.55 r_{pbi} : 0.47

A		* B		C		D		E	
r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%
-0.27	28.28	0.47	54.92	0.00	0.00	-0.24	11.07	-0.16	5.74

No. : 13 p Value : 0.81 r_{pbi} : 0.32

A		B		* C		D		E	
r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%
-0.23	5.33	-0.16	9.84	0.32	81.15	-0.13	3.28	-0.06	0.41

No. : 14 p Value : 0.45 r_{pbi} : 0.39

A		B		C		D		* E	
r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%
-0.22	34.84	-0.09	1.64	-0.17	11.89	-0.08	6.15	0.39	45.49

No. : 15 p Value : 0.73 r_{pbi} : 0.32

A		* B		C		D		E	
r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%
-0.24	2.46	0.32	72.95	-0.17	2.05	-0.17	21.72	-0.07	0.41

No. : 16 p Value : 0.09 r_{pbi} : -0.03

A		B		C		D		* E	
r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%
-0.14	11.89	0.15	70.08	-0.18	3.28	0.08	5.74	-0.03	8.61

No. : 17 p Value : 0.36 r_{pbi} : 0.13

A		B		* C		D		E	
r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%
-0.05	4.10	0.06	22.13	0.13	35.66	-0.07	9.43	-0.12	28.69

No. : 18 p Value : 0.83 r_{pbi} : 0.06

* A		B		C		D		E	
r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%
0.06	82.79	0.01	0.82	-0.05	2.05	-0.10	4.92	0.01	9.43

Item Analysis and Option Analysis

Faculty of Medicine Siriraj Hospital
Mahidol University

No. : 19 p Value : 0.25 r_{pbi} : 0.04

A		B		C		* D		E	
r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%
-0.10	51.23	0.04	13.11	0.00	0.00	0.04	24.59	0.05	11.07

No. : 20 p Value : 0.36 r_{pbi} : 0.55

A		* B		C		D		E	
r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%
-0.21	22.54	0.55	35.66	-0.12	2.46	-0.25	34.43	-0.19	4.92

No. : 21 p Value : 0.81 r_{pbi} : 0.20

* A		B		C		D		E	
r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%
0.20	80.74	-0.07	3.69	-0.13	11.89	-0.05	1.64	-0.11	2.05

No. : 22 p Value : 0.46 r_{pbi} : 0.47

* A		B		C		D		E	
r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%
0.47	45.90	-0.14	6.15	-0.11	4.92	-0.18	17.21	-0.24	25.82

No. : 23 p Value : 0.00 r_{pbi} : -0.06

A		B		* C		D		E	
r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%
-0.03	0.41	0.00	0.41	-0.06	0.41	-0.14	4.10	0.16	94.26

No. : 24 p Value : 0.64 r_{pbi} : 0.40

A		B		* C		D		E	
r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%
-0.08	5.33	-0.16	9.43	0.40	64.34	-0.20	9.02	-0.21	11.89

No. : 25 p Value : 0.61 r_{pbi} : 0.40

A		B		C		* D		E	
r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%
-0.15	2.87	-0.10	13.11	-0.23	14.34	0.40	60.66	-0.19	9.02

No. : 26 p Value : 0.70 r_{pbi} : 0.47

A		B		C		D		* E	
r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%
-0.15	7.38	-0.22	9.84	-0.26	7.79	-0.18	5.33	0.47	69.67

No. : 27 p Value : 0.51 r_{pbi} : 0.35

A		* B		C		D		E	
r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%
-0.15	9.02	0.35	50.82	-0.26	25.82	-0.05	5.33	-0.02	9.02

No. : 28 p Value : 0.50 r_{pbi} : 0.17

* A		B		C		D		E	
r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%
0.17	49.59	-0.17	20.49	-0.03	4.51	-0.04	15.98	0.01	9.43

No. : 29 p Value : 0.75 r_{pbi} : 0.17

A		B		C		D		* E	
r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%
-0.09	14.34	-0.16	3.28	-0.01	2.87	-0.06	4.92	0.17	74.59

No. : 30 p Value : 0.58 r_{pbi} : 0.37

A		B		* C		D		E	
r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%
-0.22	6.15	-0.30	31.15	0.37	57.79	0.05	4.92	0.00	0.00

No. : 31 p Value : 0.86 r_{pbi} : 0.28

* A		B		C		D		E	
r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%
0.28	86.07	-0.05	2.05	-0.21	9.43	-0.10	1.23	-0.17	1.23

No. : 32 p Value : 0.88 r_{pbi} : 0.32

A		B		* C		D		E	
r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%
-0.30	8.20	-0.16	2.87	0.32	87.70	0.03	1.23	0.00	0.00

No. : 33 p Value : 0.44 r_{pbi} : 0.37

A		* B		C		D		E	
r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%
0.09	4.92	0.37	44.26	-0.41	45.08	0.01	2.46	-0.03	3.28

No. : 34 p Value : 0.73 r_{pbi} : 0.25

* A		B		C		D		E	
r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%
0.25	72.54	-0.22	9.02	-0.15	6.15	-0.05	1.23	-0.02	11.07

No. : 35 p Value : 0.45 r_{pbi} : 0.42

A		B		C		D		* E	
r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%
0.06	9.02	-0.18	12.30	-0.38	18.44	-0.06	15.16	0.42	45.08

No. : 36 p Value : 0.68 r_{pbi} : 0.35

A		B		* C		D		E	
r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%
-0.15	4.51	-0.29	16.39	0.35	68.03	-0.04	6.97	-0.07	4.10

Item Analysis and Option Analysis

Faculty of Medicine Siriraj Hospital
Mahidol University

No. : 37 p Value : 0.29 r_{pbi} : -0.02

A		B		C		D		* E	
r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%
-0.05	2.05	0.22	52.05	-0.14	7.38	-0.20	9.84	-0.02	28.69

No. : 38 p Value : 0.75 r_{pbi} : 0.11

* A		B		C		D		E	
r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%
0.11	74.59	-0.11	22.95	-0.14	0.82	0.08	0.82	0.08	0.82

No. : 39 p Value : 0.51 r_{pbi} : 0.23

A		B		* C		D		E	
r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%
-0.02	10.25	-0.21	27.46	0.23	51.23	-0.07	9.02	0.09	1.64

No. : 40 p Value : 0.21 r_{pbi} : 0.13

A		* B		C		D		E	
r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%
0.00	40.57	0.13	20.90	0.00	4.51	0.07	17.62	-0.21	16.39

No. : 41 p Value : 0.42 r_{pbi} : -0.03

A		B		C		* D		E	
r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%
0.02	7.38	0.07	43.03	-0.02	0.41	-0.03	41.80	-0.10	7.38

No. : 42 p Value : 0.79 r_{pbi} : 0.33

A		* B		C		D		E	
r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%
-0.22	5.33	0.33	79.10	-0.20	4.92	-0.02	2.87	-0.15	7.79

No. : 43 p Value : 0.81 r_{pbi} : 0.37

* A		B		C		D		E	
r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%
0.37	80.74	-0.33	14.75	0.01	0.82	-0.14	2.05	-0.07	1.64

No. : 44 p Value : 0.56 r_{pbi} : 0.34

A		B		* C		D		E	
r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%
-0.14	1.64	-0.18	6.56	0.34	55.74	-0.22	20.08	-0.05	15.98

No. : 45 p Value : 0.86 r_{pbi} : 0.39

A		B		C		D		* E	
r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%
-0.16	2.05	-0.11	0.82	-0.04	1.23	-0.33	9.84	0.39	86.07

No. : 46 p Value : 0.81 r_{pbi} : 0.31

A		* B		C		D		E	
r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%
-0.19	10.66	0.31	80.74	-0.09	2.87	-0.15	1.64	-0.15	4.10

No. : 47 p Value : 0.93 r_{pbi} : 0.26

A		* B		C		D		E	
r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%
-0.14	2.46	0.26	93.44	-0.01	0.82	-0.17	1.64	-0.15	1.64

No. : 48 p Value : 0.07 r_{pbi} : -0.20

A		B		C		* D		E	
r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%
-0.20	12.70	-0.08	4.51	-0.18	2.87	-0.20	6.56	0.37	73.36

No. : 49 p Value : 0.95 r_{pbi} : 0.21

A		B		C		* D		E	
r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%
0.00	0.00	0.00	0.00	-0.21	4.92	0.21	95.08	0.00	0.00

No. : 50 p Value : 0.83 r_{pbi} : 0.24

A		B		* C		D		E	
r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%
0.00	0.00	0.00	0.00	0.24	83.20	-0.23	15.98	-0.09	0.82

No. : 51 p Value : 0.76 r_{pbi} : 0.26

* A		B		C		D		E	
r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%
0.26	76.23	-0.14	2.87	-0.04	2.46	0.07	0.41	-0.23	18.03

No. : 52 p Value : 0.70 r_{pbi} : 0.24

A		B		C		* D		E	
r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%
-0.15	0.82	-0.21	11.89	0.01	12.70	0.25	70.08	-0.16	4.51

No. : 53 p Value : 0.51 r_{pbi} : 0.31

A		* B		C		D		E	
r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%
0.02	4.51	0.31	50.82	-0.07	2.05	-0.07	2.87	-0.28	39.75

No. : 54 p Value : 0.37 r_{pbi} : 0.28

A		* B		C		D		E	
r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%
-0.07	9.43	0.28	36.89	-0.19	13.52	-0.09	16.80	-0.04	23.36

Item Analysis and Option Analysis

Faculty of Medicine Siriraj Hospital
Mahidol University

No. : 55 p Value : 0.71 r_{pbi} : 0.25

A		B		C		* D		E	
r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%
-0.18	2.87	-0.20	14.75	-0.08	5.74	0.25	70.90	0.01	5.74

No. : 56 p Value : 0.81 r_{pbi} : 0.29

A		* B		C		D		E	
r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%
-0.02	1.23	0.29	81.15	-0.15	7.38	-0.10	4.92	-0.22	5.33

No. : 57 p Value : 0.26 r_{pbi} : 0.19

A		B		C		* D		E	
r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%
-0.08	6.15	-0.17	29.51	-0.01	15.57	0.19	26.23	0.03	22.54

No. : 58 p Value : 0.66 r_{pbi} : 0.29

A		B		C		* D		E	
r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%
-0.16	25.00	-0.14	2.46	-0.22	0.41	0.29	65.98	-0.14	6.15

No. : 59 p Value : 0.73 r_{pbi} : 0.36

A		B		C		* D		E	
r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%
-0.13	0.82	-0.25	19.67	-0.26	5.33	0.36	73.36	0.10	0.82

No. : 60 p Value : 0.93 r_{pbi} : 0.28

A		B		C		D		* E	
r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%
0.00	0.00	-0.13	4.10	-0.27	2.87	-0.03	0.41	0.28	92.62

No. : 61 p Value : 0.89 r_{pbi} : 0.26

A		B		C		D		* E	
r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%
0.05	0.41	-0.30	2.46	-0.13	5.74	-0.06	2.46	0.26	88.93

No. : 62 p Value : 0.89 r_{pbi} : 0.38

A		B		C		* D		E	
r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%
-0.32	7.38	-0.09	0.82	-0.17	3.28	0.38	88.52	0.00	0.00

No. : 63 p Value : 0.69 r_{pbi} : 0.05

A		B		C		* D		E	
r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%
0.00	0.00	-0.12	1.64	-0.02	29.51	0.05	68.85	0.00	0.00

No. : 64 p Value : 0.81 r_{pbi} : 0.20

A		B		* C		D		E	
r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%
-0.09	0.82	0.05	2.46	0.20	80.74	-0.16	11.89	-0.10	3.69

No. : 65 p Value : 0.68 r_{pbi} : 0.10

A		B		* C		D		E	
r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%
-0.06	9.43	-0.15	1.64	0.10	68.44	-0.04	1.23	-0.01	19.26

No. : 66 p Value : 0.55 r_{pbi} : 0.32

A		B		* C		D		E	
r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%
-0.22	23.36	-0.08	11.48	0.32	54.92	-0.11	6.15	-0.07	4.10

No. : 67 p Value : 0.45 r_{pbi} : 0.29

A		B		C		* D		E	
r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%
-0.20	26.64	-0.07	17.62	-0.05	1.23	0.29	45.49	-0.06	8.61

No. : 68 p Value : 0.28 r_{pbi} : -0.03

A		B		* C		D		E	
r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%
0.02	14.34	0.07	1.64	-0.03	27.87	0.06	10.25	-0.04	45.90

No. : 69 p Value : 0.39 r_{pbi} : 0.37

A		B		C		* D		E	
r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%
-0.05	23.77	-0.07	13.93	-0.22	0.41	0.37	38.93	-0.28	22.95

No. : 70 p Value : 0.25 r_{pbi} : 0.13

A		* B		C		D		E	
r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%
-0.02	7.79	0.13	24.59	-0.10	1.64	0.06	10.66	-0.10	54.92

No. : 71 p Value : 0.80 r_{pbi} : 0.09

* A		B		C		D		E	
r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%
0.09	80.33	-0.03	1.64	-0.13	3.28	0.00	5.74	-0.03	9.02

No. : 72 p Value : 0.65 r_{pbi} : 0.37

A		B		C		D		* E	
r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%
-0.25	6.97	-0.05	6.56	-0.23	20.08	-0.05	1.23	0.37	65.16

รศ. ดร.นพ.เชิดศักดิ์ ไอรมณีรัตน์

หัวข้อ : Grading

GRADING

รศ.นพ. เชิดศักดิ์ ไอรมณีรัตน์
ภาควิชาศัลยศาสตร์ คณะแพทยศาสตร์ศิริราชพยาบาล
มหาวิทยาลัยมหิดล

"A lot of current grading practice is shamefully inadequate. We persist in the use of particular practice not because we've thought about them in any depth, but, rather because they are tradition that has remained unquestioned for years."

Thomas Guskey

Objectives

- เมื่อสิ้นสุดการบรรยายแล้ว ผู้เข้าอบรมสามารถ
  - อธิบายถึงข้อดี ข้อด้อยของการตัดสินผลการเรียนแบบอิงเกณฑ์และอิงกลุ่มได้
  - เลือกใช้วิธีการตัดเกรดที่เหมาะสมกับบริบทของสถาบันในการตัดสินผลการศึกษาของนักศึกษา
  - บอกถึงแนวทางที่จะพัฒนาคุณภาพการตัดสินผลการศึกษาของนักศึกษาในสถาบันและหน่วยงานของตนได้อย่างเหมาะสม

Outline

- What is grading?
- Why do we grade our students?
- How can we grade our students?
- How should we combine test scores?
- What does research tell us about grading?
- Guidelines for fair grading

What is grading?

- Grading is an exercise in professional judgment. It involves the collection and evaluation of evidence on students' achievement or performance over a specified period of time. Through this process, various types of descriptive information and measures of students' performance are converted into grades that summarize students' accomplishments.

Why do we grade our students?

- Functions of grading
  - Instructional uses: Grading system should focus on the improvement of student learning.
    - Clarifies the instructional objectives
    - Indicates the students' strengths and weaknesses
    - Provides information concerning students' development
    - Contributes to the students' motivation
  - Reports to parents
  - Administrative uses
    - Promotion and graduation
    - Awards

How can we grade our students?

- Letter grading system
  - A, B, C, D, F
  - S, U, (H)
- Pass-fail system
- Checklists of objectives
- Descriptive report

Who should receive an A?

- Absolute grading
  - A = 90 – 100 points
  - B = 80 – 89 points
  - C = 70 – 79 points
  - D = 60 – 69 points
  - F = below 60
- Relative grading
  - A = 15 %
  - B = 25%
  - C = 45%
  - D = 10 %
  - F = 5%

Absolute Grading

- Strengths
  - Grades relate directly to student performance
  - All students can obtain high grades
  - Students have clear vision of how to get good grades
- Limitations
  - Standards can be arbitrary.
  - Performance standards tend to vary due to variations in test difficulty, student ability, and instructional effectiveness.

Relative Grading

- Strengths
  - Guarantee a constant proportion of grades in every group of students.
- Limitations
  - The percent of students receiving each grade is arbitrary.
  - The meaning of grades varies with the students' ability.
  - Prevent students from helping each other.
  - Cannot link students' grades to the accomplishment of medical competencies

How should we combine test scores?

- The Department of Anatomy wants to grade M2 students based on 4 paper examinations, each receives 25% weight
  - Ex 1: full score 100, range 40 – 80, SD 10
  - Ex 2: full score 50, range 40 – 45, SD 2
  - Ex 3: full score 50, range 10 – 40, SD 8
  - Ex 4: full score 100, range 70 – 80, SD 5

Standardization of Scores

$$Z = \frac{x - M}{SD}$$

Z = standard score

X = raw score

M = mean

SD = standard deviation

What does research tell us about grading?

- Grading is not essential to instruction.
  - Teachers do not need grades to teach well, and students can learn quite well without them.
- Grades have some value as rewards, but no value as punishments
  - Instead of prompting greater effort, low grades more often cause students to withdraw from learning.
- Grading should be done in reference to learning criteria.
  - Normative grading makes learning a highly competitive activity.

Guidelines for Fair Grading

1. Inform students at the beginning of the course what grading procedures is used.
2. Base grades on student achievement, and achievement only.
3. Base grades on a wide variety of valid assessment data.
4. Use a proper technique to combine scores.
5. If there is no quota limitation, use absolute grading.
6. Review all borderline cases by reexamining all test scores.

Summary

- What is grading?
- Why do we grade our students?
- How can we grade our students?
- How should we combine test scores?
- What does research tell us about grading?
- Guidelines for fair grading

"The time to repair the roof is when the sun is shining."

John F. Kennedy

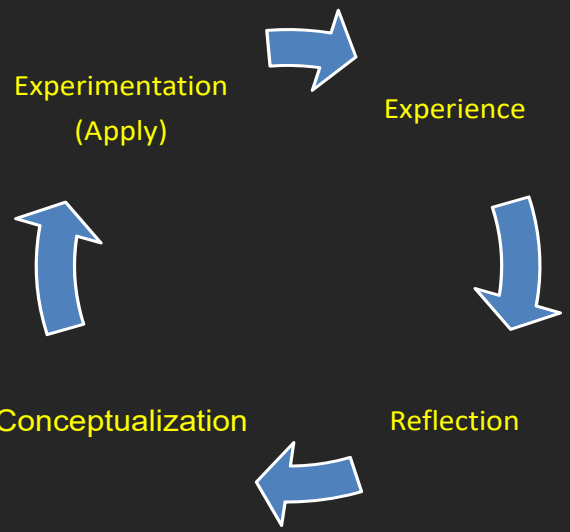
รศ. ดร.นพ.เชิดศักดิ์ ไอรมณีรัตน์

หัวข้อ : Summary

Summary

นพ. เชิดศักดิ์ ไอรมณีรัตน์
ภาควิชาศัลยศาสตร์ คณะแพทยศาสตร์ศิริราชพยาบาล
มหาวิทยาลัยมหิดล

Experiential Learning Theory



Kolb DA. Experiential learning. Englewood cliffs, NJ: Prentice-Hall, 1984.
Schön, D. The Reflective Practitioner, New York: Basic Books, 1983.

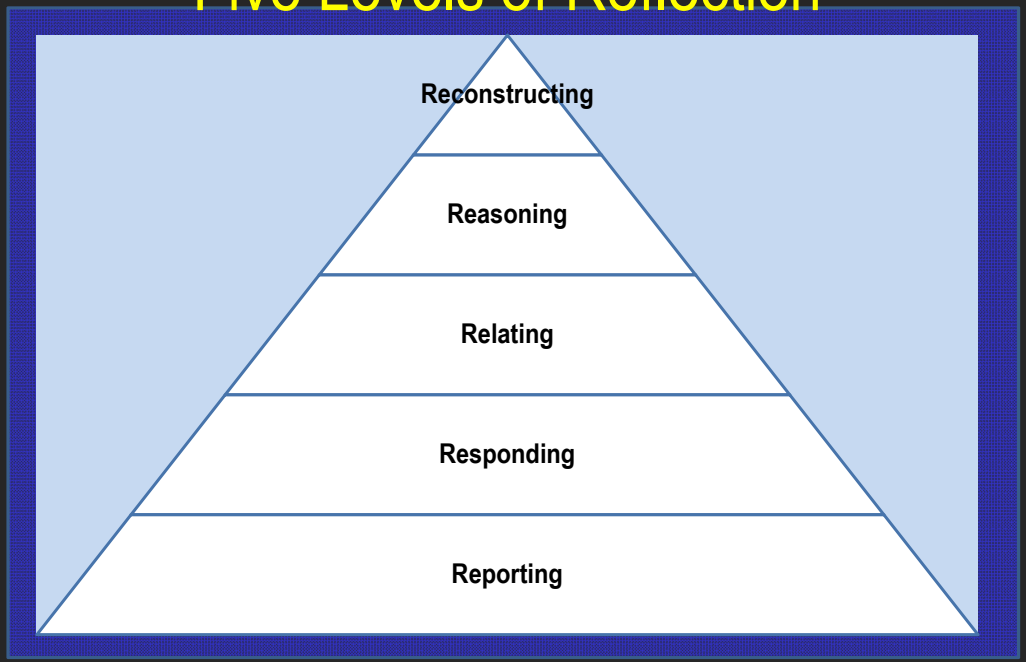
A complex and deliberate process of thinking about and interpreting experience in order to learn from it.

This is a conscious process which does not occur automatically, but is in response to experience and with a definite purpose.

Reflection is a highly personal process, and the outcome is a changed perspective, or learning.

Atkins and Murphy (1995)

Five Levels of Reflection



Bain JD, et al. Reflecting on practice: Student teachers' perspectives, Flaxton, 2002..

Examples

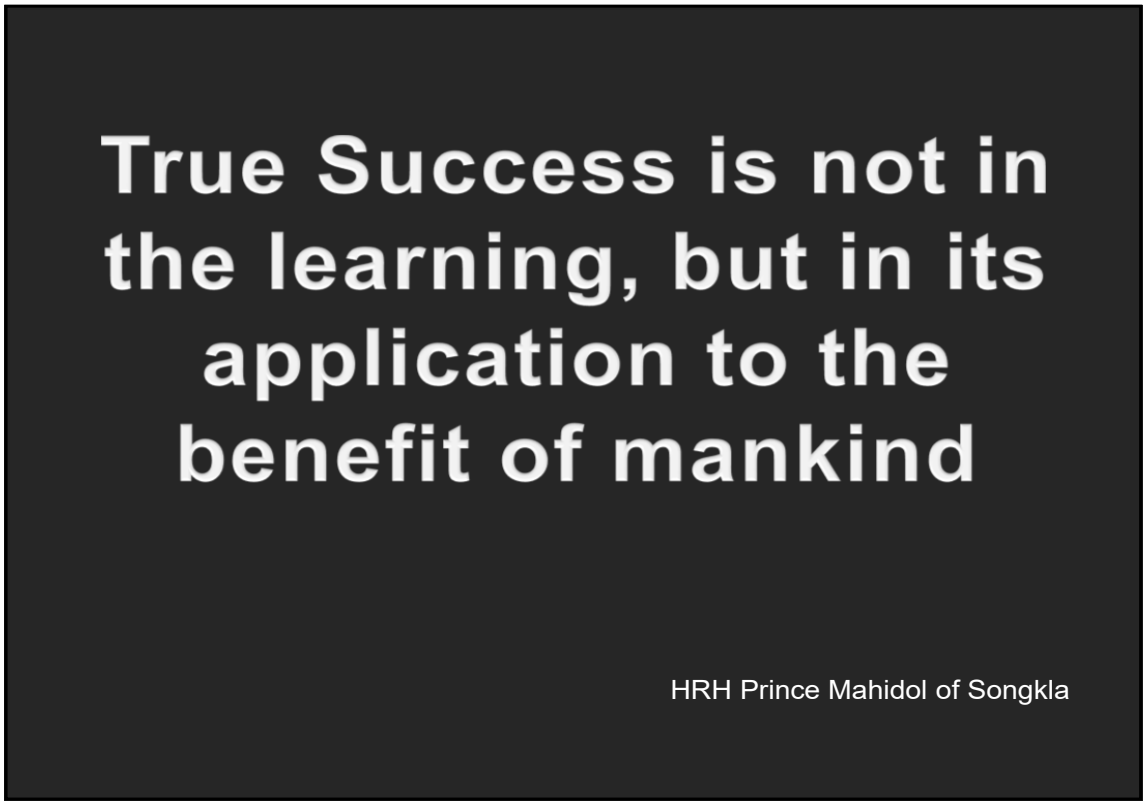
- Reporting: *วันนี้ได้เรียนเรื่อง...*
- Responding: *ฉันรู้สึกชอบแนวคิดเรื่อง...*
- Relating: *...ฉันมีปัญหาเรื่องความเที่ยงคะแนนสอบ ...*
- Reasoning: *เหตุที่คะแนนสอบในวิชาของฉันมีความเที่ยงต่ำเป็นเพราะ... ฉันควรจะแก้ไขโดย...*
- Reconstructing: *ฉันออกแบบการประเมินผลวิธีใหม่โดย...*

Bain JD, et al. Reflecting on practice: Student teachers' perspectives, Flaxton, 2002..

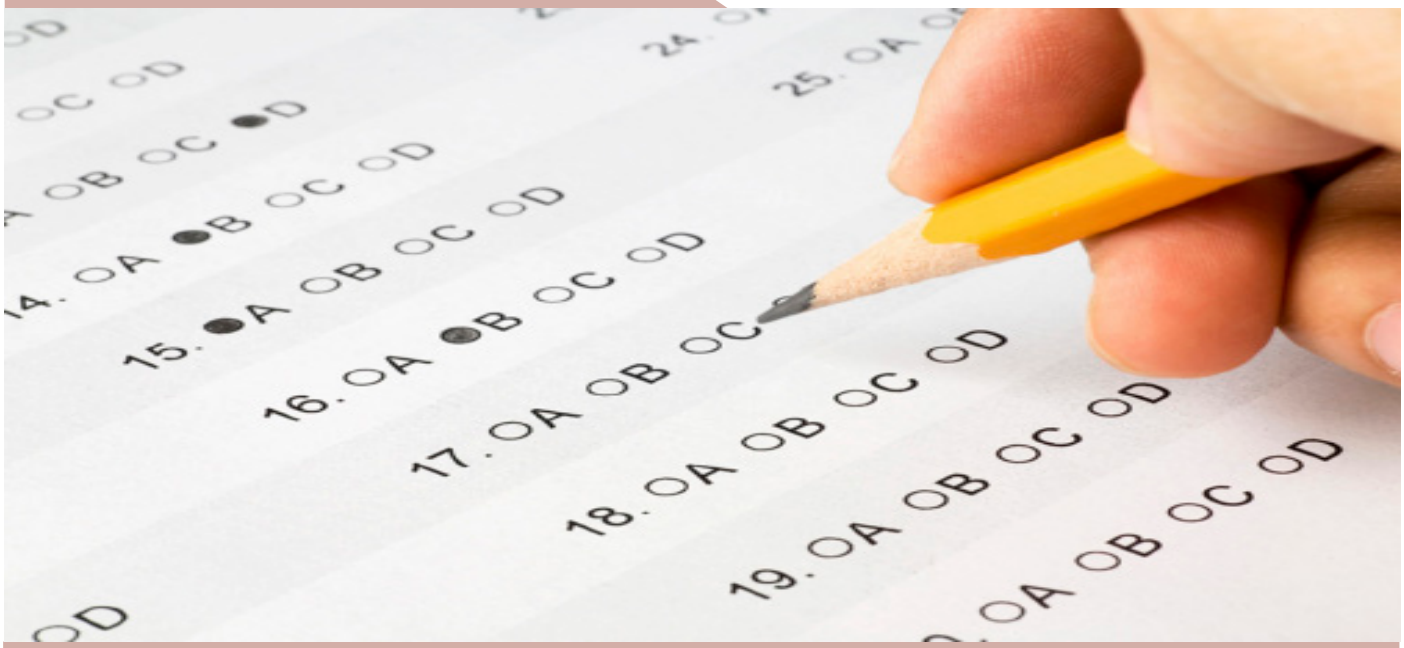
Summary of the Workshop

- Morning
  - Basic principles of assessment
  - Standard setting
- Afternoon
  - Test score analysis
  - Grading


SHEE
Siriraj Health science Education Excellence center
Shee.si.mahidol.ac.th


True Success is not in the learning, but in its application to the benefit of mankind
HRH Prince Mahidol of Songkla

เอกสารประกอบการอบรม



29 October 2020

Part 2 : การพัฒนาข้อสอบ

รศ. ดร.นพ.เชิดศักดิ์ ไอรมณีรัตน์

หัวข้อ : Multiple-choice questions item development

การสร้างข้อสอบปรนัย
MCQ Item Development

รศ.นพ. เชิดศักดิ์ ไอรมณีรัตน์
ภาควิชาศัลยศาสตร์ คณะแพทยศาสตร์ศิริราชพยาบาล
มหาวิทยาลัยมหิดล

Multiple-Choice Questions

- Selected Response Exam
  - True/False
    - Simple True/False items
    - Multiple true/false items (K-type)
  - One best response
    - Standard MCQ
    - Extended matching items

MCQ in Thai Medical Education

- Medical school admission
- Classroom tests
- Comprehensive exam
- National licensing exam steps 1, 2
- Postgraduate exam
  - Basic science exam
  - Board exam

Question

- ในการออกข้อสอบครั้งหนึ่งๆที่อาจารย์ช่วยกันออกข้อสอบกันหลายท่าน อาจารย์จะมั่นใจได้อย่างไรว่าเมื่อนำเอาข้อสอบของอาจารย์ทุกท่านมารวมกันแล้วจะได้เนื้อหาครอบคลุมวัตถุประสงค์การเรียนรู้ของรายวิชานั้นๆ

4

Categorization of the Test Items

1. Nature of the content
2. Nature of learning

5est specification

Cognitive Hierarchy

- Knowledge (รู้)
- Comprehension (เข้าใจ)
- Application (ประยุกต์ใช้)
- Analysis (วิเคราะห์)
- Synthesis (สังเคราะห์)
- Evaluation (ประเมินค่า)

6est specification

A Simplified Cognitive Hierarchy

- **Recall (ความจำ)**
- **Comprehension (ความเข้าใจ)**
- **Application (การประยุกต์ใช้)**

7 Test specification

การสร้างโจทย์และตัวเลือกข้อสอบ

รศ.นพ. เชิดศักดิ์ ไอรมณีรัตน์
ภาควิชาศัลยศาสตร์ คณะแพทยศาสตร์ศิริราชพยาบาล
มหาวิทยาลัยมหิดล

A Good MCQ Item

1. Content
2. Structure

Guidelines for MCQ items

- Content guidelines
- Format guidelines
- Stem guidelines
- Option guidelines

10

Content Guidelines

- Focus on a single idea for each item
- Avoid trivial content
- Avoid opinion-based items
- Avoid direct quotes from textbooks
- Keep item content independent from one another

11

Format Guidelines

- Simplify vocabulary and sentence structures
- Avoid presenting unrelated information, minimize reading time
- Proofread each item for correct grammar, punctuation, and spelling

12

Stem Guidelines

- Make the question as clear as possible
- Avoid using negative words (not, except)
- Place the main idea of an item in the stem, not in options

13

Option Guidelines

- Develop as many effective options as you can
- Vary the location of the correct answers
- Keep options independent
- Keep options homogeneous
- Keep the length of options about the same
- Avoid “none of above” or “all of above”
- Avoid giving clues

14

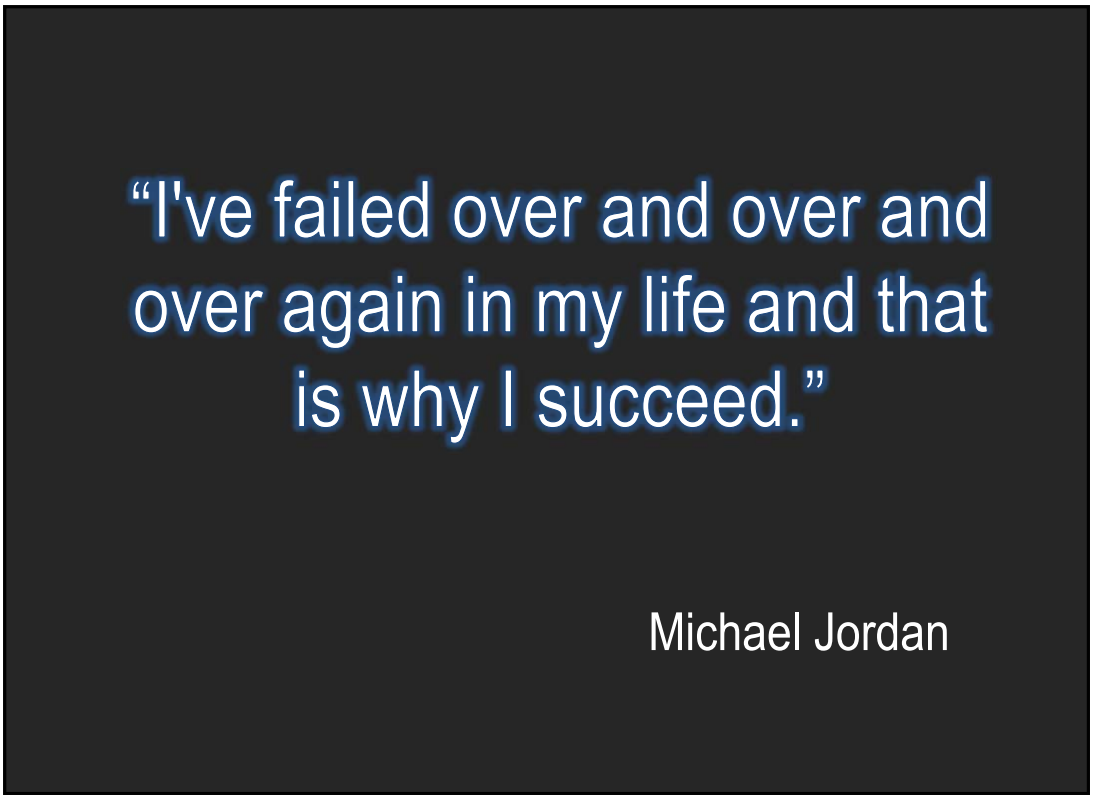
Common Cues

- Grammatical cues
- Logical cues
- Absolute terms
- Long correct option
- Repitition
- Convergence
- Suggestion by other item

Activity

- ให้อาจารย์ทุกท่านสร้างข้อสอบปรนัยสำหรับประเมินความรู้ผู้เรียน จำนวนหนึ่งข้อ
  - ผู้เรียน
  - วัตถุประสงค์
  - โจทย์
  - ตัวเลือก
  - เฉลย
- อภิปรายแนวทางการพัฒนาข้อสอบภายในกลุ่ม

ศูนย์ความเป็นเลิศด้านการศึกษาวิทยาศาสตร์สุขภาพ(ศศว)คณะแพทยศาสตร์ศิริราชพยาบาล Tel. 02-4199978, 02-4196637


“I've failed over and over and over again in my life and that is why I succeed.”
Michael Jordan

การสร้างข้อสอบปรนัย

ผู้ช่วยศาสตราจารย์ นายแพทย์เชิดศักดิ์ ไอรมณีรัตน์

ภาควิชาศัลยศาสตร์, คณะแพทยศาสตร์ศิริราชพยาบาล, มหาวิทยาลัยมหิดล, กรุงเทพมหานคร ๑๐๗๐๐.

ข้อสอบปรนัย (multiple-choice question) เป็นรูปแบบการประเมินผลที่นิยมใช้กันอย่างแพร่หลายในวงการแพทยศาสตรศึกษาเนื่องด้วยคุณสมบัติที่ดีหลายประการด้วยกัน ได้แก่ ประสิทธิภาพในการประเมินความรู้ปริมาณมากในเวลาอันสั้น ผลการประเมินที่ไม่มีผลกระทบจากความรู้สึกส่วนตัวของผู้ตรวจให้คะแนน คะแนนที่มีความเที่ยงสูง รวมถึงผลการวิจัยจำนวนมากที่สนับสนุนความถูกต้องของผลการประเมินด้วยข้อสอบปรนัย[๑-๒] ข้อสอบปรนัยที่พัฒนาขึ้นอย่างดีนั้นสามารถวัดความรู้ได้ทั้งระดับการจดจำ การทำความเข้าใจ และการประยุกต์ความรู้ไปใช้ในการดูแลคนไข้[๓-๕] อย่างไรก็ดีผลการศึกษาวิจัยเกี่ยวกับคุณภาพของข้อสอบปรนัยที่พัฒนาขึ้นใช้ในโรงเรียนแพทย์หลายแห่งพบว่าข้อสอบจำนวนไม่น้อยมีลักษณะที่ไม่เหมาะสม[๕-๖] ข้อสอบปรนัยที่ถูกพัฒนาขึ้นอย่างไม่ถูกหลักการนั้นส่งผลเสียหลายอย่าง เช่นทำให้ข้อสอบยากขึ้นโดยไม่จำเป็น ทำให้ผู้สอบเกิดความสับสน ทำให้ผู้สอบบางกลุ่มเสียเปรียบผู้สอบคนอื่น ทำให้การตัดสินผลสอบผิดพลาด เป็นต้น[๖-๗] ดังนั้นการออกข้อสอบปรนัยที่ดี วางอยู่บนหลักการที่ถูกต้องจึงมีความสำคัญมากในการควบคุมคุณภาพการศึกษาในโรงเรียนแพทย์ บทความนี้จึงถูกเขียนขึ้นเพื่อเป็นการรวบรวมหลักการพื้นฐานในการออกข้อสอบปรนัยที่ได้รับการยอมรับกันทั่วไปในวงการวัดและประเมินผล ผู้นิพนธ์หวังว่าข้อแนะนำต่าง ๆ ที่ได้นำเสนอในบทความนี้จะเป็นแนวทางที่เป็นประโยชน์ในการพัฒนาข้อสอบปรนัยที่มีคุณภาพให้ผู้อ่านไม่มากก็น้อย

รูปแบบพื้นฐานของข้อสอบปรนัย

ข้อสอบปรนัยคือข้อสอบชนิดที่มีคำถามแล้วมีตัวเลือกให้ผู้สอบเลือกตัวเลือกที่เหมาะสมเพื่อตอบคำถามดังกล่าว ข้อสอบปรนัยสามารถแบ่งออกได้เป็น ๒ รูปแบบ ได้แก่

๑. ข้อสอบถูกผิด (True/false item)

ในข้อสอบประเภทนี้จะมีข้อความให้ผู้สอบพิจารณาว่าถูกหรือผิด ในยุคแรกข้อสอบเหล่านี้แต่ละข้อจะแยกเป็นอิสระจากกัน ผู้สอบตัดสินใจว่าข้อความแต่ละข้อถูกหรือผิดโดยไม่เกี่ยวข้องกับข้อความในข้ออื่น ต่อมามีผู้พัฒนาข้อสอบเป็นชุดของข้อความ (multiple true/false หรือ K-type item) โดยในแต่ละข้อจะมีสี่ข้อความ ผู้สอบต้องพิจารณาว่าแต่ละข้อความถูกหรือผิด แล้วทำการเลือกตัวเลือกที่บรรยายจำนวนข้อความที่ถูกต้องได้อย่างเหมาะสม (เช่น ตอบ ก. เมื่อข้อความที่ ๑,๒, และ ๓ ถูกต้อง, ตอบ ข. เมื่อข้อความที่ ๑ และ ๓ ถูกต้อง ฯลฯ)

ข้อสอบชนิดถูกผิดนี้เคยเป็นที่นิยมมากในวงการแพทยศาสตรศึกษาอยู่ระยะหนึ่งเนื่องจากสามารถทดสอบความรู้ได้ปริมาณมาก แต่ข้อสอบชนิดนี้มีข้อจำกัดที่สำคัญคือสามารถใช้ได้เฉพาะกับเนื้อหาที่มีความถูกผิดชัดเจนเท่านั้น ซึ่งการตัดสินใจทางการแพทย์ส่วนมากไม่เป็นเช่นนั้น การตัดสินใจในการวินิจฉัย การตรวจค้นเพิ่มเติม หรือการรักษาผู้ป่วยส่วนใหญ่นั้นแพทย์ตัดสินใจเลือกระหว่างทางเลือกที่แตกต่างกันสามสี่อย่างซึ่งทุกทางเลือกมีความเป็นไปได้ มีส่วนถูก หรือมีความเหมาะสมในบางด้าน

แต่ก็มีความไม่เหมาะสมในด้านอื่นด้วย เช่นการเลือกใช้ยาในผู้ป่วยที่มีการติดเชื้อ นักศึกษาแพทย์มักรู้ว่าควรใช้ยาปฏิชีวนะ ซึ่งยาปฏิชีวนะหลายชนิดก็รักษาการติดเชื้อชนิดนั้น ๆ ได้ แต่นักศึกษาต้องเลือกระหว่างยาที่ล้วนใช้ได้ในการรักษานั้นว่ายาใดที่มีประสิทธิภาพสูงสุด เหมาะสมที่สุดกับชนิดของเชื้อก่อโรคที่พบบ่อยในการติดเชื้อนั้น มีผลข้างเคียงน้อยที่สุด และราคาเหมาะสมด้วย ซึ่งในสถานการณ์นี้ข้อสอบชนิดถูกผิดจะนำมาใช้ได้ยาก ด้วยเหตุนี้ทำให้ข้อสอบชนิดถูกผิดไม่เป็นที่นิยมกันมากนักในปัจจุบัน

๒. ข้อสอบเลือกคำตอบที่ถูกที่สุด (one best response item)

ในข้อสอบประเภทนี้จะมีคำถามแล้วตามด้วยตัวเลือกจำนวนหนึ่งให้ผู้สอบเลือกตัวเลือกที่เหมาะสมที่สุดเป็นคำตอบ ข้อสอบประเภทนี้ที่เป็นที่นิยมกันมากที่สุดคือข้อสอบที่มีตัวเลือก ๔-๕ ตัวเลือก (A-type) แต่นอกจากข้อสอบมาตรฐานนี้แล้วก็มีผู้ใช้ข้อสอบประเภทที่มีลักษณะเป็นการจับคู่ (extended matching item) โดยให้ผู้สอบเลือกตัวเลือกที่เหมาะสม (จากตัวเลือกจำนวนมาก ๘ – ๒๐ ตัวเลือก) ไปจับคู่กับโจทย์ (stem) ซึ่งมีหลายข้อ เช่นจับคู่ระหว่างคำบรรยายอาการของผู้ป่วยจำนวน ๕ – ๑๐ ราย กับการวินิจฉัยโรคที่เหมาะสม จำนวน ๑๕ โรค เป็นต้น

เนื่องจากข้อสอบชนิดที่มีใช้กันแพร่หลายในวงการแพทยศาสตรศึกษาในประเทศไทยในปัจจุบันคือข้อสอบประเภทที่มีตัวเลือก ๔-๕ ตัวเลือก (A-type) ผู้นิพนธ์จะขอเน้นหลักการสำหรับการออกข้อสอบประเภทนี้เป็นสำคัญ

องค์ประกอบของข้อสอบปรนัยชนิดเลือกคำตอบที่ถูกที่สุด

ข้อสอบปรนัยแต่ละข้อมีส่วนประกอบสำคัญ ๒ ส่วนด้วยกันคือ

๑. โจทย์ (stem) เป็นข้อมูลของโรค หรือภาวะหรือผู้ป่วยตามด้วยคำถาม หรือเว้นช่องว่างสำหรับเติมคำหรือข้อความที่เหมาะสมลงไป

๒. ตัวเลือก (options) คือคำ หรือข้อความที่ผู้ออกข้อสอบนำเสนอตามหลังจากโจทย์เพื่อให้ผู้สอบเลือกไปใช้ตอบคำถาม หรือเติมลงในช่องว่างในโจทย์

๒.๑ ตัวเลือกที่ถูกต้อง (correct option) เป็นคำตอบที่ถูกต้องมีเพียงตัวเลือกเดียวต่อข้อสอบข้อหนึ่ง

๒.๒ ตัวลวง (distractors) เป็นคำตอบที่ผิด หรือไม่เหมาะสม มีไว้ลวงให้ผู้สอบที่ไม่มีความรู้ หรือมีความเข้าใจไม่ถูกต้องในเนื้อหาที่นำมาออกข้อสอบเลือกตอบ ตัวลวงไม่จำเป็นต้องเป็นคำตอบที่ผิดชัดเจนเสมอไป ตัวลวงที่ดีมักมีส่วนถูกบ้าง แต่มีระดับของความถูกต้องเหมาะสมน้อยกว่าคำตอบที่ถูก

ข้อแนะนำพื้นฐานของการเขียนข้อสอบปรนัย

มีผู้เชี่ยวชาญทางการประเมินผลให้ข้อแนะนำจำนวนมากในการเขียนข้อสอบปรนัย เคยมีผู้รวบรวมไว้ถึง ๔๓ ข้อ[๒,๙] ในที่นี้ผู้นิพนธ์ขอนำเสนอเฉพาะข้อแนะนำที่ได้รับการยอมรับอย่างกว้างขวางและสามารถประยุกต์ใช้ได้ชัดเจนในการพัฒนาข้อสอบทางการแพทย์ โดยจะทำการจัดหมวดหมู่ของข้อแนะนำเหล่านี้ออกเป็น ๔ กลุ่มด้วยกัน ได้แก่ (๑) เนื้อหาข้อสอบ, (๒) การจัดรูปแบบข้อสอบ, (๓) การเขียนโจทย์, และ (๔) การเขียนตัวเลือก

๑. เนื้อหาข้อสอบ

๑.๑ ข้อสอบหนึ่งข้อควรมุ่งเน้นประเมินความรู้เพียงเรื่องเดียว

ก่อนเริ่มเขียนข้อสอบอาจารย์ผู้ออกข้อสอบควรตั้งวัตถุประสงค์ให้ชัดเจนว่าต้องการประเมินความรู้ของผู้สอบในเรื่องใด และเขียนโจทย์เพื่อตอบสนองวัตถุประสงค์ดังกล่าวเท่านั้น เนื่องจากเนื้อหาวิชาทางการแพทย์มีมาก อาจารย์แต่ละท่านเมื่อทำการสอนไปแล้วจึงอยากจะทดสอบความรู้ในหลายเรื่องที่ตนได้สอนไป แต่กลับมีโควต้าจำกัดในการออกข้อสอบ ทำให้อาจารย์จำนวนไม่น้อยเขียนข้อสอบหนึ่งข้อถามทั้งเรื่องการวินิจฉัยโรค การตรวจค้นเพิ่มเติม การรักษาโรค และ ภาวะแทรกซ้อนของโรค ไปพร้อมกัน ลักษณะข้อสอบเช่นนี้ไม่ควรใช้ เพราะมักซับซ้อนเกินไป เมื่อผู้สอบตอบข้อสอบผิด ก็ไม่สามารถวินิจฉัยได้ว่าผู้สอบขาดความรู้ ความเข้าใจในเรื่องใด

๑.๒ หลีกเลี่ยงการถามความรู้ในรายละเอียดปลีกย่อยที่ไม่มีที่ใช้ทางคลินิก (trivial content)

องค์ความรู้ทางการแพทย์นั้นมีปริมาณมาก ไม่มีผู้ใดที่จดจำเนื้อหาที่มีในตำรา หรือวารสารทางการแพทย์ได้ทั้งหมด แม้ว่าองค์ความรู้หลายเรื่องมีความน่าสนใจ แต่มีประโยชน์ในการประยุกต์ใช้ทางคลินิกค่อนข้างน้อย องค์ความรู้ดังกล่าวจัดเป็นรายละเอียดปลีกย่อย (trivial content) ซึ่งไม่แนะนำให้ทำการทดสอบ สิ่งที่ควรทำการประเมินคือความสามารถในการประยุกต์ใช้ความรู้ในทางคลินิก (application of knowledge) ไม่แนะนำการทดสอบวัดความสามารถในการจดจำเป็นหลัก อย่างไรก็ตามการที่แนะนำให้ออกข้อสอบที่เน้นการประยุกต์ใช้ความรู้ ไม่ได้หมายความว่าการแก้ปัญหาผู้ป่วยนั้นไม่ต้องใช้ความจำเลย ตรงกันข้ามการจดจำเนื้อหาเป็นพื้นฐานที่สำคัญในการแก้ปัญหาทางคลินิก ผู้สอบย่อมต้องจำเนื้อหาได้บ้าง จึงจะประยุกต์องค์ความรู้ดังกล่าวไปแก้โจทย์ปัญหาที่นำเสนอได้

๑.๓ หลีกเลี่ยงการถามความรู้ในเรื่องที่ยังมีความขัดแย้งกันในแนวทางปฏิบัติ (controversy)

ความรู้ทางการแพทย์ในหลายหัวข้อยังเป็นเรื่องที่ผู้เชี่ยวชาญยังมีความเห็นแตกต่างกัน ผู้ป่วยรายเดียวกันไปพบแพทย์สองคนอาจได้รับการรักษาที่แตกต่างกันซึ่งวิธีการรักษาทั้งสองวิธีก็มีงานวิจัยสนับสนุนด้วยกันทั้งคู่ อย่างนี้จัดว่ายังคงมีความขัดแย้ง (controversy) ในเรื่องดังกล่าวอยู่ เนื้อหาในลักษณะนี้ไม่ควรนำมาออกสอบด้วยข้อสอบปรนัย เนื่องจากในขณะที่ทำข้อสอบอยู่นั้น ผู้สอบไม่มีทางรู้ได้เลยว่าอาจารย์ผู้ออกข้อสอบอ้างอิงจากตำราหรือบทความวิชาการใด เนื้อหาที่ยังมีความขัดแย้ง ที่ผู้เชี่ยวชาญจากต่างสถาบันมีแนวทางในการปฏิบัติที่ต่างกันนี้แนะนำให้ใช้ข้อสอบในรูปแบบอื่นในการทดสอบเช่น ข้อสอบอัตนัย เป็นต้น

๑.๔ หลีกเลี่ยงการลอกประโยคหรือข้อความจากตำราโดยตรง

ดังได้กล่าวแล้วว่าข้อสอบที่ดีควรมุ่งเน้นการประเมินความเข้าใจ หรือ การประยุกต์ใช้ความรู้ ไม่ควรออกข้อสอบที่ประเมินความสามารถในการจำรายละเอียดปลีกย่อย การออกข้อสอบโดยวิธีการเปิดตำราแล้วคัดลอกประโยคจากตำราโดยตรงมักจะลงเอยด้วยข้อสอบที่ทดสอบความจำว่าผู้สอบท่องเนื้อหาในตำราตรงส่วนนั้นได้หรือไม่ ข้อสอบที่ดีควรได้จากการดูผู้ป่วย โจทย์ที่ดีควรเป็นปัญหาของผู้ป่วยที่พบในการทำงานนั่นเอง ตัวเลือกก็ได้จากข้อผิดพลาดที่นักศึกษาหรือแพทย์ประจำบ้านมักปฏิบัติกับผู้ป่วยแล้วทำให้ผลการรักษาไม่ดีนั่นเอง

๑.๕ หลีกเลี่ยงการนำเสนอข้อสอบที่ประเมินความรู้ในเรื่องเดียวกันสองข้อในข้อสอบชุดเดียวกัน

เนื่องจากเนื้อหาวิชาที่ต้องทำการประเมินในการสอบแต่ละครั้งนั้นมีมาก ดังนั้นองค์ความรู้ในแต่ละเรื่องแต่ละโรคจึงมักมีสัดส่วนของข้อสอบที่จะออกได้เพียงหนึ่งหรือสองข้อเท่านั้น การที่อาจารย์ออกข้อสอบในเรื่องหรือโรคเดียวกันซ้ำสองข้อในชุดข้อสอบเดียวกันจึงมักเป็นการลดโอกาสในการประเมินความรู้เรื่องอื่นซึ่งก็มีความสำคัญเช่นกัน การออกข้อสอบที่ดีนั้นควรต้องครอบคลุมวัตถุประสงค์การเรียนรู้ตามที่กำหนดในหลักสูตร หรือในเกณฑ์มาตรฐานผู้ประกอบวิชาชีพเวชกรรมอย่างสมดุล การที่จะบรรลุเป้าหมายดังกล่าวได้นั้นต้องเริ่มต้นจากการกำหนดสัดส่วนข้อสอบสร้างเป็นตารางกำหนดจำนวนข้อสอบ (table of specification) เมื่ออาจารย์ได้รับมอบหมายให้ออกข้อสอบควรต้องตรวจสอบให้ชัดเจนว่าเนื้อหาที่ต้องออกข้อสอบนั้นอยู่ในส่วนใดของตารางดังกล่าว การออกข้อสอบซ้ำซ้อนในเนื้อหาเรื่องเดียวกันเป็นสัญญาณบอกว่าอาจไม่ได้สร้างข้อสอบตามข้อกำหนดในตาราง นอกจากนี้การมีโจทย์สองข้อประเมินความรู้เรื่องเดียวกันมีความเป็นไปได้สูงที่เนื้อหาในข้อสอบข้อหนึ่งอาจบอกคำตอบในข้อสอบอีกข้อหนึ่งได้

๒. การจัดรูปแบบข้อสอบ

๒.๑ เลือกใช้คำศัพท์หรือรูปประโยคที่ง่ายต่อการทำความเข้าใจ

อาจารย์ผู้ออกข้อสอบต้องระลึกไว้เสมอว่าข้อสอบที่อาจารย์ออกเพื่อใช้ในการประเมินผลนักศึกษาแพทย์หรือแพทย์ประจำบ้านนั้นมีวัตถุประสงค์เพื่อทดสอบความรู้ทางการแพทย์เป็นสำคัญ มิใช่การประเมินความรู้ทางภาษาศาสตร์ ดังนั้นการเขียนข้อสอบของอาจารย์ควรเลือกใช้รูปแบบประโยคที่ง่ายต่อการทำความเข้าใจ อย่าเขียนประโยคซับซ้อนที่มีความยาวประโยคละหลายบรรทัด มุ่งเน้นให้ภาษาเป็นสื่อในการนำเสนอความคิดของอาจารย์ผู้ออกข้อสอบไปยังผู้สอบ อย่าให้

ภาษาเป็นอุปสรรคในการสื่อสาร การจะเลือกใช้ภาษาใดในการเขียนข้อสอบนั้นให้พิจารณาตามข้อกำหนดขององค์กรหรือหน่วยงานที่ควบคุมการสอบที่อาจารย์ส่งข้อสอบไปให้ใช้ ข้อสอบที่ใช้ในระดับการศึกษาหลักสูตรแพทยศาสตรบัณฑิตทั้งในระดับคณะ หรือข้อสอบที่ใช้ในการสอบระดับประเทศในปัจจุบันยังนิยมใช้ข้อสอบที่เขียนด้วยภาษาไทยโดยมีการใช้ศัพท์เทคนิคเป็นภาษาอังกฤษเหมือนดังภาษาที่แพทย์ใช้สื่อสารกันในการทำงานปกติ ส่วนข้อสอบในระดับหลังปริญญามีหลายการสอบที่ภาควิชา หรือราชวิทยาลัยที่เกี่ยวข้องกำหนดให้ใช้ภาษาอังกฤษทั้งหมด ก่อนที่อาจารย์จะสร้างข้อสอบต้องมีการศึกษาข้อกำหนดของแต่ละการสอบให้ดี

๒.๒ หลีกเลี่ยงการนำเสนอข้อมูลที่ไม่เกี่ยวข้องกับการแก้ปัญหาของโจทย์ข้อนั้น

โจทย์แต่ละข้อควรเขียนให้กระชับ ไม่ยาวเยิ่นเย้อโดยไม่จำเป็น นำเสนอเฉพาะข้อมูลที่จำเป็นในการแก้ปัญหาโจทย์ดังกล่าว อาจารย์บางท่านนำเสนอข้อมูลเยอะมากในโจทย์หนึ่งข้อ บางครั้งข้อสอบข้อหนึ่งมีความยาวถึงครึ่งหน้า โดยให้เหตุผลว่าเป็นเหมือนสถานการณ์จริงที่แพทย์ต้องตัดสินใจบนข้อมูลทางคลินิกปริมาณมาก แพทย์ต้องพิจารณาเองว่าข้อมูลใดสำคัญกับการแก้ปัญหาโจทย์ข้อนั้น ๆ แต่อาจารย์ก็ต้องไม่ลืมว่าเวลาที่ผู้สอบมีในการทำข้อสอบแต่ละข้อนั้นมีจำกัด ในการสอบทางการแพทย์ในประเทศไทยส่วนใหญ่ผู้สอบจะมีเวลาราว ๑ นาทีในการทำข้อสอบ ๑ ข้อ หากเนื้อหาโจทย์ข้อใดมีความยาวมาก ผู้สอบจำนวนไม่น้อยจะเลือกที่จะข้ามข้อสอบข้อนั้นไปก่อนด้วยเกรงว่าจะเสียเวลาอ่านและคิดแก้ปัญหาในข้อนั้นนานเกินไปทำให้ทำข้อสอบไม่ทัน ดังนั้นหากอาจารย์ต้องการให้ข้อสอบที่อาจารย์เขียนขึ้นมานั้นได้ถูกใช้จริง และผู้เข้าสอบได้คิดแก้ปัญหาจริงในการสอบ ไม่ถูกอ่านข้ามไป อาจารย์ควรเขียนข้อสอบให้กระชับ ไม่นำเสนอข้อมูลที่ไม่เกี่ยวข้องกับการแก้ปัญหา

๒.๓ จัดให้มีการตรวจสอบเนื้อหา คำศัพท์ และรูปประโยคที่ใช้ในข้อสอบแต่ละข้อก่อนนำไปใช้

ถึงแม้ว่าอาจารย์ผู้เขียนข้อสอบจะได้มีการอ่านทวนสิ่งที่ตนเองเขียนแล้วเข้าใจเนื้อหาได้ดีและคิดว่าข้อสอบอยู่ในรูปแบบที่สามารถนำไปใช้ได้แล้ว ก็ไม่ควรนำข้อสอบข้อนั้นไปใช้สอบเลย ควรให้มีคณะกรรมการข้อสอบซึ่งประกอบไปด้วยอาจารย์หลายท่านช่วยกันตรวจสอบและพิจารณาปรับแก้ข้อสอบทุกข้อก่อนนำไปใช้จริงเสมอ เนื่องจากผู้เขียนข้อสอบย่อมเข้าใจสิ่งที่ตนเขียนเสมอ แต่เมื่อผู้อื่นอ่านแล้วอาจพบว่ามีเนื้อหาที่กำกวมหรือเข้าใจโจทย์ต่างออกไปได้ การปรับแก้เนื้อหาที่มีความกำกวม หรือเฉลยซึ่งอาจารย์บางท่านอาจไม่เห็นด้วยให้ได้ข้อสอบที่มีความชัดเจน และอาจารย์ทุกท่านยอมรับในคำเฉลยได้ก่อนจะนำข้อสอบไปทำการสอบจริงย่อมเป็นสิ่งที่ดีกว่าการตรวจพบปัญหาหลังจากสอบเสร็จแล้วซึ่งต้องมาตัดสินใจกันอีกว่าจะทำอย่างไรกับการคิดคะแนนของข้อสอบข้อดังกล่าว

๓. การเขียนโจทย์

๓.๑ เขียนโจทย์ให้มีความชัดเจน ผู้สอบทุกคนอ่านแล้วมีความเข้าใจตรงกัน

ข้อแนะนำนี้อาจดูเหมือนตรงไปตรงมา แต่กลับเป็นปัญหาที่พบบ่อยมากในการพัฒนาข้อสอบปรนัย ประเด็นสำคัญคือโจทย์ที่ดีนั้นต้องมีความสมบูรณ์ในตัวเองโดยไม่ต้องอาศัยตัวเลือก โจทย์ข้อสอบที่ดีนั้นเมื่ออ่านโจทย์เสร็จแล้ว หากผู้สอบมีความรู้ในเรื่องที่ทำการประเมินนั้น เขาจะบอกคำตอบได้โดยไม่จำเป็นต้องอ่านตัวเลือกเลย ดังนั้นเมื่ออาจารย์เขียนข้อสอบเสร็จแล้วแนะนำให้ลองปิดตัวเลือกแล้วอ่านเฉพาะโจทย์ดู หากอาจารย์อ่านแล้วบอกได้ว่าโจทย์ถามอะไรและบอกได้ว่าควรตอบอะไรโดยไม่ต้องอ่านตัวเลือกจัดว่าข้อสอบข้อดังกล่าวมีโจทย์ที่มีความชัดเจน

๓.๒ เรียบเรียงเนื้อหาให้ใจความสำคัญของข้อสอบอยู่ในโจทย์

เนื่องจากข้อสอบปรนัยมีตัวเลือกที่อาจารย์ต้องสร้างขึ้นหลายตัวเลือก บางครั้งอาจารย์ผู้พัฒนาข้อสอบอาจเผลอเรอนำเอาใจความสำคัญไปใส่ไว้ในตัวเลือก ซึ่งทำให้เนื้อหาในโจทย์ขาดสาระสำคัญ อ่านโจทย์แล้วไม่เข้าใจว่าผู้ออกข้อสอบต้องการถามความรู้เรื่องอะไร ตัวอย่างข้อสอบที่ไม่เป็นไปตามข้อแนะนำนี้คือข้อสอบที่ถามว่า ข้อใดต่อไปนี้ถูกต้อง หรือข้อใดต่อไปนี้ไม่ถูกต้อง แล้วเขียนรายละเอียดเกี่ยวกับโรค หรือการรักษาบางอย่างในตัวเลือกแต่ละข้อ ข้อสอบในลักษณะนี้มักทำให้

ผู้สอบต้องอ่านข้อสอบย้อนไปมาหลายรอบกว่าจะเข้าใจจุดประสงค์ของข้อสอบ แล้วจึงตัดสินใจเลือกคำตอบ โดยทั่วไปแนะนำให้อาจารย์นำเสนอรายละเอียดต่าง ๆ ไว้ในตัวโจทย์ให้มากที่สุด ส่วนตัวเลือกเขียนเป็นคำหรือข้อความสั้น ๆ

๓.๓ หลีกเลี่ยงการเขียนโจทย์ที่มีรูปประโยคเป็นเชิงปฏิเสธ

โจทย์ที่ดีไม่ควรอยู่ในประโยคเชิงปฏิเสธ เช่น ถามถึงสิ่งที่เป็นข้อยกเว้น สิ่งที่ไม่ควรปฏิบัติ สิ่งที่พบน้อยที่สุด หรือสิ่งที่ไม่น่านึกถึงเป็นต้น งานวิจัยส่วนใหญ่พบว่าข้อสอบที่มีโจทย์ในรูปแบบปฏิเสธเหล่านี้มีระดับความยากง่ายไม่ต่างจากข้อสอบอื่น ๆ แต่งานวิจัยบางชิ้นพบว่าข้อสอบที่มีโจทย์ในรูปแบบปฏิเสธมีความยากมากกว่าข้อสอบอื่นชัดเจนโดยเฉพาะในข้อสอบวัดความรู้ระดับสูง[๑๐-๑๒] แต่ผู้เชี่ยวชาญในการประเมินผลส่วนใหญ่มีความเห็นพ้องกันว่าข้อสอบประเภทนี้สามารถสร้างความสับสนให้กับผู้สอบได้ จึงไม่แนะนำให้ใช้ แต่หากอาจารย์ผู้ออกข้อสอบมีความจำเป็นต้องใช้ข้อสอบที่มีการใช้คำปฏิเสธในโจทย์ แนะนำให้พิมพ์คำปฏิเสธให้เด่นชัด โดยใช้ตัวหนาและขีดเส้นใต้เพื่อให้ผู้สอบเห็นชัด[๑๐]

๔. การเขียนตัวเลือก

๔.๑ เขียนตัวเลือกที่มีประสิทธิภาพให้มีจำนวนมากที่สุดเท่าที่เหมาะสมกับบริบท

เรื่องจำนวนตัวเลือกที่เหมาะสมนี้เป็นเรื่องที่ผู้เชี่ยวชาญด้านการประเมินผลจำนวนมากสนใจ มีงานวิจัยเกี่ยวกับเรื่องจำนวนตัวเลือกที่เหมาะสมในข้อสอบปรนัยอยู่มากมาย[๙] อาจารย์ผู้ออกข้อสอบส่วนมากจะคุ้นเคยกับข้อสอบปรนัยชนิดที่มีห้าตัวเลือก บ่อยครั้งที่อาจารย์ออกข้อสอบแล้วนึกตัวเลือกได้เพียงสามหรือสี่ตัว จึงเกิดคำถามว่าจำเป็นต้องมีตัวเลือกครบห้าตัวเลือกหรือไม่ งานวิจัยบางชิ้นพบว่าการลดจำนวนตัวเลือกลงทำให้ข้อสอบง่ายขึ้น[๑๓-๑๔] แต่งานวิจัยบางชิ้นพบว่าการลดจำนวนตัวเลือกลงทำให้ได้ข้อสอบยากขึ้น[๑๕-๑๖] ผู้เชี่ยวชาญในการประเมินผลเสนอว่าข้อสอบปรนัยที่มีตัวเลือกเพียงสามตัวเลือกก็สามารถทดสอบความรู้ได้อย่างมีประสิทธิภาพ[๑, ๙-๑๐, ๑๗] แต่มีอาจารย์จำนวนไม่น้อยที่ไม่สบายใจที่มีตัวเลือกในข้อสอบแต่ละข้อน้อยกว่าห้าตัวเลือกด้วยกังวลว่าจะทำให้มีโอกาสสูงที่ผู้สอบที่ไม่มีความรู้จะเดาสุ่มได้คำตอบที่ถูกต้อง แต่จากข้อมูลที่ปรากฏในปัจจุบันพบว่าผู้สอบในการสอบในระดับสูงนั้นพฤติกรรมการเดาสุ่มโดยที่ผู้สอบปราศจากความรู้นั้นน่าจะมีบทบาทน้อยมาก ผู้สอบส่วนใหญ่มักพอมีความรู้บ้างและสามารถตัดตัวเลือกที่ไม่สมเหตุสมผลอย่างชัดเจนได้[๑] ในการศึกษาข้อสอบปรนัยส่วนใหญ่พบตัวเลือกที่ไม่ทำงานเป็นจำนวนไม่น้อย[๑๘] ข้อมูลที่ได้จากการวิเคราะห์ข้อสอบปรนัยที่ใช้ในทางแพทยศาสตรศึกษาในประเทศไทยหลายครั้งก็สอดคล้องกับงานวิจัยในต่างประเทศที่พบว่าข้อสอบส่วนใหญ่มักมีตัวเลือกที่ทำงานจริงราวสามหรือสี่ตัวเลือก มีข้อสอบน้อยข้อมากที่ตัวเลือกทั้งห้าตัวเลือกทำงานอย่างมีประสิทธิภาพ

ด้วยข้อมูลจากการศึกษาต่าง ๆ ข้อแนะนำในการออกข้อสอบปรนัยในปัจจุบันคือให้อาจารย์เขียนจำนวนตัวเลือกมากที่สุดที่มีความเหมาะสมกับเนื้อหาโจทย์ ไม่จำเป็นต้องเขียนตัวเลือก ๕ ตัวเลือกเสมอไป เนื่องจากตัวเลือกที่ห้าที่เขียนขึ้นเพื่อเติมเต็มโดยไม่สมเหตุสมผลนั้นมักไม่ค่อยมีคนเลือก หากเนื้อหาที่อาจารย์นำมาสอบมีตัวเลือกที่เหมาะสมเพียงสามหรือสี่ตัวเลือกก็เขียนจำนวนตัวเลือกเพียงสามหรือสี่ตัวเลือก[๑๐] แต่อย่างไรก็ตามให้อาจารย์ศึกษาข้อกำหนดของแต่ละการสอบที่อาจารย์เกี่ยวข้องด้วย เนื่องจากนโยบายของแต่ละการสอบแตกต่างกันไป องค์กรที่จัดสอบทางแพทยศาสตรศึกษาจำนวนไม่น้อยยังคงตั้งข้อกำหนดให้ใช้ข้อสอบ ๕ ตัวเลือกเสมอ ซึ่งหากอาจารย์ไม่ทำตามข้อกำหนดดังกล่าวข้อสอบที่ออกไปอาจไม่ได้รับการพิจารณาได้

๔.๒ จัดให้ตัวเลือกที่ถูกต้องมีการกระจายตำแหน่งไปให้มีจำนวนพอ ๆ กันในทุกตัวเลือก

ข้อแนะนำนี้มีวัตถุประสงค์เพื่อป้องกันไม่ให้ผู้สอบที่ตอบแบบเดาสุ่มแบบเลือกตัวเลือกเดียวกันทั้งหมดสอบผ่านได้ด้วยความบังเอิญ หากอาจารย์สร้างข้อสอบที่มีสี่ตัวเลือก เป็น ก ข ค ง อาจารย์ก็ต้องกระจายให้ตัวเลือกที่ถูกมีทั้งข้อ ก ข ค และ ง ในสัดส่วนที่ใกล้เคียงกัน

๔.๓ เขียนตัวเลือกแต่ละข้อให้เป็นอิสระ ไม่ขึ้นต่อกัน

ในการเขียนตัวเลือกของข้อสอบแต่ละข้อ อาจารย์ต้องระมัดระวังให้ตัวเลือกแต่ละตัวเลือกไม่มีความซ้ำซ้อนกัน เช่นตัวเลือก ก เป็นยากลุ่มย่อยของตัวเลือก ข ตัวเลือก ก เป็นช่วงอายุ ๒ – ๑๐ ปี ตัวเลือก ข เป็นช่วงอายุ ๕ – ๑๑ ปี เป็นต้น การเขียนตัวเลือกที่ซ้ำซ้อนกันนี้ หากเกี่ยวเนื่องกับตัวเลือกที่ถูกต้องอาจมีผู้สอบแย้งว่ามีตัวเลือกที่ถูกต้องมากกว่าหนึ่งตัวเลือก หากตัวเลือกที่ซ้ำซ้อนกันนี้ไม่เกี่ยวกับคำตอบที่ถูก ก็จะทำให้ผู้สอบบางส่วนสามารถตัดตัวเลือกบางตัวเลือกได้โดยไม่ต้องมีความรู้ทางการแพทย์ในเรื่องดังกล่าวได้

๔.๔ เขียนตัวเลือกให้ทุกตัวเลือกมีความเป็นเนื้อเดียวกัน (homogeneous)

การเขียนตัวเลือกให้มีความเป็นเนื้อเดียวกันนั้นหมายถึง ตัวเลือกแต่ละตัวมีรูปร่างหน้าตาและรายละเอียดไปในทิศทางหรือเรื่องราวเดียวกัน หรือเป็นของกลุ่มเดียวกัน การเป็นเนื้อเดียวกันนี้ครอบคลุมตั้งแต่รูปร่างหน้าตา (ตัวเลือกทุกตัวเป็นภาษาแบบเดียวกัน หากตัวเลือกตัวหนึ่งเป็นคำ ตัวเลือกอื่น ๆ ก็ควรเป็นคำ ไม่ใช่วลี หรือประโยค, ตัวเลือกหนึ่งเป็นคำนาม ตัวเลือกอื่นก็เป็นคำนามเหมือนกัน ไม่ใช่กิริยา หรือคำคุณศัพท์) และเนื้อหา (โจทย์ถามการรักษา ตัวเลือกทุกตัวก็เป็นการรักษา ไม่ใช่บางตัวเป็นการตรวจค้นเพิ่มเติม, ตัวเลือกหนึ่งเป็นยาปฏิชีวนะ ตัวเลือกอื่น ๆ ก็น่าจะเป็นยาปฏิชีวนะเช่นกันไม่ใช่ยาเคมีบำบัด หรือยาต้านเชื้อรา) การที่มีตัวเลือกที่ไม่เข้าพวก ไม่มีความเป็นเนื้อเดียวกันกับตัวเลือกอื่นเป็นคำบอกใบ้ในการตัดตัวเลือกที่ผู้สอบนิยมใช้มาก ดังนั้นอาจารย์ผู้ออกข้อสอบควรหลีกเลี่ยง

ในบางบริบทของการดูแลรักษาผู้ป่วย สิ่งที่แพทย์ต้องตัดสินใจเลือกอาจมีทั้งการเลือกที่จะให้การรักษาเลยหรือจะส่งตรวจค้นเพิ่มเติมก่อน ในกรณีนี้อาจารย์สามารถเขียนตัวเลือกที่มีการรักษาและการตรวจเพิ่มเติมปะปนกันได้ แต่การเขียนรูปประโยคคำถามต้องไม่เป็นการบอกใบ้ว่าจะไปทิศทางใด แต่ต้องเลือกใช้คำถามที่เป็นกลาง เช่น ท่านจะปฏิบัติต่อผู้ป่วยอย่างไร, ท่านจะดำเนินการอย่างไรต่อไป เป็นต้น

๔.๕ เขียนตัวเลือกแต่ละข้อให้มีความยาวพอ ๆ กัน

จากการสังเกตข้อสอบปรนัยจำนวนมากจะพบว่าตัวเลือกที่ถูกต้องมักมีความยาวมากกว่าตัวเลือกอื่น ซึ่งข้อสังเกตนี้ผู้สอบจำนวนไม่น้อยก็ทราบดี และผู้สอบส่วนมากเมื่อไม่ทราบคำตอบก็มักเลือกตัวเลือกที่มีความยาวมากที่สุด ดังนั้นอาจารย์ผู้ออกข้อสอบควรระมัดระวังไม่ให้ตัวเลือกตัวใดตัวหนึ่งมีความยาวแตกต่างไปจากตัวเลือกอื่นชัดเจน เพราะจะทำให้ผู้สอบเดาคำตอบที่ถูกได้ง่าย

๔.๖ หลีกเลี่ยงการใช้ตัวเลือก “ถูกทุกข้อ” หรือ “ไม่มีข้อใดถูก”

ตัวเลือก “ถูกทุกข้อ” เป็นตัวเลือกที่ผู้เชี่ยวชาญในการประเมินผลส่วนใหญ่เห็นสอดคล้องกันว่าไม่ควรใช้เนื่องจากมักช่วยใบ้ตัวเลือกที่ถูกต้องให้กับผู้สอบ ทำให้ผู้สอบส่วนหนึ่งตอบถูกโดยไม่ต้องอาศัยองค์ความรู้ที่สมบูรณ์ในเรื่องที่ทดสอบ งานวิจัยพบว่าข้อสอบที่มีตัวเลือกชนิดนี้จะมีผลให้ค่าความเที่ยงของคะแนนสอบลดลง[๑๐] จึงแนะนำให้หลีกเลี่ยงการใช้

ตัวเลือก “ไม่มีข้อใดถูก” เป็นประเด็นที่ผู้เชี่ยวชาญในการประเมินผลยังคงถกเถียงกันอยู่บ้าง ผู้เชี่ยวชาญบางส่วนเห็นว่าไม่ควรใช้ตัวเลือกประเภทนี้ แต่ผู้เชี่ยวชาญบางส่วนให้ความเห็นว่าสามารถใช้ได้ในบางกรณี[๑๐] เหตุผลที่ตัวเลือกชนิดนี้เป็นปัญหาคือการใช้ตัวเลือกนี้มักสร้างความลำบากใจให้กับผู้สอบในการเลือกคำตอบที่ถูกในกรณีที่ตัวเลือกแต่ละตัวเลือกไม่ถูกหรือผิดชัดเจน เพราะผู้สอบจะต้องทำการเปรียบเทียบตัวเลือกที่นำเสนอในข้อสอบกับทางเลือกอื่น ๆ ที่เขานึกได้[๙] หากโจทย์ถามว่ายาใดที่ควรให้แก่ผู้ป่วย แล้วมีชื่อยาสี่ชนิด และมีตัวเลือก “ไม่มีข้อใดถูก” นอกจากที่ผู้สอบต้องนึกว่าในบรรดายา ที่ปรากฏในตัวเลือกนั้นเหมาะสมหรือไม่แล้วเขายังนึกต่อไปอีกว่ามียาอื่นใดที่สามารถให้ในผู้ป่วยรายนี้ได้อีก หากเขานึกออกว่ามียาอื่นที่น่าจะเหมาะสมกับผู้ป่วยมากกว่ายาในตัวเลือก (ด้วยเหตุผลที่อาจแตกต่างไปจากที่อาจารย์ผู้ออกข้อสอบคิด) เขาก็จะเลือก “ไม่มีข้อใดถูก”

การใช้ตัวเลือก “ไม่มีข้อใดถูก” จะยิ่งเป็นปัญหามากขึ้นในข้อสอบที่ถามถึงสิ่งที่ไม่ควรทำ เช่นยาใดไม่ควรใช้ในผู้ป่วย ซึ่งนอกจากยาที่นำเสนอในตัวเลือกแล้วย่อมมียาชนิดอื่นอีกมากมายในบัญชียาที่ไม่เหมาะสม ซึ่งไม่มี

ทางที่ใครจะรู้ได้ว่าการที่ผู้สอบเลือกตอบ “ไม่มีข้อใดถูก” นั้นเขาคิดถึงยาใด และยานั้นไม่เหมาะสมมากไปกว่ายาที่มีอยู่ในตัวเลือกหรือไม่ งานวิจัยทั้งหมดที่ศึกษาถึงตัวเลือกชนิดนี้ได้ข้อสรุปที่ตรงกันว่าข้อสอบที่ใช้ตัวเลือกประเภทนี้เพิ่มระดับความยากให้ข้อสอบ[๑๖] โดยทั่วไปแล้วจึงไม่แนะนำให้ใช้ตัวเลือกประเภทนี้ในการสอบทางแพทยศาสตรศึกษาซึ่งทางเลือกสำหรับสถานการณ์ที่นำเสนอมีได้มากและการตัดสินใจเลือกคำตอบต้องอาศัยการเปรียบเทียบข้อดีข้อเสียของแต่ละตัวเลือก

สรุป

ในบทความนี้ผู้นิพนธ์ได้กล่าวถึงข้อแนะนำขั้นพื้นฐานในการพัฒนาข้อสอบปรนัยชนิดเลือกคำตอบที่ถูกที่สุดโดยสรุปข้อแนะนำเหล่านี้ออกเป็นสี่กลุ่มด้วยกัน ได้แก่ (๑) เนื้อหาข้อสอบ, (๒) การจัดรูปแบบข้อสอบ, (๓) การเขียนโจทย์, และ (๔) การเขียนตัวเลือก ผู้นิพนธ์หวังว่าข้อแนะนำเหล่านี้คงพอเป็นแนวทางสำหรับอาจารย์แพทย์ในการพัฒนาข้อสอบปรนัยที่มีคุณภาพเพื่อใช้ในการประเมินนักศึกษาแพทย์และแพทย์ประจำบ้านได้บ้าง อย่างไรก็ตามบทความนี้เป็นการกล่าวถึงข้อแนะนำเบื้องต้นเท่านั้น ยังมีข้อแนะนำอื่น ๆ ที่ผู้นิพนธ์ไม่ได้นำมารวบรวมไว้ในบทความนี้เพื่อต้องการทำให้เนื้อหากระชับ โดยข้อแนะนำอื่น ๆ ที่ผู้นิพนธ์ไม่ได้กล่าวถึงนี้พบว่าเป็นปัญหาน้อยในการออกข้อสอบทางการแพทย์ หรือเป็นข้อแนะนำที่ไม่ได้รับการสนับสนุนอย่างกว้างขวางจากผู้เชี่ยวชาญทางการวัดและประเมินผล หากผู้อ่านสนใจรายละเอียดของข้อแนะนำอื่น ๆ ที่มีผู้กล่าวไว้สามารถศึกษาเพิ่มเติมได้จากเอกสารอ้างอิงที่แสดงไว้ท้ายบทความ

มีข้อควรพิจารณาในการประยุกต์ใช้ข้อแนะนำเหล่านี้ในการพัฒนาข้อสอบที่ผู้นิพนธ์ขอกล่าวถึงประการหนึ่งคือ แม้ว่าข้อแนะนำที่กล่าวถึงเหล่านี้หลายข้อมีการศึกษาวิจัยสนับสนุนที่ชัดเจน แต่สิ่งเหล่านี้ก็เป็นเพียงข้อแนะนำว่าผู้ออกข้อสอบควรปฏิบัติ ไม่ใช่กฎเกณฑ์ตายตัว การเขียนข้อสอบปรนัยนั้นเป็นงานที่ต้องอาศัยทั้งศาสตร์และศิลปะผสมผสานกันอย่างเหมาะสม หาใช่สูตรคณิตศาตร์ที่ไม่มีข้อยกเว้น ผู้นิพนธ์ไม่คาดหวังให้อาจารย์ผู้พัฒนาข้อสอบยึดข้อแนะนำเหล่านี้เสมือนกฎเกณฑ์ตายตัวที่ต้องทำตามในทุกกรณี หากแต่ต้องการให้อาจารย์ใช้เป็นแนวทางในการสร้างข้อสอบ ในบางบริบทผู้ออกข้อสอบอาจเลือกที่จะไม่ปฏิบัติตามข้อแนะนำบางประการได้บ้าง แต่การที่จะไม่ปฏิบัติตามข้อแนะนำเหล่านี้นั้นต้องมีเหตุผลที่เหมาะสม และควรทำไม่บ่อยนัก ยกตัวอย่างเช่นข้อแนะนำว่า โจทย์ไม่ควรเขียนถามข้อยกเว้น จะพบได้ว่ามีบางบริบทที่การรู้ข้อยกเว้น หรือข้อห้ามปฏิบัติก็เป็นองค์ความรู้ที่สำคัญในการดูแลรักษาผู้ป่วย ดังนั้นในบริบทที่เหมาะสมผู้นิพนธ์เองก็เห็นด้วยว่าอาจเขียนโจทย์ที่ถามข้อยกเว้นได้ แต่อย่างไรก็ตามการจะไม่ปฏิบัติตามข้อแนะนำนี้ต้องไม่ทำบ่อยจนเกินจำเป็น หากออกข้อสอบ ๑๐๐ ข้อ จะมีข้อสอบที่ถามข้อยกเว้น ปะปนมาบ้าง ๒-๓ ข้อ ย่อมเป็นสิ่งที่พอยอมรับได้ แต่หากในชุดข้อสอบมีข้อสอบถึงร้อยละ ๒๐ – ๓๐ ที่โจทย์เขียนในรูปประโยคปฏิเสธ ถามสิ่งที่ไม่ควรปฏิบัติ หรือสิ่งที่ไม่ถูกต้อง อย่างนี้ย่อมจัดว่าละเลยแนวทางในการพัฒนาข้อสอบอย่างไม่เหมาะสม ซึ่งย่อมส่งผลให้คุณภาพของข้อสอบด้อยลงอย่างชัดเจน

เอกสารอ้างอิง

1. Downing SM. Assessment of knowledge with written test forms. In: Norman GR, van der Vleuten C, Newble DI, editors. International handbook of research in medial education. Dordrecht: Kluwer Academic Pubishers, 2002:647 - 72.
2. Haladyna TM, Downing SM. A taxonomy of multiple-choice item-writing rules. Appl Meas Educ 1989;2:37-50.
3. Haladyna TM. Developing and validating multiple-choice test items, 3rd ed. Mahwah, NJ: Lawrence Erlbaum Assoicates, 2004.
4. Maatsch JL, Huang RR, Downing SM, Munger BS. The predictive validity of test formats and a psychometric theory of clinical competence. The 23rd Conference on Research in Medical Education. Washington, DC: Association of American Medical Colleges, 1984.
5. Jozefowicz RF, Koeppen BM, Case S, Galbraith R, Swanson D, Glew RH. The quality of in-house medical school examinations. Acad Med 2002;77(2):156-61.
6. Tarrant M, Ware J. Impact of item-writing flaws in multiple-choice questions on student achievement in high-stakes nursing assessments. Med Educ 2008;42:198-206.

7. Downing SM. The effects of violating standard item writing principles on tests and students: The consequences of using flawed test items on achievement examinations in medical education. Adv Health Sci Educ Theory Pract 2005;10:133-43.
8. Case SM, Swanson D. Constructing written test questions for the basic and clinical sciences, 3rd ed. Philadelphia, PA: National Board of Medical Examiners, 2002.
9. Haladyna TM, Downing SM. Validity of a taxonomy of multiple-choice item-writing rules. Appl Meas Educ 1989;2(1):51-78.
10. Haladyna TM, Downing SM, Rodriguez MC. A review of multiple-choice item-writing guidelines for classroom assessment. Appl Meas Educ 2002;15:309-34.
11. Downing SM, Dawson-Saunders B, Case SM, Powell RD. The psychometric effects of negative stems, unfocused questions, and heterogeneous options on NBME Part I and Part II item characteristics. the annual meeting of the National Council on Measurement in Education. Chicago, IL, 1991.
12. Tamir P. Positive and negative multiple choice items: How different are they? Stud Educ Eval 1993;19:311-25.
13. Rogers WT, Harley D. An empirical comparison of three- and four-choice items and tests: Susceptibility to testwiseness and internal consistency reliability. Educ Psychol Meas 1999;59:234-47.
14. Sidick JT, Barrett GV, Doverspike D. Three-alternative multiple choices tests: An attractive option. Pers Psychol 1994;47:829-35.
15. Cizek GJ, Rachor RE. Nonfunctioning options: A closer look. The annual meeting of the American Educational Research Association. San Francisco, CA, 1995.
16. Crehan KD, Haladyna TM, Brewer BW. Use of an inclusive option and the optimal number of options for multiple-choice items. Educ Psychol Meas 1993;53:241-7.
17. Lord FM. Optimal number of choices per item. J Educ Meas 1977; 14:33-8.
18. Haladyna TM, Downing SM. How many options is enough for a multiple-choice item? Educ Psychol Meas 1993;53:999-1010.

ข้อผิดพลาดที่ควรระวังในการสร้างข้อสอบปรนัย

ผู้ช่วยศาสตราจารย์ นายแพทย์เชิดศักดิ์ ไอรมณีรัตน์
ภาควิชาศัลยศาสตร์, คณะแพทยศาสตร์ศิริราชพยาบาล, มหาวิทยาลัยมหิดล, กรุงเทพมหานคร ๑๐๗๐๐.

ข้อผิดพลาดที่ควรระวังในการสร้างข้อสอบปรนัย

ข้อสอบปรนัย (multiple-choice question) เป็นรูปแบบการประเมินผลที่นิยมใช้กันอย่างแพร่หลาย ในวงการแพทยศาสตรศึกษา ข้อสอบชนิดนี้เป็นที่ชื่นชอบของนักศึกษาผู้เข้าสอบจำนวนมากเนื่องจากมีคำตอบให้เลือก หากไม่มีความรู้ก็สามารถเดาได้ ซึ่งต่างไปจากข้อสอบประเภทอัตนัยซึ่งผู้สอบต้องเขียนคำตอบจากความคิดของตนเอง[๑] ดังนั้นข้อสอบปรนัยจึงเป็นข้อสอบที่ผู้สอบทำได้ง่าย แต่ในทางตรงข้ามข้อสอบปรนัยเป็นข้อสอบที่สร้างปัญหาให้กับอาจารย์ผู้สร้างข้อสอบไม่น้อย เนื่องจากในกระบวนการเขียนข้อสอบปรนัยแต่ละข้อนั้นต้องใช้ทักษะอย่างมาก ต้องใช้ทั้งศาสตร์และศิลป์ และบ่อยครั้งอาจารย์ผู้สร้างข้อสอบก็ถูกขอให้ทำการปรับแก้ข้อสอบเนื่องจากคณะกรรมการพิจารณาข้อสอบมีความเห็นว่ารายละเอียดในข้อสอบไม่เหมาะสม มีการศึกษาวิจัยพบว่าคุณภาพของข้อสอบปรนัยที่พัฒนาขึ้นในโรงเรียนแพทย์หลายแห่งนั้นไม่สู้ดีนัก มีข้อสอบที่มีลักษณะไม่เหมาะสมอยู่จำนวนไม่น้อย[๒-๓] ข้อสอบปรนัยที่มีลักษณะไม่เหมาะสมเหล่านี้ส่งผลเสียต่อการสอบได้หลายประการ เช่น ทำให้ข้อสอบยากขึ้น สร้างความสับสนให้ผู้สอบ ทำให้ผู้สอบบางกลุ่มเสียเปรียบ และทำให้การตัดสินผลสอบผิดพลาด เป็นต้น[๓-๔] ดังนั้นการออกข้อสอบปรนัยที่มีคุณภาพดีจึงเป็นงานที่มีความสำคัญและท้าทายความสามารถ

การสร้างข้อสอบปรนัยที่มีคุณภาพดีนั้นควรเริ่มต้นจากการมีองค์ความรู้พื้นฐานในการสร้างข้อสอบแล้วเกิดการฝึกฝนทักษะ สั่งสมประสบการณ์ในการออกข้อสอบจนเกิดความชำนาญ ปัญหาที่พบบ่อยในโรงเรียนแพทย์หลายแห่งคือมีอาจารย์จำนวนไม่น้อยที่ได้รับมอบหมายให้ออกข้อสอบปรนัย โดยไม่ได้มีการพัฒนาองค์ความรู้พื้นฐานที่เหมาะสมก่อน ซึ่งเป็นเหตุให้มีข้อสอบปรนัยที่มีลักษณะไม่เหมาะสมตามหลักการออกข้อสอบปะปนมาในข้อสอบที่ให้นักศึกษาแพทย์และแพทย์ประจำบ้านทำอยู่บ้าง ผู้นิพนธ์จึงเห็นความสำคัญของการเผยแพร่องค์ความรู้พื้นฐานของการออกข้อสอบปรนัย องค์ความรู้พื้นฐานในการสร้างข้อสอบปรนัยนั้นมีสองส่วน ส่วนแรกเป็นหลักการของการสร้างข้อสอบทั่วไปซึ่งได้มีผู้รวบรวมเป็นข้อแนะนำตีพิมพ์ในตำราและวารสารทางวิชาการอยู่บ้าง[๑,๕-๗] ส่วนที่สองเป็นข้อผิดพลาดในการสร้างข้อสอบที่อาจารย์ผู้ออกข้อสอบพึงหลีกเลี่ยง ในบทความนี้ผู้นิพนธ์จะมุ่งเน้นในส่วนที่สองนี้ โดยจะรวบรวมข้อผิดพลาดในการสร้างข้อสอบปรนัย ที่อาจเป็นตัวบอกใบ้ให้ผู้สอบที่ไม่มีความรู้ในเรื่องที่ทำการทดสอบสามารถเลือกคำตอบที่ถูกต้องได้ ดังนั้นการที่อาจารย์ผู้ออกข้อสอบทราบถึงสิ่งเหล่านี้และหลีกเลี่ยงเสียจะส่งผลให้ข้อสอบปรนัยที่สร้างขึ้นสามารถใช้วัดองค์ความรู้ทางการแพทย์ได้จริง โดยปราศจากปัจจัยรบกวนจากการสังเกตพบสิ่งบอกใบ้คำตอบ

ข้อสอบปรนัยที่กล่าวถึงในบทความนี้มุ่งประเด็นไปที่ข้อสอบปรนัยชนิดเลือกคำตอบที่ถูกที่สุด (one best response) เป็นสำคัญ เนื่องจากเป็นข้อสอบที่ใช้กันแพร่หลายมากที่สุดในการวัดผลการศึกษาในโรงเรียนแพทย์ไทยปัจจุบัน ในข้อสอบชนิดนี้แต่ละข้อจะมีโจทย์ (stem) ตามด้วยตัวเลือก (options) จำนวน ๔-๕ ตัวเลือก ผู้สอบต้องเลือกคำตอบที่ถูกต้องที่สุดเพียงคำตอบเดียวจากตัวเลือกเหล่านี้ ตัวเลือกอื่น ๆ ที่ไม่ใช่คำตอบเรียกว่าตัวลวง (distractors)

ในบทความนี้ผู้นิพนธ์ขอนำเสนอข้อผิดพลาดในการออกข้อสอบ ๗ กลุ่มด้วยกัน ได้แก่ (๑) ข้อผิดพลาดในไวยากรณ์, (๒) การใบ้คำตอบด้วยหลักตรรกะ, (๓) การใช้คำคุณศัพท์บอกระดับของความแน่ชัด, (๔) ความยาวของตัวเลือก, (๕) การใช้คำซ้ำในโจทย์และตัวเลือก, (๖) การเข้าพวกของคำ หรือข้อความที่ปรากฏในตัวเลือก, และ (๗) การบอกใบ้คำตอบโดยโจทย์ข้ออื่น

๑. ข้อผิดพลาดในไวยากรณ์

ตัวเลือกทุกตัวต้องสามารถตอบโจทย์ได้อย่างถูกต้องตามหลักไวยากรณ์ บ่อยครั้งอาจารย์ผู้ออกข้อสอบมุ่งความสนใจไปที่คำตอบที่ถูก และให้ความสนใจกับตัวลวงน้อยไปจนทำให้ตัวลวงผิดหลักไวยากรณ์ โดยมักพบบ่อยในข้อสอบที่เป็นภาษาอังกฤษ ข้อผิดพลาดที่พบได้บ่อยเช่น ความไม่เข้ากันของ article (A, An, The) กับคำนามที่ตามหลัง, คำนามกับกิริยาที่ไม่เข้ากันในเชิงเอกพจน์หรือพหูพจน์, การเติมคำในประโยคที่เว้นว่างไว้สำหรับเติมคำนามแต่ตัวลวงเป็นกิริยาหรือเป็นคำนามในลักษณะที่ไม่เข้ากับรูปประโยค เป็นต้น

ตัวอย่างที่ ๑. A 70-year-old woman was brought in an emergency room with alteration of consciousness. Her vital signs were stable, but her Glasgow coma score was E1V1M3. After endotracheal intubation, the next step is to provide intravenous administration of …

A. lumbar puncture

B. computerized scan of the brain

C. glucose with Thiamine

D. Sodium bicarbonate

ในตัวอย่างที่ ๑ นี้โจทย์ให้ผู้สอบเลือกตัวเลือกไปเติมในช่องว่าง ซึ่งสิ่งที่เติมลงในช่องว่างได้นั้นต้องเป็นยาที่สามารถให้ทางหลอดเลือดดำได้ ผู้สอบที่มีทักษะการทำข้อสอบดีจะตัดตัวเลือก A และ B ได้โดยไม่ต้องใช้ความรู้ทางการแพทย์

ตัวอย่างที่ ๒. Which organism is the cause of syphilis?

A. *Neisseria gonorrheae*

B. *Chlamydia trachomatis and Giardia lamblia*

C. *Treponema pallidum*

D. *Ureaplasma urealyticum and Mycoplasma genitalium*

ในตัวอย่างที่ ๒ นี้โจทย์ถามหาเชื้อก่อโรค โดยใช้รูปประโยคถามหาคำตอบที่เป็นเอกพจน์ ดังนั้นคำตอบที่ถูกต้องย่อมมีเชื้อก่อโรคตัวเดียว ผู้สอบที่มีทักษะการทำข้อสอบดีจะตัดตัวเลือก B และ D ได้โดยไม่ต้องใช้ความรู้ทางการแพทย์

๒. การใบ้คำตอบด้วยหลักตรรกะ

ในการเขียนตัวเลือก อาจารย์ผู้ออกข้อสอบต้องระมัดระวังไม่ให้ผู้สอบสามารถตัดตัวเลือกได้ด้วยหลักตรรกศาสตร์ เนื่องจากผู้สอบที่มีทักษะการทำข้อสอบดีจะสามารถพิจารณาความเป็นไปได้ของตัวเลือกต่าง ๆ และตัดตัวลวงที่ไม่มีทางเป็นไปได้ตามหลักของเหตุและผลออกไปได้โดยไม่ต้องอาศัยความรู้เรื่องที่อาจารย์ตั้งเป้าหมายว่าจะทดสอบ

ตัวอย่างที่ ๓. ภาวะไส้เลื่อนบริเวณขาหนีบ (inguinal hernia) …….

A. พบในผู้ชายบ่อยกว่าผู้หญิง

B. พบในผู้หญิงบ่อยกว่าผู้ชาย

C. พบเกิดขึ้นในผู้หญิงและผู้ชายในอัตราเท่ากัน

D. พบบ่อยในผู้ที่มีเศรษฐานะยากจน

E. พบในผู้ที่มีภูมิลำเนาในทวีปเอเชีย มากกว่าผู้ที่มีภูมิลำเนาในทวีปยุโรป

ในตัวอย่างที่ ๓ นี้อาจารย์ผู้ออกข้อสอบต้องการวัดความรู้เรื่องอุบัติการณ์ของไส้เลื่อนขาหนีบ แต่หาก

พิจารณาตามหลักตรรกศาสตร์แล้ว ตัวเลือก A, B, และ C เพียงสามตัวเลือกก็ครอบคลุมสิ่งที่เป็นไปได้ทั้งหมดแล้ว (เนื่องจากมนุษย์มีสองเพศ ภาวะไส้เลื่อนนี้หากไม่มีอัตราการเกิดเท่ากันในสองเพศแล้วก็ต้องมีเพศใดเป็นมากกว่าอีกเพศหนึ่ง) ดังนั้นผู้สอบที่มีทักษะการทำข้อสอบดีสามารถตัดตัวเลือก D และ E ได้โดยไม่ต้องมีความรู้เรื่องไส้เลื่อนเลย

๓. การใช้คำคุณศัพท์บอกระดับของความแน่ชัด

อาจารย์ผู้ออกข้อสอบพึงระมัดระวังการใช้คำคุณศัพท์ที่บ่งบอกถึงความแน่ชัดของข้อความ ซึ่งจะมีหลายระดับ โดยทั่วไปแล้วคำคุณศัพท์ที่แสดงความแน่ชัดมาก แสดงความมั่นใจมาก (เช่น always, never) มักไม่ถูกต้อง เนื่องจากในทางการแพทย์นั้นมีความไม่แน่นอนเกิดขึ้นเป็นประจำ ข้อความที่บอกเล่าถึงสิ่งที่อาจเป็นไปได้โดยไม่ชี้ชัดลงไปว่าต้องเกิดขึ้นแน่นอน (เช่น may, might, can, could) มักเป็นข้อความที่ถูก

ตัวอย่างที่ ๔. Which of the following statements is true regarding the etiology of an inguinal hernia?

A. Some connective tissue diseases may increase the incidence of inguinal hernia.

B. Patients with Marfan syndrome always developed inguinal hernia.

C. MRI scan of pelvis is the only reliable investigation for detection of groin hernia.

D. Persistent lifting of heavy weights inevitably leads to the development of groin hernia.

ในตัวอย่างที่ ๔ นี้ผู้สอบต้องเลือกข้อความเกี่ยวกับไส้เลื่อนขาหนีบที่ถูกต้องหนึ่งข้อความ หากสังเกตดูทั้งสี่ข้อความมีการใช้คำคุณศัพท์บอกความแน่ชัดของข้อความ ได้แก่ may (ตัวเลือก A), always (ตัวเลือก B), the only (ตัวเลือก C), inevitably (ตัวเลือก D) ซึ่งจะเห็นว่าตัวเลือก B, C, และ D เป็นข้อความที่แสดงความแน่ชัดว่าต้องเป็นแน่ ต้องใช่แน่นอน ไม่มีทางเลี่ยงได้ ข้อความทำนองนี้มีโอกาสสูงที่จะผิด ในทางตรงข้ามตัวเลือก A เป็นข้อความบอกว่ามีโอกาสเป็นได้โดยไม่ชี้ชัดว่าต้องเกิด ผู้สอบที่มีทักษะการทำข้อสอบดีจะตัดตัวเลือก B, C, และ D ได้โดยไม่ต้องอาศัยความรู้ทางการแพทย์เลย

๔. ความยาวของตัวเลือก

มีการตั้งข้อสังเกตว่าอาจารย์แพทย์มักชอบสอนและอธิบายแม้กระทั่งในการสอบอาจารย์แพทย์หลายท่านก็ติดนิสัยรักการสอนนี้มาด้วย ทำให้อาจารย์มักเขียนตัวเลือกที่ถูกต้องที่มีคำอธิบายประกอบอย่างครบถ้วนทำให้ตัวเลือกที่ถูกมักมีความยาวมากกว่าตัวลวง นักศึกษาผู้เข้าสอบจำนวนไม่น้อยรู้ถึงความจริงข้อนี้และมักเลือกตัวเลือกที่มีความยาวมากที่สุด หากเขาไม่สามารถหาคำตอบได้ด้วยความรู้ทางการแพทย์ที่เขามี

ตัวอย่างที่ ๕. ผู้หญิงอายุ ๒๘ ปี แต่งงานมานาน ๑ ปี ยังไม่มีบุตร คุมกำเนิดโดยการกินยาคุมเป็นประจำ สังเกตว่าตนเองน้ำหนักตัวเพิ่มขึ้นหลังจากกินยาคุม มาขอคำแนะนำเรื่องการคุมกำเนิด ท่านจะแนะนำอย่างไร

A. ให้เปลี่ยนไปใช้การใส่ห่วงอนามัย

B. ให้ใช้ถุงยางอนามัย

C. ให้กินยาคุมกำเนิดต่อได้เนื่องจากมีการศึกษาแล้วว่ายาคุมกำเนิดชนิดกินไม่ส่งผลให้เกิดการเพิ่มขึ้นของน้ำหนักตัว

D. ให้รับประทานยาลดความอ้วน

ในตัวอย่างที่ ๕ นี้จะสังเกตเห็นว่าตัวเลือก C มีการอธิบายเหตุผลประกอบส่งผลให้มีความยาวมากกว่าตัวเลือกอื่นชัดเจน ลักษณะเช่นนี้จะเป็นการบอกใบ้ให้นักศึกษาเลือกตัวเลือกนี้

๕. การใช้คำซ้ำในโจทย์และตัวเลือก

การใช้คำเดียวกัน หรือคำที่มีความหมายเหมือนกันในโจทย์และตัวเลือก มักเป็นการบอกใบ้ว่าตัวเลือกดังกล่าวเป็นตัวเลือกที่ถูกต้อง

ตัวอย่างที่ ๖. Which of the following statements is true regarding saccular theory of indirect inguinal hernia formation?

A. An increased intra-abdominal pressure is the cause of inguinal hernia.

B. A developmental diverticulum associated with a patent processus vaginalis is the cause of inguinal hernia.

C. All persons with a persistent processus vaginalis will develop an inguinal hernia.

D. A direct inguinal hernia is caused by the weakness of the posterior inguinal wall.

ในตัวอย่างที่ ๖ นี้โจทย์ถามถึง saccular theory ซึ่งหากแปลความหมายก็น่าจะเป็นเรื่องที่เกี่ยวข้องกับถุง (sac) ผู้สอบที่มีทักษะการทำข้อสอบดีจะหาตัวเลือกที่มีคำที่มีความหมายเกี่ยวกับถุง แล้วเลือกตัวเลือกดังกล่าวทันที ซึ่งในที่นี้จะพบคำว่า diverticulum ซึ่งมีความหมายว่าถุงในข้อ B การที่มีคำที่มีความหมายซ้ำกันเช่นนี้เป็นตัวบอกใบ้คำตอบที่อาจารย์ผู้ออกข้อสอบต้องตรวจตราให้ดีก่อนนำข้อสอบไปใช้

๖. การเข้าพวกของคำ หรือข้อความที่ปรากฏในตัวเลือก

ข้อสอบจำนวนไม่น้อยนำเสนอรายการของหลายอย่างในตัวเลือก (เช่น ชื่อการตรวจค้นเพิ่มเติม ชื่อโรค ชื่อยา ฯลฯ) มีผู้เชี่ยวชาญในการประเมินผลตั้งข้อสังเกตว่าในข้อสอบเหล่านี้ตัวเลือกที่ถูกต้องมักมีลักษณะเข้าพวกกับตัวเลือกอื่นมากที่สุด หากเป็นรายการของตัวเลือกที่ถูกก็คือข้อที่มีจำนวนรายการซ้ำกับตัวเลือกอื่นมากที่สุด[๕] ดังนั้นในการนำเสนอตัวเลือกอาจารย์ผู้ออกข้อสอบพึงระมัดระวังอย่าให้ตัวเลือกที่ถูกต้องมีลักษณะที่เข้าพวกได้อย่างชัดเจน พยายามทำตัวลวงอื่นให้มีลักษณะเข้าพวกให้ใกล้เคียงกับตัวเลือกที่ถูกต้อง

ตัวอย่างที่ ๗. โรคที่แพทย์วินิจฉัยผิดว่าเป็นไส้ติ่งอักเสบบ่อยที่สุดเรียงลำดับจากมากไปน้อยคือ

A. acute mesenteric lymphadenitis, pelvic inflammatory disease, twisted ovarian cyst

B. acute mesenteric lymphadenitis, Meckel diverticulitis, acute cholecystitis

C. Meckel diverticulitis, twisted ovarian cyst, sigmoid diverticulitis

D. pelvic inflammatory disease, acute gastroenteritis, right ureteric calculi

ในตัวอย่างที่ ๗ นี้โจทย์ถามชื่อโรค ตัวเลือกแสดงรายการชื่อโรค ตัวเลือกละสามโรค หากนับจำนวนของคำซ้ำจะพบว่าโรคที่กล่าวถึงบ่อยที่สุดคือ acute mesenteric lymphadenitis, pelvic inflammatory disease, twisted ovarian cyst, และ Meckel diverticulitis (กล่าวถึงโรคละ ๒ ครั้ง) ส่วนโรคที่เหลือกล่าวถึงโรคละครั้งเดียว ดังนั้นตัวเลือกที่มีพวกมากที่สุดคือตัวเลือก A ซึ่งเป็นคำตอบที่ถูกต้อง

การเข้าพวกของตัวเลือกที่ถูกนั้น ไม่จำเป็นต้องเป็นลักษณะของการมีจำนวนหรือความถี่ของคำมากที่สุดเพียงเท่านั้น อาจหมายรวมถึงการมีรูปร่างลักษณะ หรือความหมายคล้ายคลึงกันได้ด้วย

ตัวอย่างที่ ๘. ชายอายุ ๕๕ ปีเป็นมะเร็งเม็ดเลือดขาว หลังได้รับยาเคมีบำบัด ๑๔ วันมีไข้สูง ได้รับการวินิจฉัยเป็น febrile neutropenia การรักษาในข้อใดเหมาะสมที่สุด

A. Amoxycillin PO

B. Ceftazidime IV + Amikacin IV

C. Amphotericin B IV + Ceftazidime IV

D. Cloxacillin IV + Metronidazole IV

ในตัวอย่างที่ ๘ นี้โจทย์ถามถึงยาที่ควรให้กับผู้ป่วย ในตัวเลือกสี่ตัวเลือกนี้มียากินเพียงข้อเดียว (A) ที่เหลือเป็นยาฉีดสองขนานควบกัน ดังนั้นตัวเลือกข้อ A ไม่เข้าพวก จะถูกตัดทิ้งได้โดยง่าย ในบรรดายาฉีดจะเห็นว่ามียาต้านเชื้อราที่ไม่เข้าพวก (ตัวเลือก C) ดังนั้นจะเหลือตัวเลือกที่นักศึกษาต้องคิดเลือกจริง ๆ เพียงตัวเลือก B กับ D ซึ่งหากดูกลุ่มยาก็จะพบว่ายากลุ่ม Cephalosporin เข้าพวกมากที่สุด ทำให้ผู้สอบที่มีทักษะการทำข้อสอบดีสามารถเลือกคำตอบที่ถูกต้อง (ตัวเลือก B) ได้โดยไม่ต้องมีความรู้เรื่องการรักษาผู้ป่วย febrile neutropenia

๗. การบอกใบ้คำตอบโดยโจทย์ข้ออื่น

ข้อผิดพลาดนี้เป็นข้อผิดพลาดที่ตัวผู้เขียนข้อสอบไม่ค่อยรู้ แต่ผู้ที่จะตรวจพบข้อผิดพลาดนี้คืออาจารย์ผู้เลือกข้อสอบไปใช้ เนื่องจากในการสอบแต่ละครั้งใช้ข้อสอบจำนวนมาก หากเลือกข้อสอบโดยไม่ระมัดระวังอาจมีข้อสอบสองข้อที่ถามเกี่ยวกับโรคหรือกลุ่มอาการเดียวกัน ซึ่งข้อมูลจากโจทย์ในข้อหนึ่งอาจเป็นตัวบอกใบ้คำตอบของข้อสอบอีกข้อได้ ดังนั้นเมื่อทำการเลือกข้อสอบเสร็จแล้วจัดหน้ากระดาษเข้ารูปเล่มข้อสอบแล้วอาจารย์ควรอ่านข้อสอบฉบับสมบูรณ์นี้อีกหนึ่งหรือสองรอบก่อนส่ง

ไปพิมพ์ ซึ่งการอ่านทวนในขั้นตอนนี้อาจทำให้ตรวจพบข้อสอบที่มีเนื้อหาซ้ำซ้อนกันได้

ตัวอย่างที่ ๙. ผู้ป่วย febrile neutropenia มักมีไข้ขึ้นหลังจากได้รับยาเคมีบำบัดเป็นเวลากี่วัน

A. 2 - 4 วัน

B. 3 - 5 วัน

C. 5 - 7 วัน

D. 10 - 14 วัน

ในตัวอย่างที่ ๙ นี้อาจารย์ผู้ออกข้อสอบต้องการวัดความรู้ของผู้สอบเรื่อง febrile neutropenia ซึ่งเนื้อหาไปซ้ำซ้อนกับโจทย์ในตัวอย่างที่ ๘ ซึ่งผู้สอบที่มีทักษะการทำข้อสอบดีสามารถย้อนกลับไปอ่านโจทย์ในข้อก่อนหน้านี้แล้วได้ข้อมูลว่าผู้ป่วยที่นำเสนอว่าเป็น febrile neutropenia มีไข้ขึ้น ๑๔ วันหลังได้ยาเคมีบำบัด ก็สามารถตอบข้อสอบข้อนี้ถูกได้โดยง่าย

สรุป

ผู้นิพนธ์ได้รวบรวมข้อผิดพลาดในการสร้างข้อสอบปรนัยที่ผู้สอบอาจใช้เป็นแนวทางในการเลือกคำตอบที่ถูกได้โดยไม่ต้องอาศัยความรู้ทางการแพทย์ที่อาจารย์ต้องการประเมินผล โดยเรียบเรียงเป็นเจ็ดกลุ่มข้อผิดพลาดด้วยกัน ผู้อ่านทุกท่านพึงตระหนักว่าสิ่งเหล่านี้ไม่ใช่หลักการทางวิทยาศาสตร์ที่ชัดเจนดังกฎทางคณิตศาสตร์หรือฟิสิกส์ หากแต่เป็นการรวบรวมข้อสังเกตและคำแนะนำของผู้เชี่ยวชาญทางการวัดและประเมินผล จึงเป็นเพียงแนวทางเบื้องต้นในการพิจารณาตรวจสอบเนื้อหาของข้อสอบเท่านั้น การประยุกต์ใช้องค์ความรู้นี้คงต้องอาศัยศิลปะพอสมควรเพื่อที่จะได้ข้อสอบที่ดีสามารถวัดองค์ความรู้ทางการแพทย์ของนักศึกษาหรือแพทย์ประจำบ้านที่เข้าสอบได้ตามวัตถุประสงค์ของการสอบ

เอกสารอ้างอิง

1. Haladyna TM. Developing and validating multiple-choice test items, 3rd ed. Mahwah, NJ: Lawrence Erlbaum Assoicates, 2004.
2. Jozefowicz RF, Koeppen BM, Case S, Galbraith R, Swanson D, Glew RH. The quality of in-house medical school examinations. Acad Med. 2002;77:156-61.
3. Tarrant M, Ware J. Impact of item-writing flaws in multiple-choice questions on student achievement in high-stakes nursing assessments. Med Educ. 2008;42:198-206.
4. Downing SM. The effects of violating standard item writing principles on tests and students: The consequences of using flawed test items on achievement examinations in medical education. Adv Health Sci Educ Theory Pract. 2005;10:133-43.
5. Haladyna TM, Downing SM. A taxonomy of multiple-choice item-writing rules. Appl Meas Educ. 1989;2:37-50.
6. Haladyna TM, Downing SM. Validity of a taxonomy of multiple-choice item-writing rules. Appl Meas Educ. 1989;2:51-78.
7. Haladyna TM, Downing SM, Rodriguez MC. A review of multiple-choice item-writing guidelines for classroom assessment. Appl Meas Educ. 2002;15:309-34.
8. Case SM, Swanson D. Constructing written test questions for the basic and clinical sciences, 3rd ed. Philadelphia, PA: National Board of Medical Examiners, 2002.

ผศ. นพ.สุประพัฒน์ สนใจพาณิชย์

หัวข้อ : Constructed response item development

Mahidol University
Faculty of Medicine
Siriraj Hospital

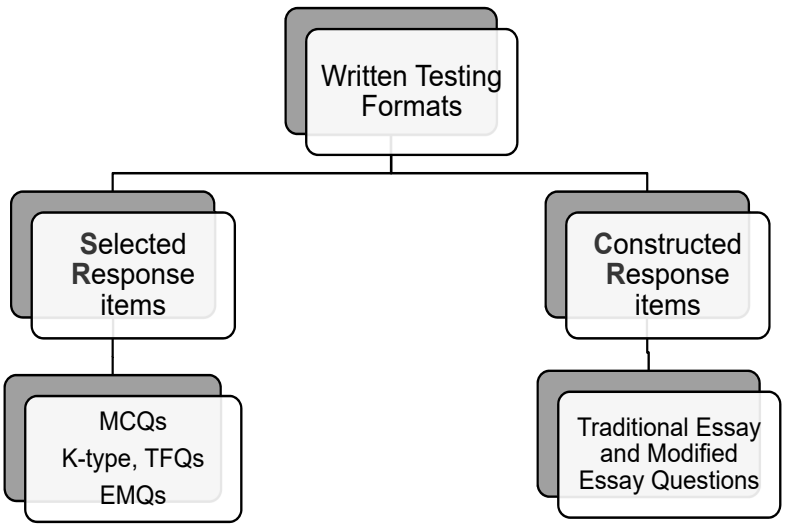
SHEE

Constructed Response Items

Suprapath Sonjaipanich MD.
Department of Pediatrics
Faculty of Medicine Siriraj Hospital
Mahidol University

Mahidol University
Faculty of Medicine
Siriraj Hospital

SHEE



Downing S.M. & Yudkowsky R. Written Tests: Constructed-Response and Selected-Response Formats. Assessment in Health Professions Education 2009

Mahidol University Faculty of Medicine Siriraj Hospital — SHEE

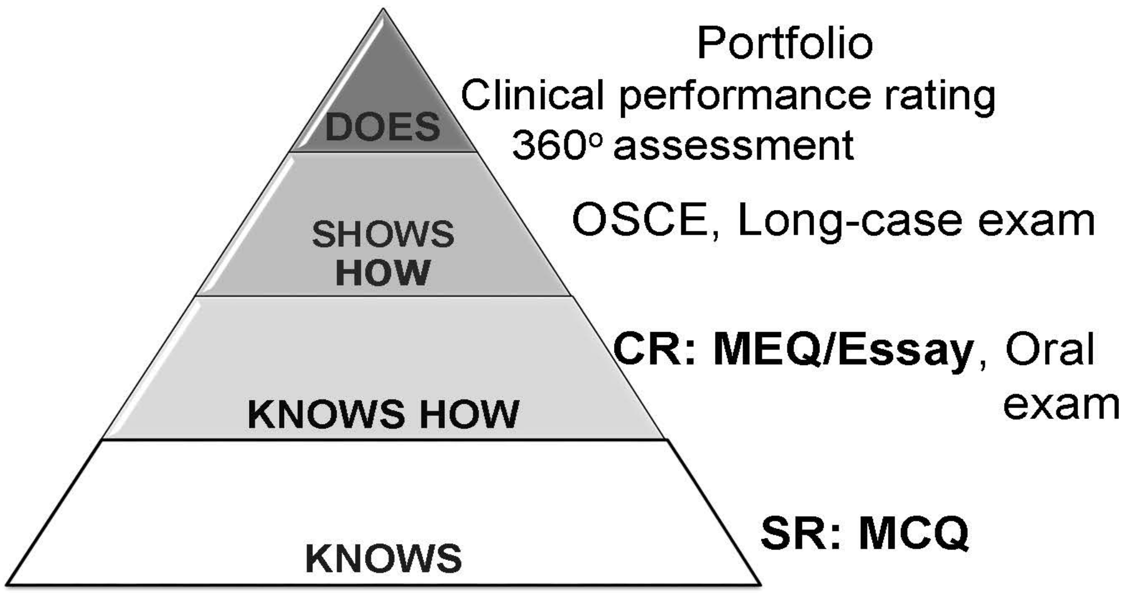
Comparison

	Selected Response	Constructed Response
Measured construct	Recall Basic interpretation, some applications	Problem solving, interpretation, decision making
Item construction	Simple	Complex
Cost of scoring	Low	Expensive
Type of scoring	Objective	Subjective
Rater effects	No effect	Significant factor
Reliability	High	Low

Adapted from Table 3.2 In Haladyna TM, Developing and validating multiple-choice Test items, 3rd ed. Mahwah, NJ: Lawrence Erlbaum Associates, 2004.

Mahidol University Faculty of Medicine Siriraj Hospital — SHEE

Assessment approach



Miller's pyramid

Mahidol University Faculty of Medicine Siriraj Hospital

SHEE

Constructed Response Items

Traditional Essay Questions

- Long essay
- Short essay

Modified Essay Questions

- Standard modified essay questions (MEQ)
- Key-feature questions (KFQ)
- Patient management problem (PMP)
- Short answer questions (SAQ)

Mahidol University Faculty of Medicine Siriraj Hospital

SHEE

Objectives

เมื่อสิ้นสุดกิจกรรม ผู้เข้าร่วมอบรมสามารถ

1. อธิบายข้อดีและข้อจำกัดของข้อสอบชนิด constructed response (CR) items
2. อธิบายขั้นตอนและประเด็นสำคัญของการสร้างข้อสอบ CR รูปแบบ Modified Essay Questions (MEQ) ได้
3. ร่วมในกระบวนการพัฒนาและทบทวนข้อสอบ CR สำหรับนักศึกษาระดับคลินิกที่แต่ละท่านเกี่ยวข้องได้

CR item development

- Clinical Problem Solving Methods
- Modified Essay Questions
  - Standard MEQ
  - Key-feature questions
- Developing an MEQ

Mahidol University
Faculty of Medicine
Siriraj Hospital
SHEE

CR Items: Strengths

- Able to measure higher-order cognitive abilities
- Uncued written responses
- Mimic actual clinical problem solving
- Motivation for clinical learning

Mahidol University
Faculty of Medicine
Siriraj Hospital

SHEE

CR Items: Limitations

- Construct underrepresentation
- Difficult to develop and score
- Unexpected responses
- Subjective scoring
- Low reliability

Mahidol University
Faculty of Medicine
Siriraj Hospital

SHEE

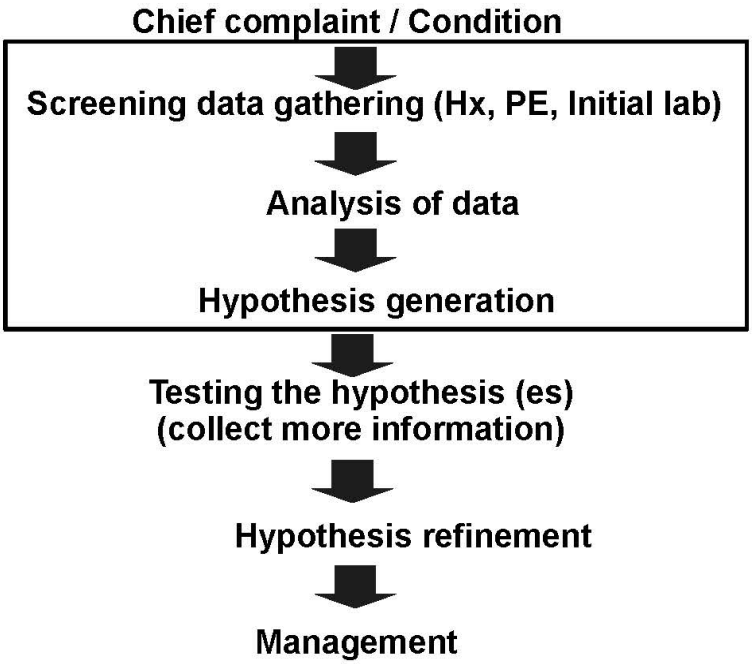
Clinical Problem Solving Methods

1. Pattern recognition
2. Algorithm
3. Forward reasoning (data driven process)
4. Backward reasoning (hypothesis driven process)

Mahidol University
Faculty of Medicine
Siriraj Hospital

SHEE

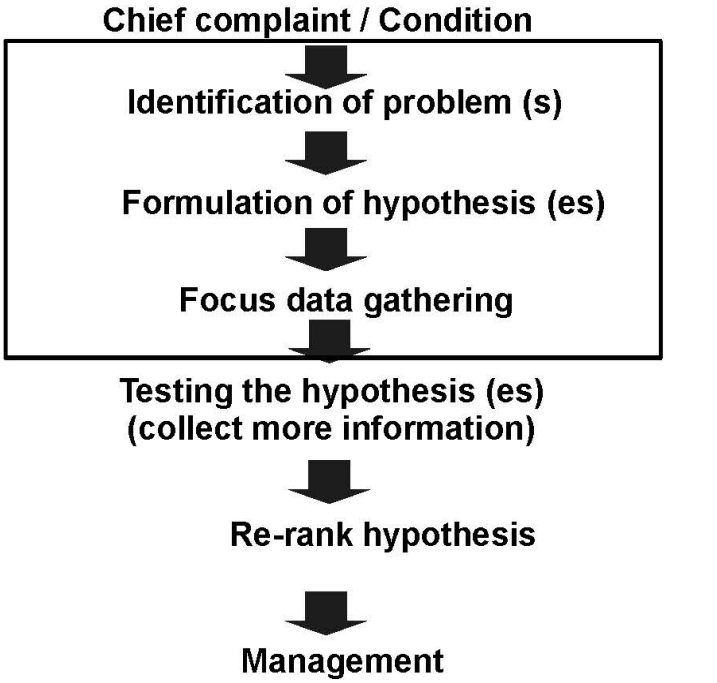
Forward Reasoning (Data driven process)



Mahidol University
Faculty of Medicine
Siriraj Hospital

SHEE

Backward reasoning (Hypothesis driven process)



Mahidol University
Faculty of Medicine
Siriraj Hospital

SHEE

Forward Reasoning
(Data driven process)

- เด็กชายอายุ 4 ปี มีอาการนอนกรนมา 1 ปี

Q#1 ท่านจะซักประวัติเพิ่มเติมอะไรบ้างเพื่อการวินิจฉัยโรค

Chief complaint / Condition

↓

Screening data gathering (Hx, PE, Initial lab)

↓

Analysis of data

↓

Hypothesis generation

Mahidol University
Faculty of Medicine
Siriraj Hospital

SHEE

Backward reasoning
(Hypothesis driven process)

- เด็กชายอายุ 4 ปี มีอาการนอนกรนมา 1 ปี ช่วง 3 เดือนนี้นอนกรนเกือบทุกคืน กระสับกระส่าย สะดุ้งตื่นบ่อย มีหายใจสะดุดเป็นระยะ

Q#1 จงบอกสาเหตุที่เป็นไปได้ของการนอนกรนในผู้ป่วยเด็กรายนี้มา 3 ข้อ

Chief complaint / Condition

↓

Identification of problem (s)

↓

Formulation of hypothesis (es)

↓

Focus data gathering

Mahidol University
Faculty of Medicine
Siriraj Hospital

SHEE

MEQ Process
(Hypothesis driven)

A brief clinical situation (scenario, vignette)

↓

Hypothesis generation / Differential diagnosis

↓

(Focused) Data gathering (Hx, PE., Lab)

↓

Hypothesis refinement

↓

Management/Intervention

↓

Question related to clinical/basic sciences
(optional)

Mahidol University
Faculty of Medicine
Siriraj Hospital

SHEE

MEQ: Serial Question Test

- เสมือนการแก้ปัญหาผู้ป่วยในชีวิตจริง
- การแก้ปัญหาของผู้ป่วยประกอบด้วยหลายขั้นตอน
  - มีข้อมูลผู้ป่วยบางส่วนในช่วงแรก
  - ต้องสืบค้นหาข้อมูลเพิ่มเติมและวิเคราะห์ ตัดสินใจแก้ปัญหาทีละขั้นตอน
  - เมื่อทำแต่ละขั้นตอนแล้ว ไม่สามารถย้อนกลับไปแก้ไขสิ่งที่ทำไปก่อนหน้านี้ได้

Mahidol University
Faculty of Medicine
Siriraj Hospital

SHEE

Physician tasks / Competencies

- Data Gathering (Hx, PE, Lab)
- Hypothesis Generation (Differential Dx)
- Hypothesis Refinement (Dx)
- Management (Emergency, Acute, Long-term)
- Health promotion and maintenance
- Counseling education
- Medical ethics
- Evidence-based
- Mechanism of diseases

Mahidol University
Faculty of Medicine
Siriraj Hospital

SHEE

Standard MEQ

- Chief complaint
- A question on differential diagnosis
- Questions to collect additional information
- Additional clinical information
- Provisional diagnosis
- A question on management plan
- Additional clinical information
- Interpretation of laboratory findings
- Exploring knowledge, reasoning

วิษณุ ธรรมลิขิตกุล การประเมินความรู้ในการแก้ปัญหาผู้ป่วยทางคลินิก. สารศิริราช 2534, 43(2): 123 – 134.

Mahidol University
Faculty of Medicine
Siriraj Hospital

SHEE

Key-feature Questions (KFQs)

- Key features
  - critical steps in the resolution of each problem
- Focus on
  - a step in which examinees are most likely to make errors
  - a difficult aspect of the identification and management of the problem in clinical practice

1. Page G, Bordage G. The Medical Council of Canada's key features project: A more valid written examination of clinical decision-making skills. *Acad Med* 1995
2. Farmer EA, Page G. A practical guide to assessing clinical decision-making skills using the key features approach. *Med Educ* 2005

Mahidol University
Faculty of Medicine
Siriraj Hospital

SHEE

Key-feature Questions (KFQs)

- Allow for more cases, items for testing a broader content domain

*"In any clinical case, there are **a few essential elements** in decision making which are the **critical steps** in the resolution of the clinical problem."*

- Reliability of 0.8 in 4 hours of testing had been demonstrated

Page G, Bordage G. The Medical Council of Canada's key features project: A more valid written examination of clinical decision-making skills. *Acad Med* 1995; 70: 104-10.

Mahidol University
Faculty of Medicine
Siriraj Hospital

SHEE

Key features: Example

Topic Anaphylaxis (Food-induced anaphylaxis)

Key features

I. Diagnosis

II. Emergency management

III. Prevention

Mahidol University
Faculty of Medicine
Siriraj Hospital

SHEE

KFQ: example

เด็กชายอายุ 5 ปี มาที่ห้องฉุกเฉิน ด้วยอาการผื่นทั่วตัว 30 นาที หลังกินอาหารที่ร้านอาหารแห่งหนึ่ง

ปกติเป็นเด็กแข็งแรง ไม่มีโรคประจำตัว

ตรวจร่างกายแรกรับ

Pulse 150/min (weak pulse), capillary refill 3 sec, SpO_2 90% (room air), generalized wheal and flare rash, bilateral wheezing

Mahidol University
Faculty of Medicine
Siriraj Hospital

SHEE

KFQ: Example

คำถาม

Q1: จงให้การวินิจฉัยโรค/ภาวะที่เป็นไปได้มากที่สุด

Q2: จงเขียนคำสั่งการรักษาเบื้องต้นที่ห้องฉุกเฉิน
(หากใช้ยา ให้ระบุชื่อ ขนาดและวิธีบริหาร)

Q3: จงบอกคำแนะนำเพื่อการป้องกันการเกิดซ้ำ

Mahidol University
Faculty of Medicine
Siriraj Hospital

SHEE

Developing an MEQ

- Assembling problem-writing groups
- Selecting a problem
- Defining the key features
- Writing the questions
- Selecting question formats
- Specifying the number of required answers
- Preparing scoring keys
- Validation and references

Farmer EA, Page G. A practical guide to assessing clinical decision-making skills using the key features approach. Med Educ 2005, 39: 1188 – 1194.

Mahidol University
Faculty of Medicine
Siriraj Hospital

SHEE

Assembling Problem-Writing Groups

- Item writers
  - Clinical expertise
  - Multidisciplinary approach / combined expertise
- The written problem
  - well grounded in practice
  - represent a wide range of real-life practice
- Review the content by a group of writers

Mahidol University
Faculty of Medicine
Siriraj Hospital

SHEE

Select A Problem

- Refer to test specification table
- Select an appropriate clinical problem

1. พบบ่อยในเวชปฏิบัติ
2. ความยากง่ายเหมาะสมกับระดับของผู้เรียน
3. ประเมินทักษะการแก้ปัญหาและการตัดสินใจ
4. เกี่ยวข้องกับหลายระบบ
5. มีการบูรณาการของสาขาวิชา
6. แพทย์มักตัดสินใจผิดพลาด

Mahidol University
Faculty of Medicine
Siriraj Hospital

SHEE

Defining Key Features

- ปรึกษาในกลุ่มผู้เขียนโจทย์จนได้ consensus
- Critical steps
  - ประเด็นสำคัญในการตัดสินใจ/จัดการกับปัญหาของผู้ป่วย
  - ขั้นตอนที่ขาดไม่ได้ในการดูแลรักษาผู้ป่วย
  - อาจเป็นประเด็นเกี่ยวกับเรื่อง medical ethics, medico-legal

Mahidol University
Faculty of Medicine
Siriraj Hospital

SHEE

Defining Key Features (cont.)

- Typical KFs
  - ประวัติเพิ่มเติมที่สำคัญ
  - การตรวจร่างกายที่ต้องมองหาหรือตรวจเพิ่มเติม
  - การสืบค้นเพิ่มเติมเพื่อ confirm หรือ exclude การวินิจฉัย
  - การรักษาที่เฉพาะเจาะจงกับโรค

<u>ไม่จำเป็น</u>ต้องเริ่มต้นด้วยการถามประวัติ หรือ ตรวจร่างกาย

Mahidol University
Faculty of Medicine
Siriraj Hospital

SHEE

From A Problem to A Case

- Select a case scenario
  - age, gender
  - setting of the encounter: OPD, IPD, ER
  - brief case: KFQ on diagnosis
  - longer case: KFQ on management

Mahidol University
Faculty of Medicine
Siriraj Hospital

SHEE

Writing the Questions

- Number of questions
  - Most case scenario: 2 – 4 questions
  - Each question test one key feature
- Number of answers for each question
  - Vary: 1 – 10
  - Typical: 3 – 5 answers

Mahidol University
Faculty of Medicine
Siriraj Hospital

SHEE

Specify the Number of Required Answers

ระบุคำถามให้ชัดเจนว่าจะให้ทำอะไร อย่างไร เช่น

- บอกชื่อโรคที่ผู้ป่วยรายนี้น่าจะเป็นมากที่สุด 1 โรค
- บอกสิ่งตรวจพบจากการตรวจร่างกายที่สำคัญที่จะช่วยในการยืนยันการวินิจฉัยโรค มา 3 ประการ
- เขียนคำสั่งการรักษาสำหรับผู้ป่วยรายนี้ในใบคำสั่งการรักษา

Mahidol University
Faculty of Medicine
Siriraj Hospital

SHEE

Preparing Scoring Keys (1/4)

- List of correct and incorrect responses
- Scores to be assigned to each response
  - Multiple acceptable answers

Key answer	Score
Viral / Rotavirus gastroenteritis	5
Acute gastroenteritis / Infectious diarrhea	3
Acute diarrhea	0

  - Only one acceptable answer

Key answer	Score
Acute post-streptococcal glomerulonephritis / Post-infectious glomerulonephritis	10
Glomerulonephritis	0

Mahidol University
Faculty of Medicine
Siriraj Hospital

SHEE

Preparing Scoring Keys (2/4)

- Partial credit system

Complete score	คำตอบถูกต้องและสมบูรณ์
Partial score	คำตอบถูกต้องและสมบูรณ์ เพียงบางส่วน
No score	คำตอบไม่ถูกต้อง

Mahidol University
Faculty of Medicine
Siriraj Hospital

SHEE

Preparing Scoring Keys (3/4)

- Partial credit system

e.g. Investigation

Complete score (5)	AST, ALT
Partial score (3)	LFT

Treatment

Complete score (10)	IV Ceftazidime
Partial score (5)	IV 3rd generation cephalosporin
No score (0)	IV antibiotic

Mahidol University Faculty of Medicine Siriraj Hospital

SHEE

Preparing Scoring Keys (4/4)

Penalty

- Absence of “must have” answers
  - score of “0” despite the presence of other less important answers
- Presence of “unnecessary” investigations or treatment
  - no score
  - negative score (but not cross items)
- Harmful treatment
  - negative score (but not cross items)

Mahidol University Faculty of Medicine Siriraj Hospital

SHEE

Time

- ควรกำหนดเวลาให้เพียงพอสำหรับแต่ละคำถาม
  1. อ่านข้อมูลเพิ่มเติมในแต่ละหน้า ที่อาจมีเนื้อหามาก
  2. วิเคราะห์คำถาม
  3. เขียนคำตอบ
- เวลาที่นักศึกษาใช้ในการตอบคำถามนั้นๆ จะมากกว่าเวลาที่อาจารย์ใช้ 30 – 50 %
  - ทดลองตอบคำถามด้วยตนเองและจับเวลา หรือ ให้เพื่อนอาจารย์ทดลองทำ

Mahidol University
Faculty of Medicine
Siriraj Hospital

SHEE

Validation and References

- Validation
  - pilot the problem with colleagues new to the problem: discussion, revision
- References
  - especially in the field of rapidly developing intervention and discovery

Mahidol University
Faculty of Medicine
Siriraj Hospital

SHEE

Conclusion

- CR item is a written test which can be used to measure ability of solving the clinical problems.
- KFQ is one of CR formats that aims to assess clinical decision making skills.
- Developing an MEQ should be based on hypothesis driven approach in clinical problem solving.

the metric of medical education

A practical guide to assessing clinical decision-making skills using the key features approach

Elizabeth A Farmer[1] & Gordon Page[2]

AIM This paper in the series on professional assessment provides a practical guide to writing key features problems (KFPs). Key features problems test clinical decision-making skills in written or computer-based formats. They are based on the concept of critical steps or 'key features' in decision making and represent an advance on the older, less reliable patient management problem (PMP) formats.

METHOD The practical steps in writing these problems are discussed and illustrated by examples. Steps include assembling problem-writing groups, selecting a suitable clinical scenario or problem and defining its key features, writing the questions, selecting question response formats, preparing scoring keys, reviewing item quality and item banking.

CONCLUSION The KFP format provides educators with a flexible approach to testing clinical decision-making skills with demonstrated validity and reliability when constructed according to the guidelines provided.

KEYWORDS *decision making; clinical competence/*standards; educational measurement/*methods/standards; problem-based learning; *education, medical; questionnaires; Canada.

*Medical Education 2005; **39**: 1188–1194*
doi:10.1111/j.1365-2929.2005.02339.x

[1]Royal Australian College of General Practitioners, Melbourne, Victoria, Australia
[2]Department of Medicine, Division of Educational Support and Development, College of Health Disciplines, University of British Columbia, Vancouver, British Columbia, Canada

Correspondence: Associate Professor Elizabeth A Farmer BSc, MBBS, PhD, FRACGP, Department of General Practice, Level 7, Flinders Medical Centre, Bedford Park, South Australia 5042, Australia.
Tel: 00 61 88 204 5606; Fax: 00 61 88 276 3305;
E-mail: liz.farmer@flinders.edu.au

INTRODUCTION

In this article, we introduce the concept of a key feature, which is the cornerstone of a problem format known as the key features problem used in written examinations of clinical decision-making skills.[1] We then focus on practical guidance in creating key features problems to test clinical decision-making skills at both undergraduate and postgraduate levels.

Bordage and Page[2] first introduced the term 'key feature' in 1987, following a critical analysis of research on the nature and assessment of clinical decision-making skills published in 1985.[3] At that time, most assessments of these skills used small numbers of lengthy clinical problems (sometimes only 1), on the premise that the skills were generic and largely independent of the factual knowledge and procedural skills demanded in any particular problem.[4] The most popular such assessment format was the patient management problem (PMP), a written problem which consisted of a clinical scenario, followed by sections of items which elicited candidates' responses in relation to history taking, physical examination, investigations and diagnosis. One PMP could take up to 90 minutes to complete.[5]

Although its high authenticity and face validity made it popular, it became clear that the PMP format had serious drawbacks. First, the reliability of the test was very low[3] and it was evident that content specificity was just as much a factor in testing clinical decision-making skills as in all other areas of clinical competence. In practical terms, this required many hours of testing in order to obtain a reliable result. In addition, the scoring of PMPs often rewarded thoroughness of data gathering, rather than ability to make appropriate decisions. Moreover, the expected differences in performance between junior and experienced doctors were not found. Finally, scores

Overview

What is already known on this subject

The value of testing clinical decision-making skills using the key features problem format has been increasingly recognised over the last decade. The approach is feasible and offers high reliability and support for face and content validity if items are well constructed.

What this study adds

The key features approach is gaining interest amongst educators in health sciences curricula; however, few have practical experience in writing high quality problems. In this paper we present a practical guide to writing and scoring key features problems in health sciences. Various attributes of the approach are highlighted, including the flexibility of the format in testing decision-making skills in a wide variety of domains.

Suggestions for further research

Further examination of predictive validity and effects on candidates' preparation for testing would be valuable.

on PMP tests correlated highly with scores on knowledge tests, suggesting that they added little additional measurement information.[4,6]

A NEW APPROACH

In order to overcome these difficulties, Page and Bordage[6] suggested that, in any clinical case, there are a few unique, essential elements in decision making which, alone or in combination, are the critical steps in the successful resolution of the clinical problem. They labelled these elements 'key features'.[2] This concept led to the creation of a new test of clinical decision-making skills, which elicited candidates' responses concerning only the critical steps in the resolution of each problem – the problem's key features. Testing only critical steps enabled candidates to be tested on a much larger number of clinical problems than was the case with the PMP format. The new test format was called the 'key features problem' (KFP) and was shown to have a potential reliability of 0.8 in 4 hours of testing.[6]

The KFP format proposed by Page and Bordage[6] also added to other written test formats in that it allowed more than 1 correct answer as required by the question. These involved either 1 or more very brief written answers, or 1 or more items selected from a long list. The flexibility in allowing for more than 1 correct answer often mirrors real-life practice more closely than is possible in single answer written formats, such as multiple-choice questions (MCQs) or extended matching questions. In addition, the KFP format also maintained the advantages of the longitudinal nature of the PMP format in that following a problem through various stages enabled testing of candidates' clinical decisions over the course of a clinical scenario. This is similar to other sequential formats, such as the modified essay question format, and again mirrors real-life clinical practice more closely than is possible in more basic test constructions such as MCQs. Key features problem test formats may be presented in either paper-based or computer-based formats. The latter suits high volume, high stakes testing, and allows for low cost incorporation of pictures into the problems, but overall is more expensive to deliver.

Key features problems are now used in a variety of testing situations. While the reliability of the format is good, in high stakes testing the format is presented as part of a suite of assessment approaches. For example, the Medical Council of Canada uses a 4-hour KFP format test in the Part 1 Qualifying Examination for licensure, together with a 3.5-hour MCQ test. Candidates for the Royal Australian College of General Practitioners (RACGP) Fellowship Examination for certification sit a 3-hour KFP paper, together with a 4-hour written test and a 3-hour objective structured clinical examination (OSCE). Key features problem formats are also employed by the University of Toronto as part of its internal examinations for medical students and by the American College of Physicians in the Medical Knowledge Self-Assessment Program (MKSAP) for continuing medical education purposes.

SAMPLE KEY FEATURES PROBLEM: —DIARRHOEA

The following problem (Fig. 1) has been reproduced from a guide to writing KFPs prepared for the

A 35-year-old mother of 3 presents to your office at 17.00 hours with complaints of severe, watery diarrhoea. On questioning, she indicates that she has been ill for about 24 hours. She has had 15 watery bowel movements in the past 24 hours, has been nauseated, but not vomited. She works during the day as a cook in a longterm care facility but left work to come to your office. On her chart, your office nurse notes a resting blood pressure of 105/50 mmHg supine (a pulse of 110/minute), 90/40 standing, and an oral temperature of 36·8 °. On physical examination, you find she has dry mucous membranes and active bowel sounds. A urinalysis (urine microscopy) was normal, with a specific gravity of 1·030.

1 What clinical problems would you focus on in your immediate management of this patient? List up to 3

2 How should you treat this patient at this time? Select up to 3

1	Antidiarrhoeal medication
2	Antiemetic medication
3	Intravenous 0.9% NaCl
4	Intravenous 2/3–1/3
5	Intravenous gentamicin
6	Intravenous metronidazole
7	Intravenous Ringer lactate
8	Nasogastric tube and suction
9	Nothing by mouth
10	Oral ampicillin
11	Oral chloramphenicol
12	Oral fluids
13	Rectal tube
14	Send home with close follow-up
15	Surgical consultation
16	Transfer to hospital

3 After management of the patient's acute condition, what additional measures, if any, would you take? Select up to 4 or select #11, none, if none are indicated

1	Avoid dairy products
2	Colonoscopy
3	Enteric precautions
4	Gastroenterology consultation
5	Give immune serum globulin to patients at longterm care facility
6	Infectious disease consultation
7	Notify Public Health Authority
8	Stool cultures
9	Strict isolation of patient
10	Temporary absence from work
11	None

Figure 1 A sample key features problem.

Medical Council of Canada.[7] The key features tested by the questions are:

1 recognise dehydration (tested) and its level of severity (not tested);
2 manage dehydration appropriately, and
3 evaluate the possible communicability of the underlying disease (family or hospital spread, possible common source).

Each question directly tests 1 of these key features, and each challenges the candidate to apply his or her knowledge in making clinical decisions.

DEVELOPING KEY FEATURES PROBLEMS

The first section of this article highlighted the rationale, nature and main advantages of the key features approach. The sections that follow outline a practical guide to the steps involved in developing KFPs, which build upon the guidelines for writing KFPs presented by Page and Bordage.[1]

Assembling problem-writing groups

Both face validity and content validity require the use of problem writers whose backgrounds and clinical expertise are pertinent to the context of the examination. In Australia, for example, the RACGP employs general practitioners from diverse metropolitan, rural and remote practices across the country, who work in small guided groups to create draft KFPs for use in part of the fellowship examination.[8] This ensures that the problems written are well grounded in practice and experience and represent a wide range of real-life Australian general practice contexts. Using the writing process outlined below, problems are written so that they do not represent mere abstractions or generalisations from textbooks.[9] This is an important step in supporting the content validity of the format and applicability to real-life practice, as perceived by the candidate group.[10]

Selecting a problem, defining its key features

First, problem writers are asked to select a clinical problem (e.g. diarrhoea), usually selected from a blueprint for a key features examination. They are asked to think of several instances (real cases) of the problem in practice. Relative to these cases, they are then asked to address the most important question they face as a problem writer: 'What are the essential steps in the resolution of this problem?'[7] This fundamental question prepares writers to concentrate on only the most critical decisions within each case – the problem's key features. It is essential to differentiate between decisions or steps that are appropriate, but not critical, and those that *must* be present. Coming to grips with this distinction is the

single biggest issue for novice writers. This step usually requires discussion amongst a small group or panel of writers to clarify which steps are critical and achieve consensus. Secondary considerations which can guide the identification of a problem's key features involve asking problem writers to also identify the elements or steps most likely to result in errors by candidates at particular levels of training (e.g. graduating medical students), and to identify the difficult aspects of the identification and management of the problem in clinical practice.

Key features are unique for each clinical problem, and may pertain to any component of the work-up and management of a case; for example, in initial data gathering and diagnostic steps, in longterm management, or in prevention of complications. Key features focus on clinical decisions (e.g. 'include depression in a differential diagnosis') or clinical actions (e.g. 'elicit risk factors', 'order a mammogram') where the clinical action is an expression of a clinical decision. Figure 2 illustrates typical decisions or actions tested in KFPs.

- Elicit history or reasons for patient request
- Interpret symptoms
- Seek critical physical findings
- Interpret physical findings
- Make a diagnosis or differential
- Order investigations to confirm or deny differential diagnoses
- Specify management goals or decisions
- Prescribe drugs
- Specify follow-up

Figure 2 Critical clinical decisions or actions tested in KFPs.

A final component of a key feature is a qualifier that may reflect such issues as the urgency of a decision (e.g. 'What *initial* action…?'), or a decision-making priority (e.g. 'What are the *most important*…?'). Figure 3 presents some common qualifiers.

- Immediate
- Initial
- Longterm
- Definitive
- Urgent
- Most important
- Most likely
- Must not miss

Figure 3 Common qualifiers in key features.

It is important to note that key features may pertain to a broad range of clinical decisions in addition to the biomedical. Key features problems can be constructed to assess ethical, medico-legal, population, preventive and organisational decisions, and in a range of health care settings. This flexibility is a useful attribute of KFP formats in contrast to the more limited multiple-choice and extended matching approaches.

Following their discussion of key features, the problem writers select 1 case for development into a problem scenario and related questions. The clinical scenario for the problem usually begins by stating a patient's age, gender and setting for the encounter. If the key features for that problem focus on the diagnostic component of the problem, the case scenario is often brief (e.g. patient demographics, presenting complaint and limited clinical information). Where the KFP focuses on the management of the problem, the case scenario is typically longer and includes laboratory and diagnostic information. The KFP format is flexible in that additional clinical information can be inserted between questions. This sequential format enables the problem to be followed longitudinally. This attribute allows writers to produce realistic scenarios that evolve over time as required. In this respect, the format is similar to the flexibility found in other sequential formats, such as the modified essay question. Figure 4 gives some examples of the kinds of clinical scenarios that lend themselves to the KFP approach.

- A reason for attendance (e.g. chest pain, check-up, follow-up)
- A request (e.g. sick note, preventive care)
- Symptoms (e.g. cough)
- Signs (e.g. abdominal tenderness)
- Results (e.g. biochemistry, imaging, haematology, audiology, ECG, spirometry)
- Photographs (e.g. clinical signs, rashes)
- Complications of therapy or management

Figure 4 Typical elements in KFP clinical scenarios.

Writing the questions

With the key features defined and the case scenario written, the next step in KFP development is to write the questions that test those key features. Most KFPs consist of a case scenario, typically followed by 2 or 3 questions, each question testing 1 or more key

features. The questions request that candidates record their clinical decisions, which, depending upon the problem's key features, can relate to data gathering (e.g. 'What investigations would you order at this consultation?'), diagnosis ('What are the most likely differential diagnoses?'), management ('What are your longterm management steps?'), etc. Most questions have several answers, which comprise the critical steps in resolving this specific problem. The number of answers may vary from 1 to 10; typically there are 3 to 5.

Selecting question formats

Two question formats are used in KFPs. These are the write-in (WI) format, where candidates supply their responses in very short note form (e.g. they write in 'insulin-dependent diabetes', or 'prescribe penicillin'), and the short menu (SM) format, where candidates select responses from a list of prepared options. The length of the options list varies and may contain up to 25 items. To reduce guessing effects, the list must contain all correct responses plus common misconceptions or likely mistakes. In practice, to reduce cueing, this requires at least 4 or 5 incorrect options for each correct item.

Write-in questions must be marked by hand, whereas SM questions may be marked by computer. The WI question is strictly limited to very short notes or single words, in contrast to the modified essay or short answer question formats, thereby reducing marking time to the minimum. While the feasibility of WI questions could be a problem, data from the Medical Council of Canada and the RACGP suggest that WI formats are more effective in identifying weaker candidates and are more discriminating.[11] In addition, it is often harder to write sequential questions purely in SM formats because of backward cueing of candidates to correct answers. Therefore, most KFPs continue to contain both formats.

Specifying the number of required answers

Each question must contain an instruction that stipulates the number of responses to select or supply. Common instructions are:

- write, in note form only, one (1)...
- select up to 'x'...
- select 'x'...
- select as many as are appropriate, and
- select none if none are indicated.

PREPARING SCORING KEYS

The scoring key for a question consists of the list of correct and incorrect responses, and scores to be assigned to each response.

Some scoring keys can contain only a single required response, such as the scoring key for question 1 of the diarrhoea problem shown in Fig. 1 (Fig. 5).

Score	Response	Synonyms
1	Dehydration	Hypovolaemia fluid loss fluid depletion
0	Listing more than 3 items	

Figure 5 Scoring key for question 1 of the diarrhoea problem shown in Fig. 1.

To emphasise that candidates must not give more than the required number of responses to a question, a forfeit is applied if this occurs. In Fig. 5, up to 3 answers were specified. A candidate who provides say, 4 answers, will receive no marks for the question.

Other scoring keys contain several responses clustered on the basis of logical considerations regarding the correct clinical actions to be taken. A simple scoring key for question 3 of the diarrhoea problem is shown in Fig. 6.

This scoring key illustrates a partial credit system of scoring, where a weight is assigned to each response – in this case the same weight of 1 mark to each response.

Score	Correct responses	
1 each	**# 3**	Enteric precautions
	# 8	Notify Public Health Authority
	# 11	Stool cultures
	# 13	Temporary absence from work
0	**# 5**	Give immune serum globulin to patients at longterm care facility
	# 12	Strict isolation of patient
	or	
	Selecting more than 4 items	

Figure 6 Scoring key for question 3 of the diarrhoea problem shown in Fig. 1.

Specifying different scores for responses allows for the instances where problem writers regard some correct answers as more important clinically than others. Starting with a default option of each correct answer scoring equally, (e.g. 1 point), more important answers may be weighted more highly (e.g. be awarded 2 or even 3 points). Simple weighting systems are preferable, as more complex systems do not improve reliability. Similarly, negative marking is not used because it does not contribute to reliability and may discriminate between students simply on the basis of their risk-taking behaviour.[12] However, an especially important answer can be specified as 'must be present'. In this case a penalty is applied such as 'no marks for the question if answer not present'. Similarly, a dangerous or negligent response (e.g. unnecessary invasive investigation, unnecessary or harmful treatment) may result in the candidate forfeiting the marks for the question involved, no matter what other responses the candidate makes to that question. Items 5 and 12 in the scoring key shown in Fig. 6 are examples of such actions. Such a penalty, if applied, results in the forfeit of marks only for the relevant question within a KFP. In most cases, where a problem consists of 2 or 3 questions, this penalty results in the forfeit of half or a third of the total marks for that problem. Whether or not such an approach is used depends on the views of the examining body and possibly partly on the stakes associated with the examination.

Total examination scores are simply the sum of the scores on each problem. Problem scores are the sum of the scores on the questions within the problem. Each problem is given the same weight in the calculation of the total mark. This can be easily achieved by transforming problem scores into a percentage.

VALIDATION AND REFERENCES

With questions and answer keys defined, the next step is their validation. Validation entails piloting the problem with discussion, review and editing by colleagues new to the problem, and confirmation of the correctness of answers through reference to suitable literature. Markers particularly appreciate evidence from the literature if questions test a new or rapidly developing area. This process is cited as enjoyable and challenging by writers, and the lively debate and sharing of clinical practice contributes to writers' own continuing education.

COMPUTERISED PRESENTATION OF KFP FORMATS

Presenting KFP in a computerised format offers 2 immediate benefits: ease of presentation of high quality pictorial material such as photographs and imaging, and a mechanism to prevent backward cueing if additional clinical information is given between questions. However, this approach requires additional resources.

QUALITY ASSURANCE ISSUES IN ITEM DEVELOPMENT

Problems that perform well can be maintained in an item bank where the performance of a problem in each examination in which it is used may be recorded. Similarly, question writers may receive feedback on the performance of a problem, and may be involved in review of their problems after use. Candidate feedback is another important source of quality assurance.

STANDARD SETTING OF KFP FORMATS

The issues of standard setting for high stakes KFP examinations are comparable to those in other written tests. The Medical Council of Canada uses the modified Angoff method while the RACGP currently employs a new approach, the Angoff at question level (AQL) method. These methods require multiple judges and are based on the concept of the borderline candidate as presented by Norcini in a previous article in the series *the Metric of Medical Education*.[13]

CONCLUSION

Writing key features problems is challenging and enjoyable. Following the steps in this guide will help ensure that KFP examination papers possess high levels of face and content validity and demonstrate levels of test score reliability that are acceptable for making decisions about individual candidates' clinical decision-making ability.

Contributors: EAF and GP conceived the paper. Both authors contributed substantially to writing and revisions. EAF took responsibility for finalising the manuscript. *Acknowledgement:* we thank Brian Jolly for his helpful comments on earlier drafts of the manuscript.

Funding: there was no external funding for this manuscript.
Conflicts of interest: none.
Ethical approval: not required.

REFERENCES

1 Page G, Bordage G, Allen T. Developing key features problems and examinations to assess clinical decision-making skills. *Acad Med* 1995;**70**:194–201.
2 Bordage G, Page G. An alternate approach to PMPs, the key feature concept. In: Hart I, Harden R, eds. *Further Developments in Assessing Clinical Competence.* Montreal: Can-Heal Publications 1987;57–75.
3 Norman G, Bordage G, Curry L *et al.* Review of recent innovations in assessment. In: Wakeford R, ed. *Directions in Clinical Assessment. Report of the Cambridge Conference on the Assessment of Clinical Competence.* Cambridge: Office of the Regius Professor of Physic, Cambridge University School of Clinical Medicine, Addenbrooks Hospital 1985;8–27
4 van der Vleuten C, Newble DI. How can we test clinical reasoning? *Lancet* 1995;**345**:1032–4.
5 McGuire CH, Solomon LM, Bashook PG. *Construction and Use of Written Simulations.* New York: Psychological Corporation of Harcourt, Brace, Jovanovich 1976.
6 Page G, Bordage G. The Medical Council of Canada's key features project: a more valid written examination of clinical decision-making skills. *Acad Med* 1995;**70**:104–10.
7 Page G. *Writing Key Feature Problems for the Clinical Reasoning Skills Examination: a Guide for CRS Committee Members in their Understanding and Preparation of Key Feature Problems.* Ottawa: Medical Council of Canada 1999.
8 Farmer EA. Writing key feature problems for general practice. Melbourne: Royal Australian College of General Practitioners 1998.
9 Jolly B, Spencer J. Letter to the editor: reply from the authors. *Med Educ* 2003;**37**(5):472.
10 Farmer EA, Joske FM, Lew SR, McDonald EA, Page GG. Performance of candidates on key features problems in the certification examination for Australian general practice. [Abstract.] In: *Proceedings of the 10th International Ottawa Conference on Medical Education.* Ottawa, Canada 2002.
11 Page G, Farmer E, Spike N, McDonald E. The use of short answer questions in the key features problems in the Royal College of General Practitioners Fellowship examination. Combining marks, scores and grades. [Abstract.] In: *Proceedings of the 9th International Ottawa Conference on Medical Education.* Cape Town, South Africa 2000.
12 Fowell SL, Jolly B. Reviewing common practices reveals some bad habits. *Med Educ* 2000;**34**:785–6.
13 Norcini JJ. Setting standards on educational tests. The metric of medical education series. *Med Educ* 2003;**37**:464–9.

Received 12 November 2004; editorial comments to authors 7 December 2004, 24 June 2005; accepted for publication 29 July 2005

Medical Teacher

ISSN: 0142-159X (Print) 1466-187X (Online) Journal homepage: https://www.tandfonline.com/loi/imte20

Twelve tips for developing key-feature questions (KFQ) for effective assessment of clinical reasoning

Marla Nayer, Susan Glover Takahashi & Patricia Hrynchak

To cite this article: Marla Nayer, Susan Glover Takahashi & Patricia Hrynchak (2018) Twelve tips for developing key-feature questions (KFQ) for effective assessment of clinical reasoning , Medical Teacher, 40:11, 1116-1122, DOI: 10.1080/0142159X.2018.1481281

To link to this article: https://doi.org/10.1080/0142159X.2018.1481281

MEDICAL TEACHER
2018, VOL. 40, NO. 11, 1116–1122
https://doi.org/10.1080/0142159X.2018.1481281

TWELVE TIPS

Twelve tips for developing key-feature questions (KFQ) for effective assessment of clinical reasoning

Marla Nayer[a], Susan Glover Takahashi[a] and Patricia Hrynchak[b]

[a]University of Toronto, Toronto, ON, Canada; [b]University of Waterloo, Waterloo, ON, Canada

ABSTRACT

Clinical reasoning is the cognitive process that makes it possible for us to reach conclusions from clinical data. "A key feature (KF) is defined as a significant step in the resolution of a clinical problem. Examinations using key-feature questions (KFQs) focus on a challenging aspect in the diagnosis and management of a clinical problem where the candidates are most likely to make errors." KFs have been used at different levels of medical education and practice, from undergraduate to certification examinations. KFQs illuminate the strengths and limits of an individual's clinical problem-solving ability. These types of items are more likely than other forms of assessment to discriminate among stronger or weaker candidates in the area of clinical reasoning. The 12 tips in this article will provide guidance to faculty who wish to develop KFQs for their tests.

Introduction

Clinical reasoning is the cognitive process that makes it possible for us to reach conclusions from clinical data, and come to a clinical decision. "A key feature (KF) is defined as a significant step in the resolution of a clinical problem. Examinations using key-feature questions (KFQs) focus on a challenging aspect in the diagnosis and management of a clinical problem where the candidates are most likely to make errors" (Hrynchak et al. 2014). KFQs have been used for undergraduate medical education, graduate medical education, and licensure examinations (Farmer and Hinchy 2005; Fischer et al. 2005; Leung et al. 2016). KFQs, by their nature, are focused on clinical reasoning and move away from the assessment of rote knowledge or comprehension towards synthesis and evaluation of information in Bloom's cognitive taxonomy (Armstrong 1956; Anderson and Krathwohl 2001; Krathwohl 2002).

Some authors use the terms clinical reasoning and clinical decision making and problem solving interchangeably (Van der Vleuten and Newble 1995; Page 1999 Introduction), or have different definitions of these terms (van Bruggen, Manrique-van Woudenbergh et al. 2012; Durning et al. 2013). For our purposes, clinical reasoning is a concept that reflects the cognitive process. It can include the assessment, diagnosis, and management of a patient. This includes, but is not limited to, clinical decision making (Hrynchak et al. 2014; Escudier et al. 2018). KFQs measure clinical reasoning (Eva 2005; Ilgen et al. 2012).

Research suggests that clinical reasoning skills are specific to the case or problem encountered (case specificity, also referred to as context or content specificity) (Norman et al. 2006). Successful clinical reasoning is contingent on understanding and using the few elements of the problem that are crucial to its successful resolution. KFs represent the critical information needed in the identification or management of a clinical problem. KFQs are focused on case scenarios, often with two to five items for each scenario, and illuminate the strengths and limits of an individual's clinical reasoning. This enables the instructor to have accurate information about the learner's clinical decision making ability. For example, a KFQ will focus on those key elements in a case history that are most likely to lead to a correct diagnosis, either by ruling in or ruling out specific differential diagnoses. These types of items are more likely than other forms of assessment to discriminate among stronger or weaker candidates in the area of clinical reasoning (Schuwirth et al. 2001; Leung et al. 2016).

KFQs have been validated by being administered to practicing clinicians, with positive results. These include physicians (Bordage et al. 1997), and physical therapists and occupational therapists (Glover Takahashi et al. 2012). These types of items appear to have predictive ability for future regulatory complaints (Tamblyn et al. 2007) as well as for quality of care (Wenghofer et al. 2009; Tamblyn et al. 2010). They have been used successfully with clinical clerks (Hatala and Norman 2002; Fischer et al. 2005; Lang et al. 2014), and junior doctors (Leung et al. 2016), as well as in licensure or certification examinations and maintenance of competence programs (Bordage, Brailovsky, et al. 1995; Page and Bordage 1995; Page et al. 1995; Farmer and Hinchy 2005; Lawrence et al. 2011; Glover Takahashi et al. 2012; Brailovsky et al. 2014). They have also been used for jurisprudence content, as well as various intrinsic CanMEDS roles (Royal College of Physicians and Surgeons of Canada 2005): e.g. Communicator, Collaborator, Health Advocate, Scholar, and Professional (Glover Takahashi et al. 2012). Incorporating KFQs into assessment programs will enhance the assessment programs and provide additional information to faculty on learner abilities (Hrynchak et al. 2014).

CONTACT Marla Nayer Marla.nayer@utoronto.ca 500 University Avenue 6th Floor, Toronto, ON M5G 1V7, Canada

Supplemental data for this article can be accessed here.

As with any type of assessment, developing strong items will be central to how well the test functions.

Tip 1

Define the key competencies related to decision making that are to be assessed and create a blueprint

The first step in any examination development is to create an examination blueprint (Downing and Haladyna 2006; Haladyna and Rodriguez 2013). Normally a program of instruction will have established exit-level competencies that each graduate should achieve. Each instructional component will have established learning objectives that are seen to contribute toward the exit-level competencies. These objectives may include professional standards and ethics, as well as diagnosis and management. In most health professions, clinical reasoning (sometimes referred to as problem solving) is a key component of the instructional content, whether it is clinical or addressing professional standards. The frequency of use and importance of each objective will help drive the weighting process of content development and the number of KFQs needed. This will establish content validity of the examination.

The blueprint should be based on the instructional content for the course or program and, for a KF examination, should address the key reasoning areas to be covered. For a very basic example of a blueprint, see Table 1. It is not necessary to fill in every cell in the table, though the *totals* for the rows and columns are important in the creation of an examination. Examples of examinations using blueprints include the Medical Council of Canada (2014), the Medical Council, Ireland (University College Cork Ireland 2015), and the Royal College of Obstetricians and Gynecologists, England. For further information on blueprint development, see the "12 Tips" article on that subject by Coderre et al. (2009).

Tip 2

Choose a clinical presentation or situation

The type of case scenario will depend on the content area and the level of the learner. For a more junior learner, it might be a focused problem or a complaint related to a single system with a typical presentation. For a more advanced learner, it might be an undifferentiated problem or complaint or an atypical presentation, or it might include multisystem involvement.

Many organizations that have developed their own milestones or competency documents [e.g. ACGME milestones (Accreditation Council for Graduate Medical Education (ACGME) and American Board of Pediatrics 2012), the United Kingdom (General Medical Council 2014), Australian Society of Pharmacists (Pharmaceutical Society of Australia 2010), Royal Australian College of General Practitioners (2015), or the Royal College of Physicians and Surgeons of Canada (Frank et al. 2014)]. When such a document is available consider aligning or linking different KFQs to the different stages, milestones or competency statements.

Tip 3

Select the "key feature" level of difficulty that is appropriate for the learners

This is the focus for a KFQ: make sure that the KF is at the appropriate level of difficulty for the level of the learner. KF exams have been used for learners at many levels (Bordage, Brailovsky, et al. 1995; Page and Bordage 1995; Page et al. 1995; Bordage et al. 1997; Hatala and Norman 2002; Farmer and Hinchy 2005; Fischer et al. 2005; Lawrence et al. 2011; Glover Takahashi et al. 2012; Brailovsky et al. 2014; Lang et al. 2014; Leung et al. 2016). Is the learner an undergraduate medical student, a trainee in Internal Medicine, a subspecialty trainee in Cardiology? Each level would require a different KF.

It is necessary to identify the elements or steps most likely to result in errors, the challenging aspects of the identification and management of the problem in clinical practice, or the common misconceptions about the clinical scenario. This is where the writer must differentiate between decisions or steps that are appropriate but not critical, and the steps that *must* be taken to identify and manage the patient's problem. Where are the learners most likely to make an error? What is the challenge in identifying or managing this situation? It is best to make sure that each question deals with a single KF.

An understanding of the common "real-life" misunderstandings and/or errors made by the learners at the different levels comes from experience in teaching and assessing learners at a certain level. This may come out of clinical teaching or from common errors seen on other types of assessments.

Tip 4

Focus the key feature

A KF may pertain to history, physical examination results, other investigations, clinical decision making, management, or the application of professional standards (Page and Bordage 1995; Page et al. 1995; Page 1999; Glover Takahashi et al. 2012, 2013).

The KF should be stated in a single sentence. Some examples: a fourth-year clinical clerk will be able to recognize an anterior ST segment elevation MI on ECG; a junior doctor will recognize the substitute decision-maker hierarchy when a patient is unable to make decisions about

Table 1. Sample blueprint.

	Dimension of Care					
Competency Area	Assessment	Diagnosis	Management	Communication	Professional Behaviour	% of exam
Behavioural Medicine						20%
Surgical Skills						15%
Care of the Elderly						20%
Paediatrics						30%
Obstetrics						15%
% of exam	25%	20%	25%	15%	15%	100%

Table 2. Sample key feature questions in different formats.

Example 1 – Pick-N item

Which of the following are most appropriately considered 'interests' rather than 'positions'? (Pick 2)

A. "We feel that junior doctors should respond to pages in less than 10 minutes"
B. "We want to provide the best care—sometimes we can't wait for a page return."
C. "Junior doctors do not respond to pages from the ward so we call repeatedly."
D. "We all would like the best communication system we can get."
E. "We wait by the phone until calls are returned."

Answers: B & D

Example 2 – Extended Matching item

For the following patients, select the vitamin that is most likely deficient in the patient's diet:

Scenario 1 A 24-year-old woman presents with complaints of fatigue, heart palpitations and a pricking sensation in her toes. She follows a strict vegan diet.

Scenario 2 A 65-year-old patient who is alcoholic presents with difficulty seeing at nighttime. He has dry irritated eyes and keratinized growths (metaplasia) on the conjunctivae.

a. Vitamin A (retinoids)
b. Vitamin B1 (Thiamine)
c. Vitamin B12 (Cobalamin)
d. Vitamin B2 (Riboflavin)
e. Vitamin B3 (Niacin)
f. Vitamin B5 (Pantothenic acid)
g. Vitamin B6 (Pyridoxine)
h. Vitamin B9 (Folic acid)
i. Vitamin C (Ascorbic Acid)
j. Vitamin D (Calciferol, 1,25-dihydroxy vitamin D)
k. Vitamin E (tocopherol)
l. Vitamin H (Biotin)
m. Vitamin K

Answer Scenario 1: d
Answer Scenario 2: a

Example 3 – Fill-in-the-blank

A 78-year-old woman presents to the office on a Friday afternoon at 4:00 pm for an urgent appointment. She is complaining of a sudden onset of blurred and decreased vision in her right eye with distortion. She says that there is no redness or pain in the eye. She has not had any trauma. She has hypertension that is under control but denies any other health conditions.

What is the most likely diagnosis in this case?

Answer: age-related macular degeneration

Example 4 – Matching

Match each drug with the most common side-effect:

a. Drug 1	1. Side effect 1
b. Drug 2	2. Side effect 2
c. Drug 3	3. Side effect 3
d. Drug 4	
e. Drug 5	

Example 5 – Multiple True/False

Indicate whether each of the following are recommendations from Choosing Wisely Canada? (T/F)

a. Recommend routine daily self-glucose monitoring in adults with stable type 2 diabetes (F)
b. Don't routinely order a thyroid ultrasound in patients with abnormal thyroid function tests unless there is a palpable abnormality of the thyroid gland. (T)
c. Use Free T4 or T3 to screen for hypothyroidism or to monitor and adjust levothyroxine (T4) dose in patients with known primary hypothyroidism. (F)
d. Only prescribe testosterone therapy when there is biochemical evidence of testosterone deficiency. (T)
e. Routinely test for Anti-Thyroid Peroxidase Antibodies (anti – TPO). (F)

When using a long menu format (Rotthoff et al. 2006) it is important that the options are single terms and synonyms are accounted for, as well as common misconceptions.

Tip 9

Develop instructions for answering

For each item, there must be clear instructions for how the candidate is to answer the question. Is there one answer? Three? Can they pick as many as they like? Some options include:

- Select up to four
- Which one of the following …
- Select as many as appropriate
- Fill in the blank

"Which one of the following … " works best with the one-best-answer multiple-choice question. The challenge with "select up to … " or "select as many as appropriate" is that candidates find the uncertainty unsettling—they like to know *how many* they should be looking for and selecting. On the other hand, "select as many investigations as appropriate" might work well in assessing resource usage, where a candidate may be penalized for selecting too many investigations. Focus the instructions for answering the KF—what is the main concept/knowledge/skill being assessed? The instructions to be used will often be clear when viewed in reference to the KF listed.

Tip 10

Develop the scoring guideline for each item

Various scoring options have been described for KFQs (Page and Bordage 1995; Page, 1995, p. 162, Farmer and Page 2005; Rotthoff et al. 2006). It is possible to penalize critical errors. Some suggest only scoring if all correct options are selected (Rotthoff et al. 2006); however, part marks can also be used. The part mark approach, as well as the summative versus average scoring approach, have both been shown to provide higher reliability than using a dichotomous score (Page and Bordage 1995).

The various types of scoring include (Hrynchak et al. 2014):

- Dichotomous scoring: 0/1; Partial credit score: number between 0 and 1
- Part mark approach: takes into account the number of incorrect as well as the number of correct responses
- Summative problem scoring: the problem score is the sum of the question scores within a problem
- Averaging problem scoring: the problem score is the average of the question scores within a problem
- Summative approach: each problem score is weighted by the number of questions it contains
- Averaging approach: all problem scores are equally weighted

Examples of scoring might be:

Lead-in: Write down the most important differential diagnosis to rule out.

Scoring: Score 1 for the correct differential. (Note: different terms that refer to the same condition may be granted the same scores.)

Lead-in: Select three steps in the management of this patient.

Scoring: Score 1 point for each correct management; however, if option C is selected, then score the whole item as 0 points, as C is contraindicated for this patient.

Lead-in: Select seven questions to ask on the history.

Scoring: Score 1 for up to five of the following seven options. (Note: full option list includes 15 options; seven options are most important however the item is to be weighted for only 5 correct answers.)

Lead-in: Select as many as appropriate.

Scoring: Score 1 point for up to 5 options; 0 if more than 5 options are selected.

Tip 11

Make sure item-writing guidelines are followed

There are books and articles that outline item-writing guidelines. Case and Swanson (Case and Swanson 1998) is an excellent starting point as is the recently updated version of this guide (Paniagua and Swygert 2016), which is available on line through the National Board of Medical Examiners (NBME) web site.

There are also books and journal articles that address item-writing (Haladyna and Downing 1989a,b, Jozefowicz et al. 2002; Haladyna 2004; Downing and Haladyna 2006; Haladyna and Rodriguez 2013) and there is evidence that faculty development in this area is successful (Abdulghani et al. 2015, 2017; Abozaid et al. 2017; Alamoudi et al. 2017).

Here are some key points for developing items. Always pose a question in a way that allows the candidate to decide on the correct answer without looking at the options. This approach is often called the "hand over" technique (i.e. it is possible to answer even if the options are covered by a hand). Following this tip will prevent having unfocused questions. Avoid, or use extremely sparingly, negatively worded questions; these questions encourage measurement error when able candidates become confused, they are challenging to respond to, and disadvantage those who are writing the examination in a language other than their mother tongue. Avoid frequency terms, such as rarely (how rare is rare?), usually (how often is usually?), or sometimes (once a day? once a week? once a month?).

Tip 12

Consider the words/language used in the items

There is some research that indicates that the language used in items may affect how the learners respond.

Weaker students will perform better when items use medical terminology rather than lay language (Norman et al. 2003; Eva et al. 2010). In some situations, it may be quite appropriate to use lay language (e.g. "a 55-year-old patient comes in to the clinic complaining of coughing up blood"; rather than "a 55-year-old patient comes in to the clinic complaining of haemoptysis"). When reasonable, use the language that the patient would use in solving a patient interaction, and more technical language if interpreting diagnostic findings or reviewing a case with supervisor.

Conclusions

KFQs are a valuable validated assessment approach to assessing the complex knowledge and clinical reasoning that takes place in real-life practice. Developing KFQs requires sophisticated thinking, a deep understanding of candidates' likely responses to questions, an awareness of candidates' perceptions about content, and the ability to write with a high degree of precision. Additional resources on writing KFQs may be found on line (e.g. Medical Council of Canada's guide (Medical Council of Canada 2012), Page's guide (Page 1999), and the Royal Australian College of General Practitioners guide (Farmer 1998; Farmer and Page 2005)).

Integrating KFQs into current systems of assessment would add value by promoting clinical reasoning, as well as identifying learners who have gaps in their ability to apply content knowledge.

Disclosure statement

The authors report no conflicts of interest. The authors alone are responsible for the content and writing of this article.

Notes on contributors

Dr. Marla Nayer, PhD, is an assessment consultant, working in postgraduate medical education and teaches a graduate level assessment course at University of Toronto.

Dr. Glover Takahashi, PhD, works in postgraduate medical education and teaches a graduate level assessment course at University of Toronto.

Dr. Patricia Hrynchak, OD, is a clinical professor at the School of Optometry and Vision Science, University of Waterloo.

ORCID

Marla Nayer http://orcid.org/0000-0002-3249-3140
Susan Glover Takahashi http://orcid.org/0000-0003-0722-7876
Patricia Hrynchak http://orcid.org/0000-0002-3187-0338

References

Abdulghani HM, Ahmad F, Irshad M, Khalil MS, Al-Shaikh GK, Syed S, Aldrees AA, Alrowais N, Haque S. 2015. Faculty development programs improve the quality of Multiple Choice Questions items' writing. Sci Rep. 5:9556.

Abdulghani HM, Irshad M, Haque S, Ahmad T, Sattar K, Salah Khalil M. 2017. Effectiveness of longitudinal faculty development programs on MCQs items writing skills: A follow-up study. Plos One. 12: e0185895.

Abozaid H, Park YS, Tekian A. 2017. Peer review improves psychometric characteristics of multiple choice questions. Med Teach. 1–5.

Accreditation Council for Graduate Medical Education (ACGME) and American Board of Pediatrics. 2012. The Pediatrics Milestone Project; [accessed 2017 Aug 4]. https://acgme.org/Portals/0/PDFs/Milestones/PediatricsMilestones.pdf.

Alamoudi AA, El-Deek BS, Park YS, Al Shawwa LA, Tekian A. 2017. Evaluating the long-term impact of faculty development programs on MCQ item analysis. Med Teach. 39(sup1):S45–S49.

Anderson LW, Krathwohl D, editors. 2001. A taxonomy for learning, teaching, and assessing: a revision of Bloom's taxonomy of educational objectives. New York, NY: Longman Publishers.

Armstrong P. Bloom's Taxonomy; [accessed 2017 Jul 12]. https://cft.vanderbilt.edu/guides-sub-pages/blooms-taxonomy/.

Bloom BS. 1956. Taxonomy of educational objectives: The classification of educational goals. New York, NY: McKay.

Bordage G, Brailovsky C, Cohen T, Page GG. 1997. Maintaining and enhancing key decision-making skills from graduation into practice: An exploratory study. Seventh Ottawa Conference on Medical Education and Assessment: Advances in Medical Education, Maastricht, The Netherlands: Kluwer Academic Publishers.

Bordage G, Brailovsky C, Carretier H, Page G. 1995. Content validation of key features on a national examination of clinical decision-making skills. Acad Med. 70:276–281.

Bordage G, Carretier H, Bertrand R, Page G. 1995. Comparing times and performances of French- and English-speaking candidates taking a national examination of clinical decision-making skills. Acad Med. 70:359–365.

Brailovsky C, Allen T, Lawrence K, Crichton T, Laughlin T, Van der Goes T. 2014. Short answer questions based on Key Features have higher discrimination indices on a certification examination in family medicine Ottawa Conference. Ottawa, ON.

Case SM, Swanson DB. 1998. Writing written test questions for the basic and clinical sciences. Philadelphia, PA: National Board of Medical Examiners.

Cerutti B, Blondon K, Galetto A. 2016. Long-menu questions in computer-based assessments: a retrospective observational study. BMC Med Educ. 16:55.

Coderre S, Woloschuk W, McLaughlin K. 2009. Twelve tips for blueprinting. Med Teach. 31:322–324.

Downing S, Haladyna TA. 2006. Handbook of test development. Mahwah, NJ: Lawrence Erlbaum Assoc. Inc.

Durning SJ, Artino AR, Schuwirth L, van der Vleuten C. 2013. Clarifying assumptions to enhance our understanding and assessment of clinical reasoning. Acad Med. 88:442–448.

Escudier M, Woolford M, Tricio J. 2018. Assessing the application of knowledge in clinical problem solving: The structured professional reasoning exercise. Eur J Dent Educ. 22:e269–e277.

Eva K. 2005. What every teacher needs to know about clinical reasoning. Med Educ. 39:98–106.

Eva KW, Wood TJ, Riddle J, Touchie C, Bordage G. 2010. How clinical features are presented matters to weaker diagnosticians. Med Educ. 44:775–785.

Farmer E. 1998. Writing key feature problems. Australia: Royal Australian College of General Practitioners. [accessed 2018 June 12]. https://www.academia.edu/1749144/Writing_Key_Features_Problems.

Farmer EA, Hinchy J. 2005. Assessing general practice clinical decision making skills: the key feature approach. Austr Fam Phys 34: 1059–1061.

Farmer EA, Page G. 2005. A practical guide to assessing clinical decision-making skills using the key features approach. Med Educ. 39:1188–1194.

Fischer MR, Kopp V, Holzer M, Ruderich F, Junger J. 2005. A modified electronic key feature examination for undergraduate medical students: validation threats and opportunities. Med Teach. 27:450–455.

Frank JR, Snell LS, Sherbino J. 2014. The Draft CanMEDS 2015 Milestones Guide; [accessed 2017 Aug 4]. http://www.royalcollege.ca/portal/page/portal/rc/common/documents/canmeds/framework/canmeds_milestone_guide_sept2014_e.pdf.

General Medical Council. 2014. Good medical practice. United Kingdom: General Medical Council.

Glover Takahashi S, Herold J, Clark M, Nayer M, Beggs C, Corbett C, Drynan D, Cho N, Dignum T, Hudson B, Corbett K. 2012. The use of key features cases to assess clinical decision-making, CanMEDS roles & competence First Montreal Conference on Clinical Reasoning. Montreal, QC.

Glover Takahashi S, Herold J, Clark M, Nayer M, Drynan D, Cho N, Dignum T, Corbett K, Hudson B, Hynes M. 2013. Building better written exams – The use of key features cases to assess clinical decision-making, CanMEDS roles and competence International Conference on Residency Education (ICRE). Calgary, Alberta.

Haladyna TA, Downing S. 1993. How many options is enough for a multiple-choice teste item. Educ Psychol Meas. 53:999–1010.

Haladyna TM. 2004. Developing and validating multiple-choice test items. Hillsdale, NJ: Lawrence Erlbaum Associates.

Haladyna TM, Downing SM. 1989a. A taxonomy of multiple-choice item-writing rules. Appl Meas Educ. 2:37–50.

Haladyna TM, Downing SM. 1989b. Validity of a taxonomy of multiple-choice item-writing rules. Appl Meas Educ. 2:51–78.

Haladyna TM, Rodriguez MC. 2013. Developing and validating test items. New York, NY: Routledge Taylor & Francis Group.

Hatala R, Norman GR. 2002. Adapting the Key Features Examination for a clinical clerkship. Med Educ. 36:160–165.

Hrynchak P, Glover Takahashi S, Nayer M. 2014. Key-feature questions for assessment of clinical reasoning: a literature review. Med Educ. 48:870–883.

Huwendiek S, Reichert F, Duncker C, de Leng BA, van der Vleuten CPM, Muijtjens AMM, Bosse HM, Haag M, Hoffmann GF, Tönshoff B, Dolmans D. 2017. Electronic assessment of clinical reasoning in clerkships: A mixed-methods comparison of long-menu key-feature problems with context-rich single best answer questions. Med Teach. 39:476–485.

Ilgen J, Humbert A, Kuhn G, Hansen M, Norman G, Eva KW, Charlin B, Sherbino J. 2012. Assessing diagnostic reasoning: a consensus statement summarizing theory, practice, and future needs. Acad Emerg Med. 19:1454–1461.

Jozefowicz RF, Koeppen BM, Case S, Galbraith R, Swanson D, Glew RH. 2002. The quality of in-house medical school examinations. Acad Med. 77:156–161.

Kilgour JM, Tayyaba S. 2016. An investigation into the optimal number of distractors in single-best answer exams. Adv Health Sci Educ Theory Pract. 21:571–585.

Krathwohl D. 2002. A revision of Bloom's taxonomy: An overview. Theory Pract. 41:212–218.

Lang AJ, Bronander K, Harrell H, Kovach R, Monteiro S, Bordage G. 2014. Validity evidence for a key features examination to assess clinical decision making in the internal medicine clerkship 16th Ottawa Conference. Ottawa, ON.

Lawrence K, Allen Brailovsky T, Crichton C, Bethune T, Donoff C, Laughlin M, Wetmore TS, Carpentier M-P, Visser S. 2011. Defining competency-based evaluation objectives in family medicine: key-feature approach. Canadian Family Physician 57:e373–e380.

Leung F-H, Herold J, Iglar K. 2016. Family medicine mandatory assessment of progress: results of a pilot administration of a family medicine competency-based in-training examination. Can Fam Physician 62:e263–e267.

Medical Council of Canada. 2012. Guidelines for the Development of Key Feature Problems and Test Cases; [accessed 2017 Jan 26] http://mcc.ca/wp-content/uploads/cdm-guidelines.pdf.

Medical Council of Canada. 2014. Blueprint Project: Qualifying examinations blueprint and content specifications. Ottawa, ON, Medical Council of Canada.

Norman GR, Arfai B, Gupta A, Brooks LR, Eva KW. 2003. The privileged status of prestigious terminology: impact of "medicalese" on clinical judgments. Acad Med. 78:S82–S84.

Norman G, Bordage G, Page G, Keane D. 2006. How specific is case specificity? Med Educ. 40:618–623.

Page GG. 1999. Writing key feature problems for the clinical reasoning skills examination; [accessed 2017 Jan 26]. http://www.idealmed.org/workshop/SectionD-KeyFeatures.pdf.

Page GG, Bordage G. 1995. The Medical Council of Canada's Key Features Project: A more valid written examination of clinical decision-making skills. Acad Med. 70:104–110.

Page GG, Bordage G, Allen T. 1995. Developing key-feature problems and examinations to assess clinical decision-making skills. Acad Med. 70:194–201.

Paniagua MA, Swygert KA. 2016. Writing written test questions for the basic and clinical sciences. Philadelphia, PA: National Board of Medical Examiners.

Pharmaceutical Society of Australia. 2010. National Competency Standards Framework for Pharmacists in Australia. Australia: Pharmaceutical Society of Australia.

Piasentin KA. 2010. Exploring the optimal number of options in multiple-choice testing. CLEAR Exam Rev (Winter). 18–22.

Rodriguez MC. 2005. Three options are optimal for multiple-choice items: a meta-analysis of 80 years of research. Educ Meas Issues Prac (Summer). 24:3–13.

Rotthoff T, Baehring T, Dicken HD, Fahron U, Richter B, Fischer MR, Scherbaum WA. 2006. Comparison between Long-Menu and Open-Ended Questions in computerized medical assessments. A randomized controlled trial. BMC Med Educ. 6:50.

Royal Australian College of General Practitioners. 2015. Competency profile of the Australian general practitioner at the point of Fellowship. Australia: Royal Australian College of General Practitioners. [accessed 2018 June 12] https://www.racgp.org.au/download/Documents/VocationalTrain/Competency-Profile.pdf

Royal College of Physicians and Surgeons of Canada. 2005. CanMEDS 2005 Framework. Ottawa, ON: Royal College of Physicians and Surgeons of Canada.

Schneid SD, Armour C, Park YS, Yudkowsky R, Bordage G. 2014. Reducing the number of options on multiple-choice questions: response time, psychometrics and standard setting. Med Educ. 48:1020–1027.

Schuwirth LW, Verheggen MM, van der Vleuten CP, Boshuizen HP, Dinant GJ. 2001. Do short cases elicit different thinking processes than factual knowledge questions do? Med Educ. 35:348–356.

Tamblyn R, Abrahamowicz M, Dauphinee D, Wenghofer E, Jacques A, Klass D, Smee S, Blackmore D, Winslade N, Girard N, et al. 2007. Physician scores on a national clinical skills examination as predictors of complaints to medical regulatory authorities. JAMA 298:993–1001.

Tamblyn R, Abramowicz M, Dauphinee D, Wenghofer E, Jacques A, Klass D, Smee S, Eguale T, Winslade N, Girard N, Bartman I, Buckeridge D, Hanley J. 2010. Influence of physicians' management and communication ability on patients' persistence with antihypertensive medication. Arch Intern Med. 170:1064–1072.

The Royal College of Obstetricians and Gynecologists. Blueprint Grid for the Membership of the Royal College of Obstetricians and Gynecologists Examination; [accessed 2018 Feb 8]. https://www.rcog.org.uk/globalassets/documents/careers-and-training/mrcog-exam/part-1/ex-part-1-blueprinting-grid-new.pdf.

University College Cork Ireland. 2015. How to Use the Draft Blueprint for the Pre-Registration Examinations (PRES) Level 3; [accessed 2018 Feb 8]. https://www.medicalcouncil.ie/Information-for-Doctors/Examinations-/How-to-use-the-Blueprint-for-the-PRES.pdf.

van Bruggen L, Manrique-van Woudenbergh M, Spierenburg E, Vos J. 2012. Preferred question types for computer-based assessment of clinical reasoning: a literature study. Perspect Med Educ. 1:162–171.

Van der Vleuten C, Newble D. 1995. How can we test clinical reasoning? Lancet. 345:1032–1034.

Wenghofer E, Klass D, Abrahamowicz M, Dauphinee D, Jacques A, Smee S, Blackmore D, Winslade N, Reidel K, Bartman I, Tamblyn R. 2009. Doctor scores on national qualifying examinations predict quality of care in future practice. Med Educ. 43:1166–1173.

World Health Organization (WHO). 2016. International Classification of Diseases, ICD10; [accessed 2018 Mar 28]. http://apps.who.int/classifications/icd10/browse/2016/en#/V.

การสร้างข้อสอบอัตนัยประยุกต์

ผู้ช่วยศาสตราจารย์ นายแพทย์เชิดศักดิ์ ไอรมณีรัตน์ พ.บ., ป.ชั้นสูง (ศัลยศาสตร์), ว.ว. ศัลยศาสตร์, MHPE, Ph.D.
ภาควิชาศัลยศาสตร์, คณะแพทยศาสตร์ศิริราชพยาบาล, มหาวิทยาลัยมหิดล, กรุงเทพมหานคร 10700.

ข้อสอบอัตนัยประยุกต์ (modified essay question, MEQ) เป็นรูปแบบการประเมินผลที่นิยมใช้กับนักศึกษาแพทย์ระดับคลินิกเพื่อประเมินความสามารถในการแก้ปัญหา และตัดสินใจเลือกการตรวจรักษาที่เหมาะสมสำหรับผู้ป่วย ในปัจจุบันมีการใช้ข้อสอบอัตนัยประยุกต์ในการสอบของนักศึกษาแพทย์ในหลายภาควิชา รวมทั้งใช้ในการสอบขั้นตอนที่สามของการประเมินความรู้ความสามารถในการประกอบวิชาชีพเวชกรรม ของแพทยสภาด้วย อย่างไรก็ตาม จากการติดตามเนื้อหาของโจทย์ข้อสอบอัตนัยประยุกต์ ร่วมกับการพิจารณาเกณฑ์การให้คะแนนของข้อสอบเหล่านี้ที่ใช้กับการสอบของนักศึกษาแพทย์ในหลายการสอบ ผู้นิพนธ์ยังคงพบเห็นปัญหาในการสร้างข้อสอบชนิดนี้อยู่พอสมควร บทความนี้จึงได้รับการเขียนขึ้นเพื่อสร้างความเข้าใจในหลักการพื้นฐาน และแนวปฏิบัติที่เหมาะสมในการสร้างข้อสอบอัตนัยประยุกต์สำหรับการประเมินความรู้ทางการแพทย์

ลักษณะพื้นฐานของข้อสอบอัตนัยประยุกต์

ข้อสอบอัตนัยประยุกต์เป็นรูปแบบหนึ่งของข้อสอบอัตนัย (Essay question) ซึ่งในรูปแบบดั้งเดิม (traditional essay) นั้นผู้ออกข้อสอบจะเขียนโจทย์คำถามแล้วให้ผู้สอบเขียนคำตอบด้วยตนเองในขั้นตอนเดียว โดยไม่มีตัวเลือกให้ ในการเขียนคำตอบอาจเขียนตอบเป็นคำ หรือวลีสั้น ๆ (Short essay) หรือ ตอบเป็นบทความที่มีความยาวเป็นย่อหน้า หรือ หลายย่อหน้า (Long essay) ซึ่งผู้ออกข้อสอบคาดว่าการสอบในลักษณะที่ผู้สอบไม่มีตัวเลือก แต่ต้องคิดคำตอบด้วยตนเองนี้จะสามารถวัดความรู้ขั้นสูงในระดับการวิเคราะห์ สังเคราะห์ หรือประเมินคุณค่าได้[1,2]

อย่างไรก็ตามข้อสอบในรูปแบบอัตนัยแบบดั้งเดิมนั้นประสบปัญหาในการใช้ประเมินความรู้ทางการแพทย์อยู่หลายประการ ทั้งความยากในการตรวจให้คะแนน ความจำกัดในปริมาณเนื้อหาที่สามารถสอบได้ในเวลาที่มี ความเห็นที่แตกต่างกันของผู้ตรวจให้คะแนน ความไม่เที่ยงของคะแนนสอบ เป็นต้น[1,2] ปัญหาที่สำคัญยิ่งที่ทำให้การสอบอัตนัยแบบดั้งเดิมไม่ได้รับความนิยมในการประเมินความรู้ในระดับคลินิกคือ การที่ข้อสอบอัตนัยแบบดั้งเดิมนั้นมักวัดความรู้ในระดับการท่องจำ หรือความเข้าใจพื้นฐานเท่านั้น และรูปแบบการคิดวิเคราะห์เพื่อตอบโจทย์ข้อสอบอัตนัยแบบดั้งเดิมนั้นมีลักษณะแตกต่างไปจากกระบวนการแก้ปัญหาในระดับคลินิกที่แพทย์ปฏิบัติจริง

ข้อสอบอัตนัยแบบดั้งเดิมที่ดีนั้นผู้ออกข้อสอบสามารถประเมินทักษะการคิดวิเคราะห์ขั้นสูงได้ แต่อุปสรรคสำคัญที่ทำให้ไม่สามารถบรรลุวัตถุประสงค์ดังกล่าวได้คือการสร้างข้อสอบที่ผู้สอบตั้งเป้าหมายให้ตรวจให้คะแนนได้ง่ายเป็นสำคัญ ทำให้ข้อสอบอัตนัยแบบดั้งเดิมส่วนใหญ่ทำการประเมินเพียงความรู้ระดับความจำหรือความเข้าใจพื้นฐานเท่านั้น

สมมติฐานพื้นฐานในการตอบข้อสอบอัตนัยแบบดั้งเดิมคือการวิเคราะห์และหาแนวทางแก้ปัญหาเป็นกระบวนการที่ทำในขั้นตอนเดียว ดังนั้นข้อสอบจึง

นำเสนอข้อมูลทั้งหมดในขั้นตอนเดียวแล้วให้ผู้เข้าสอบแสดงการวิเคราะห์และแก้ปัญหา ซึ่งเป็นกระบวนการแก้ปัญหาทางคลินิกที่แพทย์ใช้ในกรณีเจอผู้ป่วยที่ไม่ซับซ้อนที่ไม่ต้องการกระบวนการคิดวิเคราะห์ขั้นสูงมากนัก อย่างไรก็ตามปัญหาผู้ป่วยที่มีความซับซ้อนและต้องการวิเคราะห์มากมักต้องการกระบวนการแก้ปัญหาหลายขั้นตอน แพทย์จะต้องทำการประเมินข้อมูลพื้นฐานที่ได้จากผู้ป่วย แล้วซักประวัติ หรือตรวจร่างกายเพื่อเก็บข้อมูลเพิ่มเติมอย่างเหมาะสม เมื่อได้ข้อมูลพื้นฐานมาแล้ว แพทย์ต้องทำการตั้งสมมติฐานถึงโรคที่ผู้ป่วยน่าจะเป็น แล้วทำการสืบค้นเพิ่มเติมด้วยการตรวจทางห้องปฏิบัติการหรือใช้ภาพถ่ายรังสี ในบางกรณีแพทย์จำเป็นต้องให้การรักษาเบื้องต้นก่อน พร้อมกับทำการสืบค้นเพิ่มเติม ซึ่งเมื่อเวลาผ่านไปแพทย์จะได้รับข้อมูลของผู้ป่วยมากขึ้นเรื่อย ๆ จากผลตรวจทางห้องปฏิบัติการ หรือการตอบสนองต่อการรักษาที่ให้ เมื่อได้ข้อมูลมากขึ้นแพทย์จะต้องทำการประเมินสถานการณ์ใหม่ ข้อมูลที่เพิ่มขึ้นอาจทำให้แพทย์สามารถให้การวินิจฉัยที่แน่ชัด และวางแผนการรักษาที่เหมาะสมได้ จะเห็นได้ว่ากระบวนการแก้ปัญหาของแพทย์มักทำเป็นหลายขั้นหลายตอน แต่ละขั้นตอนจะได้ข้อมูลเพิ่มขึ้นเรื่อย ๆ การตัดสินใจในแต่ละขั้นเมื่อได้เลือกที่จะตรวจหรือให้การรักษาใดแก่ผู้ป่วยแล้ว ไม่สามารถย้อนเวลากลับไปแก้ไขการตัดสินใจที่ทำผิดพลาดไปก่อนหน้านี้ได้

จากข้อจำกัดของข้อสอบอัตนัยแบบดั้งเดิมที่กล่าวมาข้างต้น ทำให้มีการพัฒนารูปแบบการสอบเป็นข้อสอบอัตนัยประยุกต์ (modified essay question, MEQ) ซึ่งเป็นข้อสอบที่เริ่มจากการให้สถานการณ์ของผู้ป่วย แล้วมีโจทย์ถามให้ผู้สอบตอบคำถามที่เกี่ยวกับการแก้ปัญหาผู้ป่วยในสถานการณ์นั้นโดยไม่มีตัวเลือกให้ เมื่อผู้สอบตอบคำถามแล้วจะมีการเปิดเผยข้อมูลเพิ่มเติมเกี่ยวกับผู้ป่วยมากขึ้นทีละน้อย และมีโจทย์ถามคำถามเพิ่มเติมเป็นลำดับ โดยที่ผู้สอบไม่มีโอกาสย้อนกลับไปแก้ไขคำตอบของตนเองที่ได้ตอบไปในขั้นตอนก่อนหน้านั้น[1,3] รูปแบบของข้อสอบอัตนัยประยุกต์ที่นิยมใช้กันมากในยุคแรก ๆ มีลักษณะเป็นการสอบถามกระบวนการดูแลผู้ป่วยตั้งแต่ต้นจนจบในรูปแบบที่เรียกว่าการจัดการปัญหาของผู้ป่วย (Patient management problem, PMP)[1, 4, 5]

เนื่องจากข้อสอบอัตนัยประยุกต์ที่ใช้ในทางการแพทย์มักมุ่งเน้นการประเมินทักษะการวินิจฉัยโรค ผู้นิพนธ์จึงขอทบทวนทฤษฎีเกี่ยวกับกระบวนการวินิจฉัยโรคสักเล็กน้อยก่อนนำเข้าสู่หลักการสร้างข้อสอบ โดยทั่วไปแล้ววิธีการที่แพทย์ใช้ในการวินิจฉัยโรคมีสามวิธีหลักได้แก่ (1) วิธีจำได้จากแบบแผนของความผิดปกติที่พบ (pattern recognition), (2) วิธีปฏิบัติตามขั้นตอนวิธีที่มีแบบแผน (algorithm), และ (3) วิธีทดสอบสมมติฐาน (hypothesis testing)[6] ซึ่งในวิธีทดสอบสมมติฐานนี้สามารถแบ่งออกเป็นวิธีการย่อยได้สองวิธีคือ (3.1) การแก้ปัญหาด้วยวิธีอุปนัย (inductive reasoning) ซึ่งแพทย์จะรวบรวมข้อมูลอย่างครบถ้วนตามแบบแผนก่อนจึงตั้งสมมติฐาน และ (3.2) การแก้ปัญหาด้วยวิธีนิรนัย (deductive reasoning) ซึ่งแพทย์จะเริ่มตั้งสมมติฐานตั้งแต่เมื่อเริ่มเก็บข้อมูลจากผู้ป่วยเพียงเล็กน้อย แล้วใช้สมมติฐานที่ได้มานั้นเป็นแนวทางในการซักประวัติ และตรวจร่างกายอย่างมีจุดหมายเพื่อทดสอบสมมติฐานที่ตั้งขึ้นจนค่อย ๆ ตัดโรคที่ไม่สอดคล้องกับข้อมูลที่ได้รับออกไปเรื่อย ๆ โดยทั่วไปแล้ววิธีอุปนัยเป็นวิธีที่มีประสิทธิภาพน้อยกว่าวิธีนิรนัย เนื่องจากการเก็บข้อมูลเป็นไปอย่างขาดจุดหมายทำให้เสียเวลาและอาจพลาดการเก็บข้อมูลที่สำคัญไป[6]

การสร้างข้อสอบอัตนัยประยุกต์ที่มีคุณภาพดีควรเริ่มจากความเข้าใจในปรัชญาพื้นฐานของการประเมินผลว่าข้อสอบอัตนัยประยุกต์นั้นได้รับการพัฒนาขึ้นเพื่อประเมินทักษะการแก้ปัญหาด้วยวิธีนิรนัยเป็นสำคัญ ข้อผิดพลาดที่พบบ่อยของการสร้างข้อสอบอัตนัยประยุกต์ประการหนึ่งคือการสร้างข้อสอบที่ให้ข้อมูลผู้ป่วยสั้นมาก (จนไม่มีทางตั้งสมมติฐานที่ชัดเจนได้) แล้วตั้งโจทย์ให้ผู้เข้าสอบเขียนรายการประวัติที่จะสอบถามหรือการตรวจร่างกายที่จะดำเนินการในผู้ป่วยดังกล่าว เช่น ให้สถานการณ์เป็นหญิงอายุ 45 ปี ปวดท้อง 1 วัน แล้วตั้งโจทย์ว่า จงทำการซักประวัติที่เหมาะสม ซึ่งการให้สถานการณ์ในลักษณะนี้มีโรคที่สามารถเป็นไปได้มากมาย ในหลายระบบ สิ่งที่จะประเมินได้จากการตอบ

คำถามลักษณะนี้คือความจำขั้นพื้นฐาน (simple recall) ว่าแบบแผนการซักประวัติผู้ป่วยปวดท้องเฉียบพลันมีอะไรบ้าง ซึ่งผู้เข้าสอบเขียนอะไรมาก็น่าจะถูกหมด ไม่มีการซักประวัติที่ไม่เข้าประเด็น เนื่องจากข้อมูลจากโจทย์ไม่มีรายละเอียดมากพอที่จะจำกัดโรคที่ควรนึกถึง ข้อสอบอัตนัยประยุกต์ที่ดีควรเริ่มจากข้อมูลที่สามารถสร้างสมมติฐานที่ชัดเจนพอได้ เช่น หญิงอายุ 50 ปี จุกแน่นลิ้นปี่และใต้ชายโครงขวาเป็น ๆ หาย ๆ 4 เดือน มีอาการปวดท้องใต้ชายโครงขวามาก ร่วมกับมีไข้ต่ำ ๆ 7 ชั่วโมง การให้ข้อมูลที่มีรายละเอียดพอสมควรนี้ผู้สอบที่มีความรู้จะตั้งสมมติฐานได้ว่าผู้ป่วยน่าจะเป็นโรคใด หากโจทย์กำหนดให้ซักประวัติเพิ่มเติม ผู้สอบที่มีความรู้จะสามารถสอบถามอาการที่สอดคล้องกับการวินิจฉัยที่เหมาะสมได้ ในกรณีนี้คำตอบที่ไม่สอดคล้อง (เช่นสมมติฐานที่เหมาะสมคือภาวะถุงน้ำดีอักเสบเฉียบพลัน แต่ผู้สอบซักประวัติประจำเดือน ประวัติเพศสัมพันธ์) ไม่ควรได้คะแนน

พัฒนาการของข้อสอบอัตนัยประยุกต์

หลังจากที่มีรายงานการใช้ข้อสอบอัตนัยประยุกต์ในการประเมินผลทางแพทยศาสตรศึกษาตั้งแต่ปี พ.ศ. 2514 โดยราชวิทยาลัยแพทย์เวชปฏิบัติทั่วไปเพื่อประเมินทักษะการแก้ปัญหาทางคลินิกแล้ว[3,7,8] ข้อสอบอัตนัยประยุกต์ก็ได้ถูกใช้ในการประเมินทางการแพทย์และสาธารณสุขในหลากหลายบริบท[9-12] โดยรูปแบบที่เป็นที่นิยมกันมากเป็นการสอบถามกระบวนการดูแลผู้ป่วยตั้งแต่ต้นจนจบในรูปแบบ การจัดการปัญหาของผู้ป่วย (Patient management problem, PMP) ซึ่งการแก้ปัญหาผู้ป่วยแต่ละรายมักใช้เวลานานมาก ทำให้การสอบแต่ละครั้งมักมีจำนวนสถานการณ์ผู้ป่วยที่นำมาสอบไม่มากนัก[13]

จากการใช้ข้อสอบอัตนัยประยุกต์ในรูปแบบการจัดการปัญหาของผู้ป่วย พบว่ามีข้อจำกัดบางประการกล่าวคือ ข้อสอบส่วนใหญ่มุ่งเน้นวัดความครบถ้วนสมบูรณ์ของคำตอบมากกว่าการตัดสินใจแก้ปัญหา จำนวนสถานการณ์ผู้ป่วยที่มีจำนวนน้อยทำให้ไม่สามารถครอบคลุมองค์ความรู้ที่ต้องการประเมินได้ครบ และความเที่ยงของคะแนนสอบที่ต่ำ[4,13,14] ปัญหาที่สำคัญยิ่งในการสอบด้วยสถานการณ์ผู้ป่วยจำนวนน้อยคือ ทักษะในการแก้ปัญหาทางคลินิกมีความจำเพาะต่อบริบทของผู้ป่วยแต่ละราย (case specificity)[15-18] การที่ผู้เข้าสอบสามารถแก้ปัญหาผู้ป่วยที่มีอาการเจ็บหน้าอกได้ดีนั้นไม่สามารถจะบอกได้ว่าผู้เข้าสอบคนดังกล่าวจะสามารถแก้ปัญหาผู้ป่วยที่มีอาการปวดศีรษะได้ดีด้วยหรือไม่ ดังนั้นหลักการที่สำคัญประการหนึ่งในการสร้างข้อสอบอัตนัยประยุกต์ก็คือการจัดทำข้อสอบให้มีหลากหลายสถานการณ์ เพื่อให้สามารถประเมินการแก้ปัญหาของผู้เข้าสอบได้ในหลากหลายบริบท ในหลายระบบอวัยวะ จากปัญหาในการใช้ข้อสอบอัตนัยประยุกต์ต่าง ๆ เหล่านี้ ทำให้นักการศึกษามีการพัฒนารูปแบบข้อสอบอัตนัยประยุกต์ให้ต่างไปจากรูปแบบดั้งเดิม รูปแบบข้อสอบที่ผู้เชี่ยวชาญในการประเมินผลแนะนำในปัจจุบันคือ การแก้ปัญหาสำคัญ (key features problems, KFP)

ข้อสอบอัตนัยประยุกต์แบบแก้ปัญหาสำคัญนี้ได้รับการพัฒนาบนหลักการสำคัญคือในการแก้ปัญหาผู้ป่วยแต่ละรายมีประเด็นปัญหาที่เป็นหัวใจสำคัญเพียงไม่กี่ประเด็นเท่านั้น ซึ่งประเด็นปัญหาเหล่านี้เรียกว่าปัญหาสำคัญ (key features)[19] ซึ่งในผู้ป่วยแต่ละรายจะมีปัญหาสำคัญที่แพทย์ต้องให้ความสนใจต่างกันไป บางรายเป็นเรื่องการซักประวัติ บางรายเป็นการเลือกการส่งตรวจทางห้องปฏิบัติการ ในขณะที่บางรายเป็นการตัดสินใจเลือกวิธีการรักษาที่เหมาะสม เป็นต้น ในข้อสอบอัตนัยประยุกต์รูปแบบการแก้ปัญหาสำคัญจะมุ่งเน้นตั้งโจทย์ถามเฉพาะประเด็นปัญหาสำคัญเหล่านี้เท่านั้น ไม่จำเป็นต้องถามกระบวนการดูแลผู้ป่วยตั้งแต่ต้นจนจบในผู้ป่วยทุกราย การสร้างข้อสอบอัตนัยประยุกต์ในลักษณะนี้ทำให้ผู้สอบใช้เวลาในการแก้ปัญหาผู้ป่วยแต่ละรายไม่มากนัก และสามารถประเมินทักษะการแก้ปัญหาได้ในหลากหลายสถานการณ์ คะแนนสอบที่ได้จึงมีความเที่ยงสูง มีรายงานค่าความเที่ยงของคะแนนสอบถึง 0.8 ในการสอบอัตนัยประยุกต์แบบแก้ปัญหาสำคัญเป็นเวลาสี่ชั่วโมง[14]

ตัวอย่างข้อสอบอัตนัยประยุกต์แบบแก้ปัญหาสำคัญ

ตอนที่ 1 ชาย 36 ปี น้ำหนักตัว 55 กิโลกรัม ท้องร่วงถ่ายเป็นน้ำ 20 ครั้งในเวลา 1 วัน ตรวจร่างกายพบ อุณหภูมิ 36.9 องศาเซลเซียส ชีพจร 112 ครั้งต่อนาที ตรวจความดันโลหิตท่านอน 104/56 มิลลิเมตรปรอท ความดันโลหิตท่านั่ง 90/50 มิลลิเมตรปรอท

คำถามที่ 1.1 ให้ผู้สอบเขียนปัญหาสำคัญที่สุดของผู้ป่วยรายนี้ 1 อย่าง

ตอนที่ 2 ผู้ป่วยได้รับการประเมินว่ามีภาวะขาดสารน้ำปานกลางถึงรุนแรง ท่านต้องการให้สารน้ำทางหลอดเลือดดำแก่ผู้ป่วย

คำถามที่ 2.1 จงเขียนคำสั่งการรักษาเพื่อให้สารน้ำที่เหมาะสมแก่ผู้ป่วย

คำถามที่ 2.2 จงส่งตรวจเพิ่มเติมทางห้องปฏิบัติการเพื่อช่วยวินิจฉัยผู้ป่วยรายนี้ 2 การตรวจ

จากตัวอย่างข้างต้นจะเห็นว่าผู้ออกข้อสอบไม่ได้เริ่มจากการถามว่าจะซักประวัติ หรือตรวจร่างกายอะไรในผู้ป่วยที่มีภาวะท้องร่วงรุนแรง เนื่องจากผู้ออกข้อสอบเห็นว่าปัญหาสำคัญในการดูแลผู้ป่วยในภาวะนี้เป็นเรื่องการประเมินความรุนแรงของการขาดสารน้ำและการให้น้ำเกลือทดแทนในปริมาณที่เหมาะสมร่วมกับการสืบค้นหาสาเหตุของท้องร่วง ดังนั้นโจทย์ข้อนี้จึงมีเพียงสองตอนและใช้เวลาสอบไม่เกินสิบนาที

ขั้นตอนการสร้างข้อสอบอัตนัยประยุกต์

การสร้างข้อสอบอัตนัยประยุกต์ที่มีคุณภาพดีควรมีการดำเนินการเป็นขั้นตอน ดังนี้[4,20]

1. ตั้งกลุ่มพัฒนาข้อสอบ

ข้อสอบอัตนัยประยุกต์ที่ดีควรเป็นการแก้ปัญหาที่อาศัยความรู้จากหลากหลายวิชา การมีทีมคณาจารย์ที่มีประสบการณ์และความชำนาญแตกต่างกันมาช่วยกันสร้างข้อสอบจะได้สถานการณ์ผู้ป่วยที่เหมือนจริงในเวชปฏิบัติและสามารถประเมินความรู้ของผู้เข้าสอบได้ครอบคลุมสหสาขาวิชา และมั่นใจได้ว่าการเฉลยคำตอบทำได้อย่างรอบคอบ

2. เลือกปัญหาทางคลินิกที่จะทำการประเมินผู้สอบ

ขั้นตอนนี้เป็นขั้นตอนที่สำคัญมาก เนื่องจากโดยลักษณะข้อสอบอัตนัยประยุกต์จะทำให้ทำการสอบได้จำนวนข้อไม่มากนัก จึงเป็นไปไม่ได้ที่จะทำให้สถานการณ์ที่เป็นปัญหาทางคลินิกทุกอย่างจะมาปรากฏอยู่ในชุดข้อสอบ ดังนั้นการเลือกปัญหาทางคลินิกที่จะทำการสอบจึงต้องทำอย่างเป็นระบบ ควรมีการจัดทำตารางกำหนดลักษณะข้อสอบที่ชัดเจนว่าในการสอบครั้งหนึ่ง ๆ จะมีข้อสอบกี่ข้อ จะประเมินความรู้ในระบบอวัยวะใด และจัดสรรให้ข้อสอบไม่ซ้ำซ้อนกัน (ไม่ควรมีข้อสอบสองข้อถามความรู้ในระบบอวัยวะเดียวกัน ในขณะที่บางระบบอวัยวะไม่มีข้อสอบเลย)

ลักษณะปัญหาทางคลินิกที่ควรเลือกมาสอบด้วยข้อสอบอัตนัยประยุกต์ ได้แก่

- ปัญหาที่พบได้บ่อยในเวชปฏิบัติ
- ปัญหาที่แพทย์เกิดความผิดพลาดในการดูแลผู้ป่วยค่อนข้างบ่อย
- ปัญหาที่ยังไม่สามารถวินิจฉัยสาเหตุได้ชัดเจน
- ปัญหาที่มีความเกี่ยวพันกับหลายระบบ

เมื่อทีมคณาจารย์กำหนดปัญหาทางคลินิกที่จะทำการประเมินได้ชัดเจนแล้ว (เช่น ปัญหาตัวเหลือง, น้ำหนักลด เป็นต้น) สิ่งที่ต้องดำเนินการต่อคือการสร้างสถานการณ์ผู้ป่วยที่แสดงถึงปัญหาดังกล่าวขึ้น โดยกำหนดรายละเอียดต่าง ๆ ให้ผู้เข้าสอบอ่านแล้วนึกภาพผู้ป่วยได้ ในสถานการณ์ควรมีรายละเอียดเกี่ยวกับอายุ เพศ อาการสำคัญ บริบทของการดูแลผู้ป่วย (เช่น ห้องฉุกเฉินของโรงพยาบาลชุมชน หรือ หอผู้ป่วยในโรงพยาบาลมหาวิทยาลัย เป็นต้น)

3. กำหนดปัญหาสำคัญ

เมื่อทีมคณาจารย์เลือกปัญหาทางคลินิกที่จะทำการสอบแล้ว คณาจารย์ต้องตั้งคำถามว่าขั้นตอนใดในการดูแลผู้ป่วยที่มีปัญหาดังกล่าวจัดเป็นขั้นตอนสำคัญที่สุดในการจัดการปัญหานั้น ซึ่งขั้นตอนดังกล่าวจะได้รับการกำหนดให้เป็น ปัญหาสำคัญของสถานการณ์ผู้ป่วยที่จะใช้สอบ ในบางกรณีที่ทีมคณาจารย์ไม่สามารถเลือกขั้นตอนสำคัญในปัญหาทางคลินิกนั้น ๆ จากวิธีดังกล่าวได้

อาจใช้คำถามว่าขั้นตอนใดในการดูแลผู้ป่วยที่มีปัญหาดังกล่าวเป็นขั้นตอนที่นักศึกษาแพทย์หรือแพทย์ประจำบ้านทำผิดพลาดมากที่สุด[4]

มีข้อแนะนำสองประการสำหรับการกำหนดปัญหาสำคัญในแต่ละสถานการณ์ ได้แก่

- สิ่งที่ต้องตัดสินใจในผู้ป่วยแม้เป็นสิ่งที่ถูกต้องและควรปฏิบัติอาจไม่ได้เป็นขั้นตอนสำคัญที่จะต้องนำมาสอบเสมอไป การปฏิบัติต่อผู้ป่วยหลายอย่างที่ทำกันเป็นปกติ โดยไม่ต้องคิดวิเคราะห์ เป็นขั้นตอนที่ไม่ค่อยทำผิดพลาด มักไม่ใช่ปัญหาสำคัญในสถานการณ์นั้น

- ปัญหาสำคัญไม่จำกัดอยู่เฉพาะประเด็นปัญหาทาง ชีววิทยาการแพทย์ (biomedical) เท่านั้น ในบางสถานการณ์ปัญหาสำคัญอาจเป็นประเด็นทางจริยธรรม กฎหมาย หรือ การส่งเสริมสุขภาพและป้องกันโรคก็ได้

4. เขียนโจทย์คำถาม

เมื่อมีสถานการณ์ผู้ป่วยและขั้นตอนที่เป็นปัญหาสำคัญในสถานการณ์นั้นแล้ว ทีมคณาจารย์ต้องเขียนโจทย์คำถามที่มีความชัดเจน เพื่อประเมินว่าผู้เข้าสอบมีความสามารถในการตัดสินใจในการแก้ปัญหาสำคัญในสถานการณ์ดังกล่าวหรือไม่ โดยทั่วไปแล้วลักษณะโจทย์คำถามที่ใช้บ่อยในข้อสอบอัตนัยประยุกต์ได้แก่

- จงสอบถามประวัติที่สำคัญเพิ่มเติม
- จงบอกการตรวจร่างกายที่สำคัญที่ต้องมองหา (หรือตรวจเพิ่มเติม) ในผู้ป่วย
- จงให้การวินิจฉัย (หรือ การวินิจฉัยแยกโรค)
- จงสั่งการตรวจค้นเพิ่มเติมเพื่อให้การวินิจฉัยโรค
- จงสั่งการรักษาที่เหมาะสมให้ผู้ป่วย

โดยทั่วไปแล้วสถานการณ์ผู้ป่วยหนึ่ง ๆ ควรมีคำถามราว 2 – 3 ข้อ แต่ละข้อประเมินความสามารถในการจัดการกับปัญหาสำคัญ 1 ประเด็น[4,21] ในการเขียนโจทย์คำถามแต่ละข้อนั้นแนะนำให้มีการกำหนดจำนวนคำตอบที่สามารถตอบได้ไว้ด้วย เช่น

- จงบอกชื่อโรคที่ผู้ป่วยรายนี้น่าจะเป็นมากที่สุด 1 โรค
- จงบอกผลการตรวจร่างกายที่สำคัญที่จะช่วยยืนยันการวินิจฉัยโรคมา 3 ประการ
- จงระบุการตรวจเพิ่มเติมทางห้องปฏิบัติการที่จะช่วยในการวินิจฉัยโรค 1 การตรวจ

การกำหนดจำนวนคำตอบนี้จะทำให้ผู้เข้าสอบต้องเลือกสิ่งที่ถูกต้องเหมาะสมที่สุดเท่านั้นมาเขียนตอบ หากผู้เข้าสอบเขียนคำตอบเกินจำนวนที่กำหนด อาจารย์ผู้ตรวจข้อสอบจะไม่อ่านคำตอบที่เกินมา การปฏิบัติเช่นนี้จะช่วยกำจัดปัญหาการตรวจกระดาษคำตอบที่ผู้เข้าสอบเขียนคำตอบแบบหว่านแห ให้ครอบคลุมทุกอย่างโดยที่ผู้เข้าสอบเองไม่มีความรู้ ความเข้าใจว่าสิ่งใดเป็นประเด็นสำคัญในการดูแลผู้ป่วยในขั้นตอนนั้น ๆ

เมื่อทำการเขียนโจทย์คำถามและจำนวนคำตอบที่ต้องการแล้ว ให้อาจารย์ระบุเวลาที่ใช้ในการตอบคำถามตอนนั้นด้วย เนื่องจากข้อสอบอัตนัยประยุกต์มีการดำเนินของสถานการณ์ผู้ป่วยที่กำหนดให้โดยมีการให้ข้อมูลทีละส่วน ผู้เข้าสอบจำเป็นที่จะต้องรู้เวลาที่มีในการทำข้อสอบแต่ละตอนก่อนที่จะต้องส่งคำตอบและสถานการณ์ผู้ป่วยดำเนินต่อไป ในการกำหนดเวลาในการทำข้อสอบแต่ละตอนให้อาจารย์ผู้ออกข้อสอบพิจารณาจากทั้งเวลาที่ต้องใช้ในการอ่าน และเวลาที่ต้องใช้ในการเขียนคำตอบในข้อสอบตอนที่ต้องอ่านเนื้อหาโจทย์มาก หรือต้องเขียนคำตอบหลายบรรทัด ควรต้องมีการให้เวลาในการทำข้อสอบมากพอ หากเป็นไปได้ควรได้มีการลองทำการอ่านโจทย์และเขียนคำตอบโดยตัวอาจารย์ผู้ออกข้อสอบเองหรือเพื่อนอาจารย์แล้วลองจับเวลาที่อาจารย์ใช้ในการทำข้อสอบตอนนั้น ๆ เวลาที่ได้จะเป็นเวลาที่ผู้เชี่ยวชาญใช้แก้ปัญหาผู้ป่วยในสถานการณ์ดังกล่าว หากให้นักศึกษาทำ ควรเพิ่มเวลาให้ร้อยละ 30 – 50 ของเวลาที่อาจารย์ใช้

5. กำหนดเกณฑ์การให้คะแนน

ขั้นตอนสุดท้ายในการสร้างข้อสอบอัตนัยประยุกต์คือการกำหนดเกณฑ์การให้คะแนน ซึ่งเป็นขั้นตอนที่มีความท้าทาย และสร้างความลำบากใจให้แก่อาจารย์ผู้ออกข้อสอบหลายท่าน เนื่องด้วยเกรงว่าจะเฉลยคำตอบไม่ครอบคลุมสิ่งที่ผู้เข้าสอบจะเขียนตอบมา หรือเกิดความไม่เป็นธรรมขึ้น ในที่นี้ผู้นิพนธ์ขอเสนอแนะแนวทางในการกำหนดเกณฑ์ให้คะแนนดังนี้

- แนะนำให้กำหนดคะแนนเต็มในการแก้ปัญหา

สถานการณ์หนึ่ง ๆ เป็น 100 คะแนน เท่ากันในทุกสถานการณ์ เพื่อให้ไม่ต้องทำการปรับคะแนนสอบหลังการตรวจข้อสอบ

- กรณีที่มีคำตอบที่ถูกต้องยอมรับได้เพียงคำตอบเดียว เช่นข้อมูลจากโจทย์มีความชัดเจนว่าผู้ป่วยเป็นโรคอะไร แล้วโจทย์ให้ผู้เข้าสอบตอบชื่อโรค หากผู้เข้าสอบตอบตรงตามเฉลยที่ตั้งไว้ให้ได้คะแนนเต็ม หากตอบคำตอบอื่นนอกจากนั้นไม่ได้คะแนน

- ในกรณีที่มีคำตอบที่เป็นไปได้หลายคำตอบ เช่น ถามการวินิจฉัยแยกโรค 3 โรค ในกรณีนี้ผู้ออกข้อสอบควรเตรียมเฉลยไว้หลายคำตอบ (มากกว่าที่กำหนดให้ตอบ) โดยแต่ละคำตอบสามารถมีน้ำหนักคะแนนไม่เท่ากันได้ โดยคำตอบที่ถูกต้องมาก สอดคล้องกับสิ่งที่ควรคิดถึงหรือปฏิบัติในขั้นตอนดังกล่าว จะได้คะแนนสูง ในขณะที่สิ่งที่สามารถเป็นไปได้หรือควรปฏิบัติน้อยกว่าจะได้คะแนนลดลงไป แต่เมื่อรวมคะแนนจากทุกคำตอบที่ผู้เข้าสอบตอบมาแล้วคะแนนสูงสุดที่ผู้เข้าสอบจะได้ต้องไม่สูงเกินคะแนนที่กำหนดไว้เป็นคะแนนเต็มของข้อสอบตอนนั้น

- คำตอบบางลักษณะมีการเขียนเนื้อหาที่มีความครบถ้วนสมบูรณ์แตกต่างกันได้ การกำหนดเกณฑ์สามารถกำหนดให้คำตอบที่มีความสมบูรณ์ได้คะแนนเต็ม ส่วนคำตอบที่ไม่สมบูรณ์จะได้คะแนนลดหลั่นลงไปตามความเหมาะสม (เช่น โจทย์ถามเรื่องการให้สารน้ำทางหลอดเลือดดำ คำตอบ Normal saline solution 1000 ml IV drip 200 ml/hr จะได้คะแนนเต็ม 4 คะแนน แต่หากเขียนตอบ Normal saline solution โดยไม่บอกอัตราเร็วของการให้ ได้เพียง 2 คะแนน หากบอกอัตราการให้ถูกต้องให้ 2 คะแนน)

- คำตอบที่ไม่ถูกต้อง ไม่สมควรปฏิบัติแก่ผู้ป่วยโดยทั่วไปแล้วพิจารณาไม่ให้คะแนน ซึ่งก็จัดเป็นการทำโทษในระดับหนึ่งแล้ว เพราะผู้สอบมีสิทธิเขียนคำตอบได้จำนวนจำกัด การที่ไม่ให้คะแนนในคำตอบที่ไม่เหมาะสม ก็จะทำให้คะแนนสูงสุดที่ผู้สอบจะทำได้ลดลงไปแล้ว การปฏิบัติที่ไม่ถูกต้องที่มีผลเสียรุนแรงต่อผู้ป่วยเท่านั้นที่ควรจะพิจารณาให้คะแนนติดลบ และแม้มีการให้คะแนนติดลบก็ไม่ควรมีการติดลบข้ามไปถึงข้อสอบข้ออื่นในชุดข้อสอบนั้น

- การกำหนดเกณฑ์การให้คะแนน ไม่ควรใช้อาจารย์ท่านเดียวในการกำหนด เพราะมักได้คำตอบที่ไม่ครอบคลุม ควรใช้ทีมคณาจารย์หลายท่านช่วยกันคิดว่าคำตอบที่ผู้เข้าสอบอาจจะตอบได้ในสถานการณ์ดังกล่าว ซึ่งจะได้เกณฑ์การให้คะแนนที่สมบูรณ์กว่า อย่างไรก็ตาม ถึงแม้ว่าจะใช้คณาจารย์หลายท่านช่วยกันคิดคำตอบแล้วก็ตาม จะพบว่าในการตรวจข้อสอบอัตนัยประยุกต์หลายครั้ง จะพบคำตอบที่ผู้เข้าสอบตอบมาที่น่าจะได้คะแนนแต่อาจารย์ผู้ออกข้อสอบไม่ได้กำหนดเกณฑ์คะแนนไว้ล่วงหน้าอยู่ประปราย ดังนั้นในการนำข้อสอบอัตนัยประยุกต์ที่สร้างขึ้นใหม่มาใช้ในการสอบ 2 – 3 รอบแรกแนะนำให้อาจารย์ผู้ที่ออกข้อสอบและมีความเชี่ยวชาญชำนาญในการดูแลผู้ป่วยในสถานการณ์นั้น ๆ เป็นผู้ทำการตรวจข้อสอบ เพื่อให้สามารถพิจารณาได้ว่าคำตอบใดที่น่าจะเพิ่มเข้าไปในเกณฑ์การให้คะแนนด้วย ซึ่งเมื่อทำไป 2 – 3 รอบการสอบแล้วมักจะได้เกณฑ์การให้คะแนนที่มีความครอบคลุมคำตอบที่ผู้สอบจะตอบมาได้ทั้งหมด แล้วจึงมอบหมายให้อาจารย์ท่านอื่นช่วยตรวจให้คะแนนข้อสอบต่อไป

เมื่อทำการกำหนดเกณฑ์การให้คะแนนในข้อสอบเสร็จทุกข้อย่อยแล้วกระบวนการขั้นตอนสุดท้ายในการสร้างข้อสอบอัตนัยประยุกต์คือการกำหนดเกณฑ์ผ่านของโจทย์สถานการณ์นั้น กล่าวคือจากคะแนนเต็ม 100 คะแนน ผู้สอบต้องทำคะแนนได้อย่างน้อยที่สุดกี่คะแนนจึงจะจัดว่าสอบผ่านในการแก้ปัญหาสถานการณ์นั้น ๆ วิธีการตั้งเกณฑ์ผ่านทำได้หลายวิธี แต่วิธีที่เป็นที่นิยมมากที่สุดสำหรับข้อสอบอัตนัยประยุกต์ และเป็นวิธีที่คณะแพทยศาสตร์ศิริราชพยาบาลใช้เป็นประจำในการตัดสินผลสอบอัตนัยประยุกต์คือวิธี Modified Angoff ซึ่งมีขั้นตอนที่สำคัญสามขั้นตอนคือ

(1) กำหนดลักษณะของผู้ที่มีความรู้ ความสามารถคาบเส้น (borderline examinee) ว่าในความเห็นของคณาจารย์แล้วผู้ที่มีความรู้เทียบเท่าระดับต่ำสุดของเกณฑ์มาตรฐานการทำงานในการแก้ปัญหาเรื่องนั้น ๆ น่าจะทำอะไรได้ ทำอะไรไม่ได้

(2) ไล่ดูโจทย์คำถามที่ละข้อพร้อมเฉลย แล้วทำสัญลักษณ์ * ไว้ในคำตอบที่คาดว่าผู้ที่มีความรู้ ความสามารถคาบเส้นจะตอบในข้อสอบแต่ละตอน

(3) ทำการรวมค่าคะแนนที่ได้รับการทำสัญลักษณ์ * ไว้ตั้งแต่ข้อแรกจนถึงข้อสุดท้าย จะได้คะแนนเกณฑ์ผ่านในการแก้ปัญหาสถานการณ์นั้น ๆ[22]

แนวทางการพัฒนาข้อสอบอัตนัยประยุกต์ในคณะแพทยศาสตร์ศิริราชพยาบาล

คณะแพทยศาสตร์ศิริราชพยาบาลมีการใช้ข้อสอบอัตนัยประยุกต์ในการประเมินความรู้ของนักศึกษาแพทย์ชั้นคลินิกมานานแล้ว โดยเริ่มต้นจากการสอบของแต่ละภาควิชา และต่อมาเมื่อศูนย์ประเมินและรับรองความรู้ความสามารถในการประกอบวิชาชีพเวชกรรมกำหนดให้การสอบอัตนัยประยุกต์เป็นส่วนหนึ่งของการประเมินขั้นตอนที่ 3 ในการขอใบประกอบวิชาชีพเวชกรรมตั้งแต่ปีการศึกษา 2550 ทางคณะแพทยศาสตร์ศิริราชพยาบาลก็ได้มีการจัดสอบประมวลความรอบรู้ทางการแพทย์สหสาขาวิชา ด้วยข้อสอบอัตนัยประยุกต์ (comprehensive MEQ examination) ในนักศึกษาแพทย์ปีที่ 6 อย่างต่อเนื่อง ตลอดช่วงเวลาที่มีการใช้ข้อสอบอัตนัยประยุกต์ในคณะฯได้มีการพัฒนาข้อสอบประเภทนี้อย่างต่อเนื่อง จากเดิมเคยจัดสอบข้อสอบอัตนัยประยุกต์ในรูปแบบข้อสอบกระดาษ จนพัฒนาให้จัดสอบอัตนัยประยุกต์ด้วยการนำเสนอข้อมูลผู้ป่วยบนจอภาพคอมพิวเตอร์ ร่วมกับการเขียนคำตอบในกระดาษคำตอบตั้งแต่ปีการศึกษา 2552 จนถึงปัจจุบัน แต่ถึงแม้ว่าฝ่ายการศึกษาจะมีการพัฒนาระบบจัดสอบข้อสอบอัตนัยประยุกต์ให้มีประสิทธิภาพมากขึ้น อำนวยความสะดวกให้ผู้เข้าสอบมากขึ้น และเพิ่มความพึงพอใจในประสบการณ์การสอบขึ้นอย่างต่อเนื่อง จากการเก็บรวบรวมข้อมูลการวิเคราะห์ข้อสอบ วิเคราะห์คะแนน และแบบสำรวจความพึงพอใจของผู้สอบที่ผ่านมาผู้นิพนธ์มีความเห็นว่าการจัดสอบประมวลความรอบรู้ทางการแพทย์ด้วยข้อสอบอัตนัยประยุกต์ของนักศึกษาแพทย์ยังสามารถพัฒนาให้มีคุณภาพดีขึ้นได้อีกในหลายด้าน ดังนี้

(1) **เนื้อหาข้อสอบ**

ข้อสอบอัตนัยประยุกต์ที่ใช้ในการสอบประมวลความรอบรู้ทางการแพทย์ของคณะแพทยศาสตร์ศิริราชพยาบาลที่ผ่านมาหลายข้อเป็นเนื้อหาวิชาที่ยากและเป็นความรู้ลึกในระดับผู้เชี่ยวชาญเฉพาะทาง แนวทางในการพัฒนาการสอบอัตนัยประยุกต์อันดับแรกคือการพัฒนาเนื้อหาให้เหมาะสมกับการประเมินความรู้ของแพทย์เวชปฏิบัติทั่วไป

เนื้อหาข้อสอบอัตนัยประยุกต์สำหรับการสอบประมวลความรอบรู้นั้นควรมุ่งเน้นเนื้อหาที่เป็นสหสาขาวิชา กล่าวคือต้องอาศัยองค์ความรู้ที่นักศึกษาได้ศึกษามาจากหลายภาควิชามาช่วยกันแก้ปัญหาผู้ป่วย ข้อสอบอัตนัยประยุกต์ที่นำมาสอบนักศึกษาแพทย์ทุกข้อในปัจจุบันล้วนมีความเป็นสหสาขาวิชาทั้งสิ้น มีอาจารย์จากหลากหลายภาควิชามาร่วมกันออกข้อสอบ แต่อย่างไรก็ตามข้อสอบบางข้ออาจมีลักษณะการใช้ความรู้สหสาขาวิชาแบบแยกเป็นส่วน ๆ กล่าวคืออาจารย์ต่างภาควิชากันใช้การแบ่งงานออกเป็นส่วน ๆ อาจารย์ภาควิชาที่หนึ่งออกข้อสอบในตอนที่หนึ่งกับสอง อาจารย์ภาควิชาที่สองออกข้อสอบในตอนที่สามกับสี่ และอาจารย์ภาควิชาที่สามออกข้อสอบในตอนที่ห้ากับหก ข้อสอบลักษณะนี้มักจะยากมาก เนื่องจากเป็นการใช้ความรู้เชิงลึกของแต่ละภาควิชาทีละเรื่อง เช่นซักประวัติ ตรวจร่างกายแล้วก็ไม่สามารถวินิจฉัยโรคได้ ต้องส่งต่อไปทำการตรวจเพิ่มเติมในอีกภาควิชาหนึ่ง ซึ่งผลการตรวจเพิ่มเติมก็แปลผลได้ยาก เมื่อได้ข้อสรุปแล้วก็ต้องส่งต่อไปให้แพทย์อีกสาขาวิชาหนึ่งทำการรักษา เมื่อรักษาแล้วก็มีภาวะแทรกซ้อนต้องส่งต่อให้แพทย์อีกสาขาวิชาหนึ่งทำการแก้ไขภาวะแทรกซ้อนให้เป็นต้น โดยทั่วไปแล้วข้อสอบอัตนัยประยุกต์ที่ใช้ความรู้สหสาขาวิชาที่เป็นที่ต้องการในการสอบประมวลความรอบรู้นั้นไม่ควรเป็นการประเมินความรู้ในเชิงลึกทีละวิชาในข้อสอบแต่ละตอน แต่ควรเป็นการผสมผสานความรู้จากหลากหลายสาขาวิชาในทุกขั้นตอน เช่น หญิงอายุ 30 ปี ปวดท้องน้อยตื้อ ๆ ตลอดเวลา 6 ชั่วโมง มีไข้ต่ำ ๆ คลื่นไส้เล็กน้อย โจทย์ให้ผู้สอบซักประวัติเพื่อการวินิจฉัยโรคซึ่งผู้สอบที่จะตอบคำถามได้ดีต้องอาศัยความรู้ทั้งโรคในระบบทางเดินอาหาร ทางเดินปัสสาวะ อวัยวะสืบพันธุ์สตรี กระดูกและกล้ามเนื้อ เป็นต้น

ข้อแนะนำในเรื่องเนื้อหาที่สำคัญคืออาจารย์ผู้ออกข้อสอบต้องตระหนักว่าการสอบนี้เป็นการประเมินความรู้เวชปฏิบัติทั่วไป มิใช่การประเมินความรู้เชิงลึกในศาสตร์ของแต่ละสาขาวิชา โรคหรือภาวะที่นำมาออก

ข้อสอบส่วนใหญ่ควรอยู่ในเกณฑ์มาตรฐานผู้ประกอบวิชาชีพเวชกรรมในกลุ่มที่ 1 หรือ 2 (โรคหรือภาวะที่แพทย์เวชปฏิบัติทั่วไปสามารถให้การดูแลด้วยตนเองได้ และพิจารณาส่งต่อในกรณีที่โรครุนแรงหรือซับซ้อน) โรคหรือภาวะที่อยู่ในเกณฑ์มาตรฐานฯ กลุ่มที่ 3 (โรคหรือภาวะที่แพทย์เวชปฏิบัติทำการดูแลเบื้องต้นแล้วให้ส่งต่อไปยังผู้เชี่ยวชาญ) ควรนำมาออกข้อสอบไม่มากนัก หากจะนำโรคหรือภาวะในเกณฑ์มาตรฐานฯกลุ่มที่ 3 มาออกสอบ ต้องมุ่งเน้นการดูแลรักษาเบื้องต้นที่แพทย์เวชปฏิบัติทั่วไปพึงทำได้ ไม่ควรมุ่งประเด็นไปที่การรักษาโดยผู้เชี่ยวชาญเฉพาะสาขามากจนเกินไป

(2) **รูปแบบคำถาม**

หลักการสำคัญของการวัดและประเมินผลคือการเลือกใช้เครื่องมือที่เหมาะสมในการวัดผลการเรียนรู้ ข้อสอบอัตนัยประยุกต์ได้รับการพัฒนาขึ้นเพื่อประเมินทักษะในการตัดสินใจทางคลินิกเป็นสำคัญ สิ่งที่ยังเป็นปัญหาในข้อสอบอัตนัยประยุกต์บางข้อคือการเลือกถามคำถามในรูปแบบที่ไม่ตรงตามเป้าประสงค์ของการสอบอัตนัยประยุกต์ เช่นถามความจำขั้นพื้นฐาน โดยไม่ต้องคิดวิเคราะห์และตัดสินใจว่าจะทำหรือไม่ทำสิ่งใดกับผู้ป่วย รูปแบบคำถามที่ไม่เหมาะสมเหล่านี้เช่น ผู้ชายอายุ 40 ปี มีไข้สองเดือน จงถามประวัติ การใช้รูปแบบคำถามลักษณะนี้จะวัดเพียงว่าผู้เข้าสอบจดจำหัวข้อทั้งหมดของการซักประวัติในผู้ป่วยที่มีไข้เรื้อรังได้หรือไม่ และผู้สอบคนใดเขียนได้เร็วและครบถ้วนกว่ากัน ซึ่งอาจารย์สามารถใช้เครื่องมือประเมินผลชนิดอื่นในการวัดความจำขั้นพื้นฐานได้ดีกว่าการใช้ข้อสอบอัตนัยประยุกต์ การใช้ข้อสอบอัตนัยประยุกต์ควรมุ่งเน้นคำถามประเมินความสามารถในการวิเคราะห์ปัญหาผู้ป่วย และตัดสินใจสั่งการตรวจ หรือรักษาผู้ป่วยอย่างเหมาะสม

(3) **จำนวนสถานการณ์ผู้ป่วยที่ใช้สอบ**

ในการสอบประมวลความรอบรู้ด้วยข้อสอบอัตนัยประยุกต์ของคณะแพทยศาสตร์ศิริราชพยาบาลที่ผ่านมามีการใช้สถานการณ์ผู้ป่วยในข้อสอบตั้งแต่ 5 ถึง 8 ราย ถึงแม้ว่าจำนวนสถานการณ์ในการสอบระยะหลังมีแนวโน้มเพิ่มขึ้น แต่หากพิจารณาในแง่ของความจำเพาะต่อบริบทของผู้ป่วย (case specificity) ที่ได้อภิปรายไปก่อนหน้านี้แล้วจะเห็นได้ว่าการที่ผู้สอบแก้ปัญหาผู้ป่วยได้ 5 ถึง 8 รายนี้น่าจะยังครอบคลุมประเด็นปัญหาทางคลินิกได้ไม่มากเพียงพอ และคะแนนสอบที่ได้มาน่าจะพัฒนาให้มีความเที่ยงสูงขึ้นได้อีกหากในการสอบมีจำนวนสถานการณ์มากขึ้น เนื่องด้วยรูปแบบข้อสอบอัตนัยประยุกต์ที่ใช้ในการสอบของคณะฯยังเน้นการสอบถามการจัดการปัญหาของผู้ป่วยตลอดตั้งแต่ต้นจนจบ (Patient management problem, PMP) จึงทำให้เวลาที่ใช้ในการสอบในแต่ละสถานการณ์ค่อนข้างนาน (แต่ละสถานการณ์มีคำถามย่อย 4 – 8 ข้อ ใช้เวลา 15 ถึง 30 นาทีต่อสถานการณ์) จึงทำให้ไม่สามารถสอบได้หลายสถานการณ์

หากพิจารณาจากข้อแนะนำของผู้เชี่ยวชาญในการประเมินผลที่ได้อภิปรายไปก่อนหน้านี้ที่แนะนำให้ใช้ข้อสอบอัตนัยประยุกต์แบบแก้ปัญหาสำคัญ แนวทางการพัฒนาข้อสอบอัตนัยประยุกต์ของคณะฯ ให้มีความครอบคลุมสถานการณ์ผู้ป่วยที่มากขึ้น และ มีความเที่ยงของคะแนนสอบมากขึ้นคือการใช้ข้อสอบแบบแก้ปัญหาสำคัญมาแทนการจัดการปัญหาของผู้ป่วยตั้งแต่ต้นจนจบ กล่าวคือในแต่ละสถานการณ์ผู้ป่วย ข้อสอบควรมุ่งถามคำถามสำคัญเพียงสองหรือสามข้อ และเพิ่มจำนวนสถานการณ์ผู้ป่วยให้มากขึ้นนั่นเอง

(4) **การนำเสนอข้อสอบ**

การทำข้อสอบอัตนัยประยุกต์ ผู้สอบต้องทำงานภายใต้ข้อจำกัดด้านเวลา เวลาที่ใช้ในการตอบข้อสอบอัตนัยประยุกต์เป็นผลรวมของเวลาที่ใช้อ่านโจทย์ คิดวิเคราะห์ และเขียนคำตอบ ปัญหาสำคัญประการหนึ่งที่สร้างความลำบากให้กับผู้สอบคือปริมาณข้อมูลที่นำเสนอให้ผู้สอบอ่านในสถานการณ์ผู้ป่วยแต่ละรายนั้นมีมาก ทำให้ผู้สอบต้องใช้เวลาในการอ่านมากและเหลือเวลาสำหรับเขียนคำตอบน้อย ถึงแม้ว่าในการนำเสนอข้อมูลของข้อสอบอัตนัยประยุกต์จะได้มีการแยกข้อมูลเดิมที่เคยนำเสนอไปก่อนหน้านี้ ออกจากข้อมูลใหม่ที่เพิ่มเติมขึ้นมาในการนำเสนอข้อสอบแต่ละตอนแล้วก็ตาม ด้วย

รายละเอียดที่นำเสนอมีมาก ผู้สอบก็ยังคงมีความจำเป็นต้องประมวลผลข้อมูลปริมาณมากอยู่ดี จากการทบทวนเนื้อหาของข้อสอบอัตนัยประยุกต์ที่ได้จัดสอบไปหลายครั้งพบว่าข้อสอบหลายข้อใช้ข้อมูลเพียงส่วนน้อยของที่นำเสนอเท่านั้นก็สามารถนำไปสู่การแก้ปัญหาและการตัดสินใจเลือกการส่งตรวจหรือให้การรักษาผู้ป่วยได้อย่างถูกต้อง ดังนั้นแนวทางในการพัฒนาคุณภาพของการสอบอัตนัยประยุกต์อีกทางหนึ่งคือการที่อาจารย์ผู้ออกข้อสอบพึงตระหนักถึงข้อจำกัดเรื่องเวลาในการทำข้อสอบของนักศึกษาและเขียนสถานการณ์ผู้ป่วยให้มีความกระชับ นำเสนอเฉพาะข้อมูลที่มีความจำเป็นในการตัดสินใจให้การดูแลรักษาผู้ป่วยเท่านั้น ในการนำเสนอข้อมูลแต่ละตอนควรต้องทบทวนว่าข้อมูลเก่าที่เคยให้ในขั้นตอนก่อนหน้านั้นมีความจำเป็นต้องนำเสนอซ้ำทั้งหมดหรือไม่ หากทำได้ควรทำการสรุปข้อมูลให้ผู้เข้าสอบ และตัดทอนข้อมูลที่ไม่จำเป็นในการแก้ปัญหาขั้นตอนนั้น ๆ ออกไป ตัวอย่างเช่น ในข้อสอบตอนที่หนึ่งมีการนำเสนอประวัติผู้ป่วยสั้น ๆ แล้วมีโจทย์ถามถึงประวัติที่จะซักเพิ่มเติม และการตรวจร่างกายที่จะทำเพื่อนำไปสู่การวินิจฉัยโรค ในข้อสอบตอนที่สองอาจารย์นำเสนอประวัติและผลการตรวจร่างกายเพิ่มเติมให้ แล้วมีโจทย์ถามถึงการวินิจฉัยโรค และการส่งตรวจทางห้องปฏิบัติการที่เหมาะสม ในข้อสอบตอนที่สามอาจารย์นำเสนอข้อมูลการวินิจฉัยโรคของผู้ป่วยพร้อมผลการตรวจทางห้องปฏิบัติการ แล้วถามแนวทางการรักษา การนำเสนอข้อสอบในลักษณะนี้ในข้อสอบหลายข้อมีการนำเสนอข้อมูลของโจทย์ซ้ำเดิมและค่อย ๆ เพิ่มข้อมูลขึ้นในทุกขั้นตอน ในข้อสอบตอนที่สองก็นำเสนอข้อมูลที่เสนอในตอนที่หนึ่งกับสอง ในข้อสอบตอนที่สามก็นำเสนอข้อมูลที่เสนอในตอนที่หนึ่ง สอง และ สาม ซึ่งเมื่อผ่านการสอบไปหลายตอนจะมีข้อมูลสะสมจำนวนมากที่ผู้สอบต้องอ่าน การนำเสนอข้อสอบที่มีประสิทธิภาพมากกว่าควรมีการสรุปข้อมูลอย่างเหมาะสม ในข้อสอบตอนที่สาม หากได้ข้อสรุปการวินิจฉัยโรคแล้ว จะถามแนวทางการรักษาโรค อาจารย์ควรพิจารณาตัดข้อมูลประวัติและการตรวจร่างกายออก หากการสั่งการรักษาจำเป็นต้องทราบข้อมูลจากประวัติ หรือการตรวจร่างกายบางอย่าง เช่น น้ำหนักตัว หรือ โรคร่วมที่ส่งผลต่อการวางแผนการรักษา ก็ให้นำเสนอเฉพาะข้อมูลที่ส่งผลต่อการตัดสินใจในขั้นตอนนั้นเท่านั้น

การนำเสนอข้อสอบอัตนัยประยุกต์ด้วยระบบคอมพิวเตอร์ก็เป็นอีกแนวทางหนึ่งที่คณะแพทยศาสตร์ศิริราชพยาบาลเห็นความสำคัญ และได้ดำเนินการพัฒนาอย่างต่อเนื่อง คณะแพทยศาสตร์ศิริราชพยาบาลมีความพร้อมในการพัฒนาด้านนี้มากพอสมควร เนื่องด้วยมีห้องคอมพิวเตอร์ที่มีจำนวนคอมพิวเตอร์มากพอที่จะจัดให้ผู้เข้าสอบทุกคนมีจอคอมพิวเตอร์ส่วนตัว มีการวางระบบเครือข่ายให้มีการส่งผ่านข้อมูลระหว่างเครื่องคอมพิวเตอร์ได้ดี และมีความเสถียรของระบบพอสมควร มีการวางมาตรการรักษาความปลอดภัยของข้อมูลในระบบที่ดี สามารถควบคุมการเข้าออกของข้อมูลจากระบบเครือข่ายคอมพิวเตอร์ได้ จึงส่งผลให้คณะได้ปรับรูปแบบการจัดสอบอัตนัยประยุกต์จากระบบสอบด้วยข้อสอบกระดาษมาเป็นการนำเสนอข้อสอบบนจอคอมพิวเตอร์ตั้งแต่ปีการศึกษา 2552 ซึ่งจากการสำรวจความเห็นของนักศึกษาผู้เข้าสอบได้รับการตอบรับดีมาก นักศึกษาพึงพอใจกับการสอบในระบบนี้ในระดับมากถึงมากที่สุด อย่างไรก็ตามระบบการสอบนี้ยังมีโอกาสที่จะพัฒนาให้ดีขึ้นได้อีก ในระบบการจัดสอบปัจจุบันของคณะฯ ยังคงเป็นรูปแบบที่ไม่ได้ใช้คอมพิวเตอร์อย่างเต็มรูปแบบ ยังคงให้ผู้สอบเขียนคำตอบลงในกระดาษคำตอบและเก็บกระดาษในตอนท้ายของการสอบในแต่ละสถานการณ์ผู้ป่วย การใช้ประโยชน์ของคอมพิวเตอร์ในการสอบปัจจุบันเน้นไปในการนำเสนอข้อมูลที่ทำให้ผู้สอบสามารถเห็นภาพถ่ายรังสี ภาพการตรวจทางห้องปฏิบัติการ แผนภาพ ตาราง รวมถึงรูปของผู้ป่วยได้ โดยผู้สอบทุกคนเห็นภาพที่มีความละเอียดสูงเท่าเทียมกัน และทำให้การบริหารการสอบทำได้มีประสิทธิภาพมากขึ้น ตัดปัญหาผู้สอบลักลอบเปิดดูข้อสอบในตอนต่อไปล่วงหน้า หรือทำข้อสอบในบางตอนเกินเวลา การแสดงเวลาที่เหลือในการทำข้อสอบแต่ละตอนบนหน้าจอทำให้ผู้สอบบริหารเวลาในการทำข้อสอบได้ดีขึ้น

ระบบจัดสอบอัตนัยประยุกต์ด้วยคอมพิวเตอร์อย่างเต็มรูปแบบที่ไม่ต้องมีการเขียนตอบในกระดาษเลยนั้นมีการจัดทำในต่างประเทศ[12,23] แต่ต้องยอมรับว่าการ

สร้างระบบการจัดสอบอัตนัยประยุกต์ด้วยคอมพิวเตอร์อย่างเต็มรูปแบบนั้นเป็นงานที่ซับซ้อนและมีความท้าทายหลายอย่าง ทั้งในด้านผู้จัดสอบ ระบบเครือข่ายคอมพิวเตอร์ และผู้เข้าสอบ ในอนาคตอันใกล้นี้ทางฝ่ายการศึกษายังไม่มีแนวทางที่จะพัฒนาการสอบอัตนัยประยุกต์เป็นระบบคอมพิวเตอร์อย่างเต็มรูปแบบ ด้วยข้อจำกัดสำคัญสามประการคือ ความพร้อมของผู้เข้าสอบ ความพร้อมของผู้ตรวจข้อสอบ และความพร้อมของระบบการสื่อสารระหว่างผู้ใช้กับคอมพิวเตอร์ กล่าวคือ ผู้เข้าสอบจำนวนไม่น้อยยังไม่คุ้นเคยกับการพิมพ์คำตอบที่มีทั้งภาษาไทยและภาษาอังกฤษผสมกันภายในเวลาที่จำกัด อาจารย์ผู้ตรวจข้อสอบจำนวนไม่น้อยยังไม่สะดวกที่จะทำการตรวจข้อสอบและกรอกคะแนนบนหน้าจอคอมพิวเตอร์ในสถานที่และเวลาที่กำหนด และการสร้างระบบการสื่อสารระหว่างคอมพิวเตอร์กับผู้ใช้ให้ทั้งนำเสนอข้อมูลผู้ป่วยที่มีรายละเอียดมาก พร้อมกับตอบรับคำตอบที่มีทั้งอักษร ตัวเลข และสัญลักษณ์พิเศษ ที่ผู้เข้าสอบจะพิมพ์เข้าเครื่องพร้อม ๆ กันหลายร้อยคนโดยมีการควบคุมเวลาอย่างรัดกุมด้วย ยังเป็นสิ่งที่ทำได้ยากในระบบเครือข่ายคอมพิวเตอร์ในปัจจุบัน ดังนั้นในอนาคตอันใกล้นี้ทิศทางการพัฒนาระบบการจัดสอบข้อสอบอัตนัยประยุกต์คงยังมุ่งเน้นไปในรูปแบบการนำเสนอข้อสอบผ่านจอภาพคอมพิวเตอร์ ร่วมกับการเขียนตอบในกระดาษคำตอบอยู่

แต่ถึงแม้ว่าจะคงการจัดสอบอัตนัยประยุกต์ในรูปแบบผสมผสานเช่นนี้ ผู้นิพนธ์ก็ยังเห็นว่ามีสิ่งที่ระบบการนำเสนอข้อมูลผ่านจอคอมพิวเตอร์สามารถทำให้ดีขึ้นได้ เช่นการทำให้ภาพมีรายละเอียดสูงขึ้น การเปิดโอกาสให้ผู้เข้าสอบสามารถขยายภาพเพื่อดูรายละเอียดในบางส่วน การปรับรูปแบบการนำเสนออักษร และพื้นหลังของจอภาพให้ผู้เข้าสอบอ่านข้อมูลได้ง่ายขึ้น เป็นต้น ซึ่งสิ่งเหล่านี้จะได้มีการศึกษาหาแนวทางในการพัฒนาในการสอบอัตนัยประยุกต์ครั้งต่อ ๆ ไป แต่อย่างไรก็ตามด้วยศักยภาพของระบบการจัดสอบในปัจจุบัน ผู้นิพนธ์ยังมีความเห็นว่าอาจารย์ผู้ออกข้อสอบก็ยังไม่ได้ใช้ศักยภาพของระบบอย่างเต็มที่ ยังมีข้อสอบหลายข้อที่ใช้การบรรยายสิ่งตรวจพบที่สามารถมองเห็นเป็นภาพได้แต่นำมาเขียนเป็นอักษรบรรยายสิ่งตรวจพบดังกล่าว ซึ่งทำให้ผู้เข้าสอบไม่ได้คิด วิเคราะห์และแปลผลการตรวจด้วยตนเอง แนวทางการพัฒนาข้อสอบอัตนัยประยุกต์ที่สมควรได้รับการส่งเสริมในระบบการจัดสอบปัจจุบันคือการใช้สื่อที่เป็นรูปภาพในข้อสอบให้มากขึ้น ไม่ว่าจะเป็นการตรวจร่างกายจากการดู การดูภาพรังสี การดูคลื่นไฟฟ้าหัวใจ การดูสิ่งส่งตรวจด้วยกล้องจุลทรรศน์ ล้วนแล้วแต่ควรนำเสนอเป็นรูปภาพทั้งสิ้น

บทสรุป

ในบทความนี้ผู้นิพนธ์ได้กล่าวถึงความรู้พื้นฐานในการสร้างข้อสอบอัตนัยประยุกต์ โดยได้สรุปลักษณะพื้นฐานของข้อสอบอัตนัยประยุกต์ พัฒนาการของข้อสอบประเภทนี้จากรูปแบบการจัดการปัญหาผู้ป่วยเป็นการแก้ปัญหาสำคัญ มีการสรุปขั้นตอนสำคัญในการสร้างข้อสอบอัตนัยประยุกต์ห้าขั้นตอน ได้แก่ (1) ตั้งกลุ่มพัฒนาข้อสอบ, (2) เลือกปัญหาทางคลินิก, (3) กำหนดปัญหาสำคัญ, (4) เขียนโจทย์คำถาม, และ (5) กำหนดเกณฑ์การให้คะแนน และในตอนท้ายได้มีการนำหลักการพัฒนาข้อสอบต่าง ๆ ที่กล่าวมาแล้วมาวิเคราะห์สถานการณ์การจัดสอบอัตนัยประยุกต์สำหรับนักศึกษาแพทย์คณะแพทยศาสตร์ศิริราชพยาบาลและเสนอแนะแนวทางในการพัฒนาคุณภาพการสอบอัตนัยประยุกต์สี่แนวทาง ได้แก่ (1) เนื้อหาข้อสอบ, (2) รูปแบบคำถาม, (3) จำนวนสถานการณ์ผู้ป่วย, และ (4) การนำเสนอข้อสอบ ผู้นิพนธ์เชื่อมั่นว่าหากการจัดสอบอัตนัยประยุกต์ได้รับการพัฒนาอย่างเหมาะสมจะนำไปสู่การประเมินความรู้ และทักษะการตัดสินใจดูแลผู้ป่วยในระดับคลินิกที่มีประสิทธิภาพ

เอกสารอ้างอิง

1. Downing SM. Assessment of knowledge with written test forms. In: Norman GR, van der Vleuten C, Newble DI, editors. International handbook of research in medial education. Dordrecht: Kluwer Academic Pubishers, 2002:647 - 72.
2. Epstein RM. Assessment in medical education. New Engl J Med 2007;356:387-96.
3. The Board of Censors of the Royal College of General Practitioners. The modified essay question. J Roy Coll Gen Practit 1971;21:373-6.
4. Farmer EA, Page G. A practical guide to assessing clinical decision-making skills using the key features approach. Med Educ 2005;39: 1188 -94.

5. McGuire CH, Babbott D. Simulation technique in the measurement of problem solving skills. J Educ Meas 1967;4:1-10.
6. จินตนา ศิรินาวิน, สาธิต วรรณแสง. ทักษะทางคลินิก, พิมพ์ครั้งที่ 2. กรุงเทพฯ: หมอชาวบ้าน, 2549.
7. Hodgkin K, Knox JDE. Problem centered learning. London, United Kingdom: Churchill Livingstone, 1975.
8. Stratford P, Pierce-Fenn H. Modified essay question. Phys Ther 1985; 65(1075-9).
9. Feletti GI, Smith EK. Modified essay quetions: Are they worth the effort? Med Educ 1986;20:126 - 32.
10. Rabinowitz HK. The modified essay question: An evaluation of its use in a family medicine clerkship. Med Educ 1987;21:114-8.
11. Wallerstedt S, Erickson G, Wallerstedt SM. Short answer questions or modified essay questions - More than a technical issue. Int J Clin Med 2012;3:28-30.
12. Lim EC, Seet RC, Oh VMS, Chia B, Aw M, S Q, et al. Computer-based testing of the modified essay question: The Singapore experience. Med Teach 2007;29:e261-8.
13. Norman G, Bordage G, Curry L, et al. Review of recent innovations in assessment. In: Wakeford R, editor. Directions in clinical assessment: Report of the Cambridge conference on the Assessment of Clinical competence. Cambridge: Office of the Regius Professsor of Physic, Cambridge University School of clinical Medicine, 1985:8-27.
14. Page G, Bordage G. The Medical Council of Canada's key features project: A more valid written examination of clinical decision-making skills. Acad Med 1995;70:104-10.
15. Neufeld VR, Norman GR, Barrows HS, Feightner JW. Clinical problem solving by medical students: A longitudinal and cross-sectional analysis. Med Educ 1981;15:315-22.
16. Perkins DN, Salomon G. Are cognitive skills context-bound? Educ Researcher 1989;18:16-25.
17. van der Vleuten CPM, Swanson DB. Assessing clinical skills with standardized patients: The state of the art. Teach Learn Med 1990;2 (58-76).
18. Eva KW. On the generality of specificity. Med Educ 2003;37(7): 587-88.
19. Bordage G, Page G. An alternate approach to PMPs, the key feature concept. In: Hart I, Harden R, editors. Further developments in assessing clinical competence. Montreal: Can-Heal Publications, 1987:57-75.
20. Page G, Bordage G, Allen T. Developing key features problems and examinations to assess clinical decision-making skills. Acad Med 1995;70:194-201.
21. Norman G, Bordage G, Page G, Keane D. How specific is case specificity? Med Educ 2006;40:618-23.
22. Hambleton RK, Pitoniak MJ. Setting performance standards. In: Brennan RL, editor. Educational measurement, 4th ed. Westport, CT: Praeger publishers, 2006:433-70.
23. Federation of State Medical Boards of the United States, National Board of Medical Examiners. USMLE Step 3: Content description and general information, Available from http://www.usmle.org/pdfs/step-3/2014content_Step3.pdf. June 2014.

ตามปกหน้าเวชบันทึกศิริราช ปีที่ 7 ฉบับที่ 2 กรกฎาคม-ธันวาคม 2557 หน้า 74-83 เรื่อง "หน้ากากครอบกล่องเสียง Laryngeal Mask Airway (LMA)" โดย อรุโณทัย ศิริอัศวกุล

ขอแก้ไขเป็น

เวชบันทึกศิริราช
ปีที่ 7 ฉบับที่ 2 กรกฎาคม-ธันวาคม 2557 หน้า 74-83 เรื่อง
"หน้ากากครอบกล่องเสียง Laryngeal Mask Airway (LMA)" โดย อังศุมาศ หวังดี
และได้ทำการแก้ไข pdf เรียบร้อยแล้ว

รศ. ดร.นพ.เชิดศักดิ์ ไอรมณีรัตน์

หัวข้อ : OSCE item development

OSCE Item Development

เชิดศักดิ์ ไอรมณีรัตน์
ภาควิชาศัลยศาสตร์ คณะแพทยศาสตร์ศิริราชพยาบาล
มหาวิทยาลัย มหิดล

OSCE

- Objective
- Structured
- Clinical
- Examination

- **มีวัตถุประสงค์ที่ชัดเจน**
- **มีการจัดโครงสร้างเป็นสถานีย่อย**
- **ประเมินทักษะทางคลินิก**
- **การสอบ**

History

- 1975: Ronald Harden (University of Dundee) proposed a series of stations in examination of clinical skills for 5 minutes per each station.
- 1988: Faculty of Medicine, Ramathibodi hospital implemented an OSCE in M3 exam (introduction to clinical medicine)
- 1991: Medical Council of Thailand implemented an OSCE in medical licensing exam for foreign graduates.
- 2009: Center for Medical Competency Assessment and Accreditation implemented an OSCE as Step 3 medical licensing exam.

OSCE

- Objective Structured Clinical Examination
- Assessment of clinical skills
  - History taking
  - Physical examination
  - Communication skills
  - Procedural skills
  - Interpretation of medical investigations
  - Ordering of medical treatment

Components of an OSCE item

1. Scenario (ภาพรวมสถานการณ์)
2. Instruction for examinees (คำแนะนำผู้เข้าสอบ)
3. Instruction for SPs (คำแนะนำผู้ป่วยมาตรฐาน)
4. Scoring rubric (ใบให้คะแนน +/- คำแนะนำอาจารย์)

Scenario

- Title
- Objectives
- Examinees
- Clinical information
- Apparatus
- SP requirements
- Time

Scenario 1

หัวข้อ : การตรวจร่างกายผู้ป่วยที่มีอาการปวดท้อง

Objective : นักศึกษาแพทย์สามารถแสดงวิธีการตรวจร่างกายผู้ป่วยที่มีอาการปวดท้องเฉียบพลัน และให้การวินิจฉัยที่ถูกต้องได้

ผู้สอบ: นักศึกษาแพทย์ชั้นปีที่ 6

สถานการณ์: สมบูรณ์ อายุ 35 ปี มีอาการปวดท้องใต้ชายโครงด้านซ้าย 6 ชั่วโมง ปวดตื้อ ๆตลอดเวลา

คำสั่ง : จงแสดงวิธีการตรวจหน้าท้องผู้ป่วย บรรยายสิ่งที่ตรวจพบและให้การวินิจฉัยโรคที่คิดถึงมากที่สุด 1 โรค

เวลา : 5 นาที (ตรวจร่างกาย 4 นาทีครึ่ง บอกสิ่งที่พบและวินิจฉัยครึ่งนาที)

Scenario 1 (cont.)

Apparatus	ผู้ป่วยสมมติ (ชายอายุ 30 – 40 ปี ไม่มีแผลผ่าตัดหน้าท้อง)	1 คน
	โต๊ะนั่งสำหรับกรรมการ	1 ตัว
	เก้าอี้นั่ง	1 ตัว
	เตียงตรวจร่างกาย	1 ตัว
	ผ้าปูเตียง หมอน และผ้าห่ม	1 ชุด
	เอกสารอธิบายและแบบฟอร์มการให้คะแนน	

Instruction for Examinees

- ผู้ป่วยหญิงไทยคู่ อายุ 22 ปี มีอาการปวดท้อง 4 ชั่วโมงก่อนมาโรงพยาบาล
- **คำสั่ง**
  1. จงซักประวัติผู้ป่วยรายนี้ (4 ½ นาที)
  2. จงบอกการวินิจฉัยโรคที่นึกถึงมากที่สุด (1/2 นาที)

Standardized Patient (SP)

- ผู้ป่วยมาตรฐาน
  - ผู้ป่วยจริง หรือ คนปกติมาแสดงเป็นผู้ป่วย
  - ได้รับการฝึกให้นำเสนออาการ หรือ อาการแสดงที่กำหนด
  - สามารถแสดงได้เหมือนบทบาทในการแสดงทุกครั้ง
  - เพื่อใช้ในการสอน หรือ ประเมินผลนักศึกษา

SP Script

- Challenges of SP script
- Types of SP script
  - Uncomplicated script
  - Complicated script

SP Script

- General information about the scenario
- Information of the portrayed patient
  - Name, age, and relevant personal information (occupation, family, etc.)
  - Dress (+/- make-up)
  - Medical history/ physical findings
    - If being asked ….., answered …
    - If being pressed ….., reacted….
    - Cue to portray or reveal special information/findings (cry, angry, guiding info., etc.)

Scoring Rubric General Format

หัวข้อการประเมิน	ปฏิบัติ		ไม่ปฏิบัติ
	สมบูรณ์	ไม่สมบูรณ์	
ตอนที่ 1. การปฏิบัติต่อผู้ป่วย	10 ครบ	6 อย่างน้อย 2	0 1 หรือ 0 ข้อ
ตอนที่ 2. รายละเอียดอาการ/การปฏิบัติ	5	3	0
ตอนที่ 3 การวินิจฉัยแยกโรค	XXXX 10 YYYY 8 ZZZZ 5		

Scoring Rubric

ขั้นตอนการประเมิน	สมบูรณ์	ไม่สมบูรณ์	ไม่ปฏิบัติ
1. <u>การแนะนำตัว</u>			
1.1 การแนะนำตัวเองอย่างสุภาพ	5	3	0
1.2 การถามชื่อผู้ป่วยอย่างสุภาพ	5	3	0
2. <u>การถามประวัติ</u>			
2.1 ถามตำแหน่งที่ปวด	10	–	0
2.2 ถามลักษณะของการปวด	10	6	0
2.3 ถามอาการปวดร้าวไปที่อื่น	10	–	0
…			
2.8 ถามประวัติประจำเดือน	10	6	0
3. <u>การวินิจฉัยโรค</u>			
Ectopic pregnancy	10		
Acute appendicitis	6		

Scoring Rubric

- กระชับ ได้ใจความ สื่อความหมายตรงกัน
- กำหนดประเด็นที่สำคัญ หรือเป็นจุดที่มักทำผิดพลาด
- บรรยายพฤติกรรมที่ผู้ประเมินสังเกตได้
- กำหนดน้ำหนักคะแนนตามความสำคัญ

Iramaneerat C. Guidelines in developing an objective structured clinical examination: Case content [Thai]. Medical Education Pamphlet 2005; 1(8): 4.

ข้อแนะนำในการจัดสอบ OSCE (Objective Structured Clinical Examination) (ตอนที่ 1)

เชิดศักดิ์ ไอรมณีรัตน์

Objective Structured Clinical Examination (OSCE) เป็นเทคนิคที่เป็นที่ยอมรับและ ได้รับการใช้มากขึ้นเรื่อยๆ ทั้งการสอนและประเมินผล ทางแพทยศาสตรศึกษาทุกระดับทั่วโลก ผมจะขอเสนอเกร็ดความรู้เกี่ยวกับการจัดสอบ OSCE โดยแบ่งออกเป็น 3 ตอนตามส่วนประกอบสำคัญของ OSCE ได้แก่ เนื้อหาของโจทย์ (content) ผู้ป่วยมาตรฐาน (standardized patient) และ อาจารย์ผู้ให้คะแนน (rater) ในบทความนี้จะขอกล่าวถึง เนื้อหาของโจทย์

1. สิ่งแรกที่ต้องคำนึงถึงคือวัตถุประสงค์ของการสอบ เนื่องจาก OSCE เป็นการสอบที่ต้องใช้ทรัพยากรมาก ควรตั้งวัตถุประสงค์การสอบเพื่อประเมินความรู้ความสามารถที่ไม่สามารถประเมินได้ด้วยวิธีอื่น เช่นทักษะในการสื่อสารกับผู้ป่วย ทักษะการให้คำแนะนำแก่ผู้ป่วย ทักษะการทำหัตถการ เป็นต้น ไม่ควรใช้ OSCE เพื่อวัดความรู้ผิวเผินที่สามารถวัดได้ด้วยข้อสอบ MCQ
2. วางแบบแปลนของเนื้อหาข้อสอบ (test blueprint) ที่ครอบคลุมเนื้อหาวิชาในทุกด้าน และทุกทักษะที่ต้องการประเมินอย่างเท่าเทียมกัน มีการระบุชัดว่าในการสอบ OSCE นี้ทดสอบความรู้เรื่องใดบ้าง (โรคปอด โรคหัวใจ โรคไต ฯลฯ) และใช้ทักษะใดบ้าง (การซักประวัติ การตรวจร่างกาย การให้คำแนะนำ ฯลฯ) อย่างละกี่ข้อ ระวังอย่าให้เนื้อหาข้อสอบมีน้ำหนักในเรื่องใดเรื่องหนึ่งมากกว่าเรื่องอื่น
3. ในการเขียนโจทย์ OSCE แต่ละข้อ ต้องเขียนให้ครอบคลุมรายละเอียดทุกด้านของการสอบ ได้แก่ คำชี้แจงสำหรับนักเรียน สำหรับผู้ป่วยมาตรฐาน และสำหรับอาจารย์ผู้คุมสอบ สถานการณ์ผู้ป่วยจำลอง ประวัติและผลการตรวจร่างกายที่ผู้ป่วยมาตรฐานต้องแสดงออก อุปกรณ์ประกอบที่ต้องใช้ ระยะเวลาที่ต้องใช้ แบบฟอร์มให้คะแนน และเกณฑ์การให้คะแนน
4. การเขียนโจทย์ผู้ป่วยควรนำข้อมูลมาจากผู้ป่วยจริง ซึ่งจะทำให้โจทย์มีความเหมือนจริง ไม่ขาดรายละเอียดในเนื้อหาของโจทย์ และประหยัดเวลาในการแต่งโจทย์ นอกจากนี้ยังทำให้มีแฟ้มประวัติและผลการตรวจเพิ่มเติมรวมทั้งฟิล์มที่สามารถนำมาใช้เสริมโจทย์ได้ง่าย
5. โจทย์สำหรับแต่ละสถานีควรมีความยาวเหมาะสม โจทย์ที่ใช้เวลานานสามารถให้ข้อมูลเกี่ยวกับความสามารถของนักเรียนในเรื่องนั้นๆ ได้ละเอียด แต่ก็ทำให้มีโอกาสวัดความสามารถของนักเรียนได้น้อยเรื่อง เนื่องจากทักษะทางการแพทย์หลายด้านมีความเจาะจงต่อภาวะโรค (นักเรียนที่ซักประวัติโรคเลือดได้ดีอาจซักประวัติผู้ป่วยโรคซึมเศร้าไม่คล่องได้) โดยทั่วไปแนะนำให้จัดเวลาที่ใช้สอบในแต่ละสถานี ให้นักเรียนได้มีโอกาสสอบในอย่างน้อย 8 – 10 สถานี (ยิ่งมีสถานีสอบมาก ผลการสอบยิ่งมีความแม่นยำมาก) หลายการศึกษาพบว่าเพื่อให้ได้ผลการสอบ OSCE ที่มีความแม่นยำพอยอมรับได้ จะต้องใช้เวลาในการสอบอย่างน้อย 3 – 4 ชั่วโมง
6. จัดให้มีการตอบคำถามตามหลังการสอบทักษะกับผู้ป่วย (post-encounter probe) เท่าที่จำเป็น ไม่มากเกินไป เนื่องจากคำถามเหล่านี้มักวัดความสามารถที่แตกต่างไปจากวัตถุประสงค์หลักของการสอบ OSCE (มักวัดความรู้ในทำนองเดียวกับ MCQ) จึงเป็นการเพิ่มเวลาสอบโดยไม่จำเป็นและยังลดความแม่นยำของผลการสอบอีกด้วย

Iramaneerat C. Guidelines in developing an objective structured clinical examination: Standardized patients [Thai]. Medical Education Pamphlet 2005; 1(9): 3.

ข้อแนะนำในการจัดสอบ OSCE (Objective Structured Clinical Examination) (ตอนที่ 2)

เชิดศักดิ์ ไอรมณีรัตน์

ในบทความนี้จะขอเสนอเกร็ดความรู้เกี่ยวกับการใช้ผู้ป่วยมาตรฐาน (Standardized patients) ใน OSCE ก่อนอื่นผมขอกล่าวถึงนิยามของศัพท์ที่สำคัญในการใช้ผู้ป่วยในการสอบก่อน เราเรียกคนปกติที่ไม่มีความเจ็บป่วย แต่แสดงบทบาทเป็นผู้ป่วยว่า ผู้ป่วยสมมติ (simulated patient) ซึ่งผู้ป่วยสมมติเหล่านี้อาจแสดงออกไม่สม่ำเสมอ เมื่อได้พบกับนักเรียนแต่ละคน หากเราทำการฝึกให้ผู้ป่วยสมมติ (หรือ ผู้ป่วยจริง) แสดงออกซึ่งอาการและอาการแสดงอย่างสม่ำเสมอ เป็นมาตรฐานเดียวกัน ไม่ว่าจะได้พบกับนักเรียนคนใด เราจะได้ ผู้ป่วยมาตรฐาน (standardized patient) การสอบ OSCE ให้ได้ผลการประเมินที่แม่นยำนั้นต้องใช้ผู้ป่วยมาตรฐาน (standardized patient, SP)

1. ผู้ป่วยมาตรฐานต้องได้รับการฝึกฝนอย่างดีจนมั่นใจว่าการแสดงออกซึ่งอาการและอาการแสดงได้มาตรฐานในทุกครั้งที่แสดงบทบาท การฝึกฝนนี้ต้องเริ่มต้นจากการมีบท (script) ที่ดี มีความละเอียดครอบคลุมข้อมูลทุกด้านที่เกี่ยวกับภาวะโรคที่สนใจ และมีการฝึกซ้อมและตรวจแก้โจทย์โดยอาจารย์ผู้แต่งโจทย์เพื่อให้มั่นใจว่าความเข้าใจบทบาทของผู้ป่วยมาตรฐานถูกต้องตามความตั้งใจของผู้แต่งโจทย์ โดยทั่วไปเมื่อได้รับการฝึกฝนแล้วผู้ป่วยมาตรฐานสามารถแสดงออกซึ่งอาการและอาการแสดงได้อย่างถูกต้องมากกว่า 90%
2. ในการสอบใหญ่บางครั้งมีความจำเป็นต้องใช้ผู้ป่วยมาตรฐานหลายคนเพื่อแสดงบทบาทเดียวกัน มีหลายการศึกษาแสดงว่าการใช้ผู้ป่วยมาตรฐานหลายคนในลักษณะนี้ไม่ลดความแม่นยำของผลสอบตราบเท่าที่เรามีสถานีสอบ OSCE มากเพียงพอ และผู้ป่วยมาตรฐานได้ถูกสุ่มกระจายตัวอยู่ตามสถานีสอบอย่างไม่ลำเอียง (randomly distributed)
3. หลายการศึกษาที่วิเคราะห์การสอบที่มีความจำเป็นต้องใช้ผู้ป่วยมาตรฐานชุดเดิมสอบนักเรียนหลายชุดต่อเนื่องกัน พบว่านักเรียนที่สอบรอบหลังไม่ได้ทำคะแนนได้ดีกว่านักเรียนที่สอบรอบแรก แสดงว่านักเรียนที่สอบก่อนไม่ให้ข้อมูลเกี่ยวกับการสอบที่เป็นประโยชน์แก่นักเรียนที่สอบรอบหลัง หรือหากนักเรียนให้ข้อมูลแก่กัน ข้อมูลเพียงที่ได้รับเกี่ยวกับคำชี้แจงโจทย์โดยไม่มีข้อมูลรายละเอียดของเกณฑ์การให้คะแนนนั้นไม่ได้ก่อให้เกิดความได้เปรียบในการสอบแก่นักเรียนรอบหลัง
4. นอกจากจะใช้ผู้ป่วยมาตรฐานเพื่อวัดทักษะของนักเรียนที่เกี่ยวเนื่องกับผู้ป่วยโดยตรง (เช่นการซักประวัติ ตรวจร่างกาย) แล้ว เรายังสามารถใช้ผู้ป่วยมาตรฐานประกอบกับแบบจำลองเพื่อทดสอบทักษะการทำหัตถการเพื่อทำให้การปฏิบัติหัตถการมีความสมจริงได้ด้วย เช่น การนำแบบจำลองสำหรับเย็บแผลมาติดกับแขนของผู้ป่วยจำลอง จะช่วยให้สามารถวัดทักษะในการเย็บแผลในขณะเดียวกันกับที่ต้องมีปฏิสัมพันธ์กับผู้ป่วยที่มีความเจ็บปวดจากบาดแผลด้วย

Iramaneerat C. Guidelines in developing an objective structured clinical examination: Scoring [Thai]. Medical Education Pamphlet 2005; 1(10): 1.

ข้อแนะนำในการจัดสอบ OSCE (Objective Structured Clinical Examination) (ตอนที่ 3)

เชิดศักดิ์ ไอรมณีรัตน์

ในบทความนี้จะขอเสนอเกร็ดความรู้เกี่ยวกับการให้คะแนนในการสอบ OSCE

1. การให้คะแนน OSCE ทำได้ 2 วิธีใหญ่ๆ ด้วยกัน คือ checklist (ให้คะแนน 1 เมื่อทำสิ่งที่ระบุในรายการ และให้คะแนน 0 เมื่อไม่ทำรายการนั้น เช่น “นักเรียนถามประวัติประจำเดือนครั้งสุดท้าย”: 0 ทำ, 1 ไม่ทำ) และ rating scale (ให้คะแนนได้หลายระดับขึ้นกับระดับความถูกต้องของการปฏิบัติ เช่น “นักเรียนอธิบายหัตถการที่จะทำได้ชัดเจน” : 1 ไม่เห็นด้วยอย่างยิ่ง, 2 ไม่เห็นด้วย, 3 เห็นด้วย, 4 เห็นด้วยอย่างยิ่ง) การให้คะแนนด้วย checklist จะได้ผลการประเมินที่ผู้ให้คะแนน (rater) มีความเห็นพ้องกัน (inter-rater agreement) มากกว่า แต่สามารถแยกแยะความแตกต่างระหว่างนักเรียนที่มีความสามารถต่างกันได้ไม่ดีเท่ากับการให้คะแนนด้วย rating scale ควรใช้ checklist สำหรับให้คะแนนโจทย์ที่ประเมินความครบถ้วนของเนื้อหาหรือขั้นตอน (เช่น ซักประวัติ ตรวจร่างกาย) แต่ควรใช้ rating scale สำหรับให้คะแนนโจทย์ที่ประเมินคุณภาพของทักษะหรือกระบวนการปฏิบัติ (เช่น ทักษะการสื่อสาร ทักษะการทำหัตถการ)
2. ไม่มีความจำเป็นต้องใช้ผู้ให้คะแนน (rater) มากกว่า 1 คน ต่อ 1 สถานี หากมีทรัพยากรบุคคลมากพอ เราควรจะเพิ่มจำนวนสถานีสอบ มากกว่า เพิ่มจำนวนผู้ให้คะแนนต่อสถานี การเพิ่มจำนวนสถานีสอบส่งผลให้คะแนนสอบ OSCE มีความแม่นยำเพิ่มขึ้นมากกว่า การเพิ่มจำนวนผู้ให้คะแนนต่อสถานี
3. นอกจากเราจะให้อาจารย์แพทย์เป็นผู้ให้คะแนนแล้ว เรายังสามารถฝึกให้ผู้ป่วยมาตรฐาน (standardized patient) ทำการให้คะแนนได้ด้วย พบว่าเมื่อได้รับการอธิบายเกณฑ์การให้คะแนนและฝึกปฏิบัติแล้ว ผู้ป่วยมาตรฐาน สามารถให้คะแนนที่มีความแม่นยำสูงไม่แพ้อาจารย์แพทย์ ข้อดีของการให้ผู้ป่วยมาตรฐานเป็นผู้ให้คะแนนคือสะดวก และประหยัด ในทางกลับกันการให้อาจารย์แพทย์เป็นผู้ให้คะแนนมีข้อได้เปรียบคืออาจารย์สามารถชี้แนะข้อบกพร่องและแนะนำแนวทางการปรับปรุงแก้ไขทักษะและวิธีคิดของนักเรียนได้ทันที
4. ไม่ควรใช้ผลการประเมินจากสถานีใดสถานีหนึ่งเป็นตัวบ่งชี้ว่านักเรียนมีความสามารถหรือไม่มีความสามารถในด้านใด เนื่องจากผลการประเมินจากสถานีเดียวมีโอกาสผิดพลาดได้มาก การตัดสินว่านักเรียนคนใดมีความสามารถหรือไม่ให้ใช้ผลการประเมินโดยรวมซึ่งมีความแม่นยำมากกว่า
5. การรายงานคะแนน OSCE แก่นักเรียนนั้นต้องคำนึงถึงวัตถุประสงค์ของการสอบ หากทำการสอบ formative test ควรบอกข้อดี ข้อด้อย ของนักเรียนแต่ละคน และชี้แจงสิ่งที่ควรปรับปรุงอย่างละเอียด ส่วนคะแนนรวมนั้นอาจไม่ค่อยมีความสำคัญนัก ในทางกลับกัน หากทำการสอบ summative test เราต้องคำนึงถึงการรักษาความลับของข้อสอบ เนื่องจากข้อสอบ OSCE ที่ดีนั้นพัฒนาขึ้นได้ยาก และควรได้รับการเก็บไว้ในคลังข้อสอบเพื่อนำมาใช้ในอนาคต ดังนั้นเราไม่ควรแจงรายละเอียด ข้อถูก ข้อผิด ของนักเรียนแต่ละคนในทุกสถานี แต่แจ้งเพียงผลสอบว่าผ่านหรือไม่ผ่าน

เอกสารประกอบการอบรม

30 October 2020

ผศ. นพ.สุประพัฒน์ สนใจพาณิชย์

หัวข้อ : Long case examination

Mahidol University
Faculty of Medicine
Siriraj Hospital

SHEE

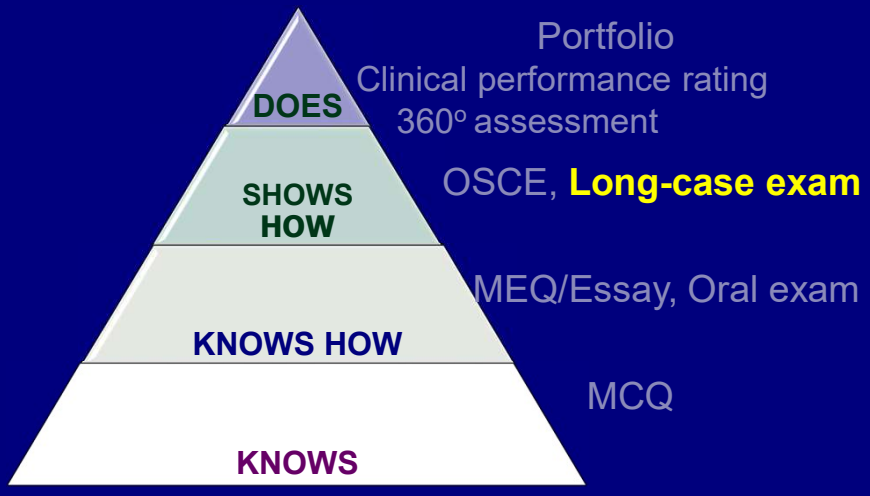
Long Case Examination

Suprapath Sonjaipanich MD.
Department of Pediatrics
Faculty of Medicine Siriraj Hospital
Mahidol University

Mahidol University
Faculty of Medicine
Siriraj Hospital

SHEE

Assessment approach



Miller's pyramid

Mahidol University Faculty of Medicine Siriraj Hospital

SHEE

Long Case Examination

- Construct underrepresentation
- Unstructured settings
- Adequacy of observation
- Fairness among students
- Subjective ratings
- Low generalizability
- Low reliability

1. Norcini JJ. The death of the long case? *BMJ*. 2002
2. Wilkinson TJ, Campbell PJ, Judd SJ. Reliability of the long case. Medical education. 2008
3. Ponnamperuma GG, Karunathilake IM, McAleer S, Davis MH. The long case and its modifications: a literature review. *Med Educ*. 2009

Mahidol University Faculty of Medicine Siriraj Hospital

SHEE

Long-case Examination

Problems

- Validity
- Reliability
- Objectivity

In high stake summative assessment
Long case exam should be avoided

"Luck of the draw:
different examiners examine
different candidates on different patients"

Stokes J F. 'The clinical examination: assessment of clinical skills'. Publisher: Dundee, Association for the Study of Medical Education 1974

Mahidol University
Faculty of Medicine
Siriraj Hospital

SHEE

Long Case Examination

Strengths

- Comprehensive competency evaluation
- In-depth exploration of knowledge, skills
- A powerful tool for providing feedback
- A unique opportunity to test the physician's tasks and interaction with a real patient

1. Vleuten C. Making the best of the "long case". Lancet 1996
2. Ponnamperuma GG, Karunathilake IM, McAleer S, Davis MH. The long case and its modifications: a literature review. *Med Educ.* 2009

Mahidol University
Faculty of Medicine
Siriraj Hospital

SHEE

National Medical Licensing Examination for medical undergraduates in Thailand

Step 1: MCQ in Basic medical science

Step 2: MCQ in Clinical science

Step 3: Clinical skills and problem solving

1. OSCE
2. MEQ
3. Long case exam

Mahidol University Faculty of Medicine Siriraj Hospital

SHEE

The Objective Structured Long Examination Record (OSLER)

- 10 items
  - 4 on history
  - 3 on physical examination
  - 3 on investigation, management, and clinical acumen
- Prior agreement on what to be examined
- Assess both processes and products
- Identification of case difficulty by an examiner

Gleeson F. AMEE Medical Education Guide No 9. Assessment of clinical competence using the Objective Structured Long Examination Record (OSLER). *Med Teach*. 1997; 19(1): 7-14.

Mahidol University Faculty of Medicine Siriraj Hospital

SHEE

OSLER's components

History taking	Physical examination	Investigations, Management, Clinical acumen
• Clarity of presentation • Communication process • Systematic presentation • Establishment of case facts	• Systematic approach • Examination technique • Establishment of correct physical findings	• Appropriate investigation • Logical sequence • Ability to solve problems

Mahidol Univers
Faculty of Medicine
Siriraj Hospital

SHEE

OBJECTIVE STRUCTURED LONG EXAMINATION RECORD (OSLER) DATE: .....

CANDIDATE'S ..... EXAMINATION NO.
Name

Examiners are required to **GRADE** each of the ten items below
And assign an overall **GRADE** and **MARK** concerning the candidate
PRIOR to discussion with their co-examiner as follows

EXAMINER:
.....

GRADE
P+ = Very good/excellent
P = Pass/Borderline pass
P- = Below pass

MARKS
(60-80+) see over page
(50-55) for specific
(35-45) mark details

CO-EXAMNIER:
.....

PRESENTATION OF HIRTORY	GRADE	AGREED GRADE
PACE/CLARITY	_____	_____
COMMUNICATION PROCESS	_____	_____
SYSTEMATIC PRESENTATION	_____	_____
CORRECT FACTS ESTABLISHED	_____	_____
PHYSICAL EXAMINATION		
SYSTEMIC	_____	_____
TECHNIQUE (including attitude to patient)	_____	_____
CORRECT Findings ESTABLISHED	_____	_____
APPROPRIATE INVESTIGATIONS IN A LOGICAL SEQUENCE (communication process option)	_____	_____
CLINICAL ACUMEN (problem identification/problem solving ability)	_____	_____

ADDITIONAL COMMENTS:

_____

Please Tick (√) for CASE DIFFICULTY

	Individual examiner	Agreed case difficulty
Standard	_____	_____
Difficult	_____	_____
Very difficult	_____	_____

GRADE **Individual examiner**

OVERALL GRADE	MARK

PAIRED OF EXAMINERS

AGREED GRADE	AGREED MARK

Mahidol Unive
Faculty of Medicine
Siriraj Hospital

SHEE

EXTENDED CRITERION REFERENCED GRADING SCHEME	EXTENDED MARKING SCHEME
P +	80 OUTSTANDINGLY clear and factually correct presentation of the patient's history, demonstration of physical signs, and organisation of the case management. Clearly, a candidate displaying outstanding communication skills and clinical acumen. First class honours. 75 EXCELLENT OVERALL case presentation, communication skills, examination technique, and demonstration of the correct facts and physical signs of the case. The candidate may even display outstanding attributes in some but not all measurable criteria. First class honours. 70 EXCELLENT IN MOST RESPECTS of overall case presentation, communication skills, examination technique, and demonstration of the correct facts and physical signs of the case. Also excellent communicator and demonstrates the ability to investigate and appropriately manage the patient with a very well developed clinical acumen. First class honours. 65 VERY GOOD OVERALL presentation covering all major aspects; few omissions, good priorities. Very clearly an above average candidate in terms of communication skills and clinical acumen. Second class honours, division 1. 60 VERY GOOD IN MOST RESPECTS of presentation and communication, but not in all respects. However, a good solid performance in most areas assessed with a well developed clinical acumen. Second class honours, division 2.
P	55 GOOD SOUND OVERALL presentation and communication of the case without displaying attributes out of the ordinary. The candidate displays an overall adequate standard of examination technique. The patient's problems are identified and a reasonable management outline suggested. 50 ADEQUATE presentation of the case and communication ability. Nothing to suggest more than just reaching an acceptable standard in physical examination and identification of the patient's problems and their management. Clinical acumen just reaching an acceptable standard. Safe borderline candidate who just reaches a pass standard.
P–	45 POOR performance in terms of case presentation, communication with the patient, and demonstration of physical signs. Inadequate attempt at a clear identification of the patient's problems. The candidate may display some adequate attributes but does not reach an acceptable pass standard overall. THE MARK 40 IS NOT USED IN CLINICALS 35 VETO MARK The candidate's performance in terms of case presentation, clinical, and communication skills is so poor that the standard required is not even remotely approached. Quite clearly this candidate requires a further period of training.

Mahidol University Faculty of Medicine Siriraj Hospital

SHEE

OSLER

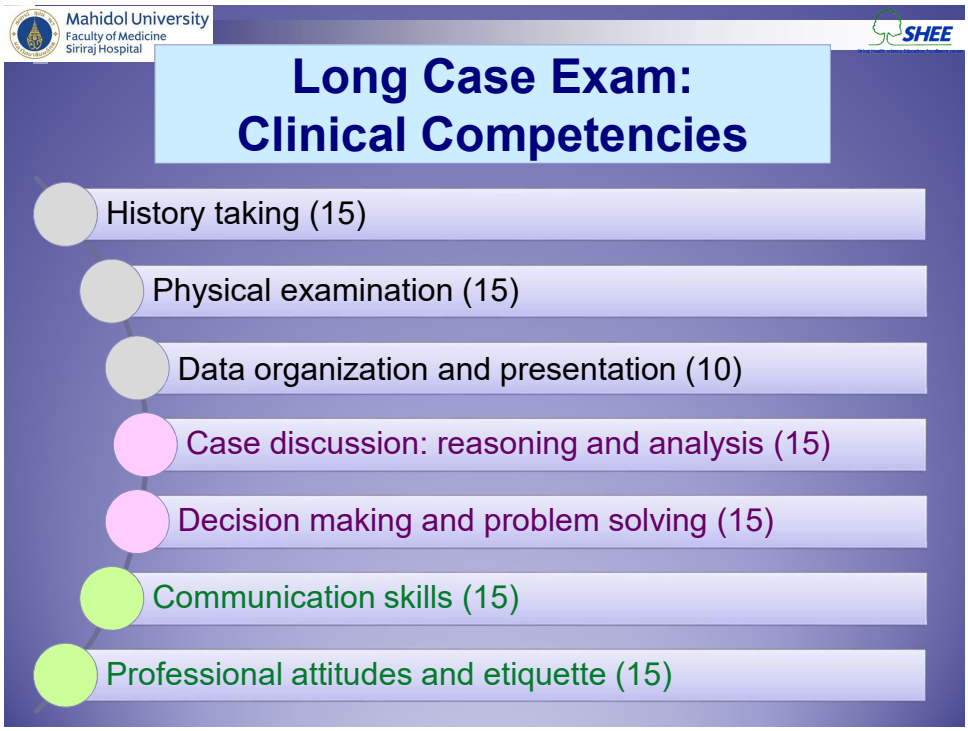
- To standardize patients
  - Real patient, No SP
  - Case difficulty
    1. Standard case: 1 problem
    2. Difficult: up to 3 problems
    3. Very difficult: > 3 problems
- To standardize examiners
  - Two examiners
  - Increased number of items and fixed structure
  - “Conscious” examiner; measure what it is supposed to measure

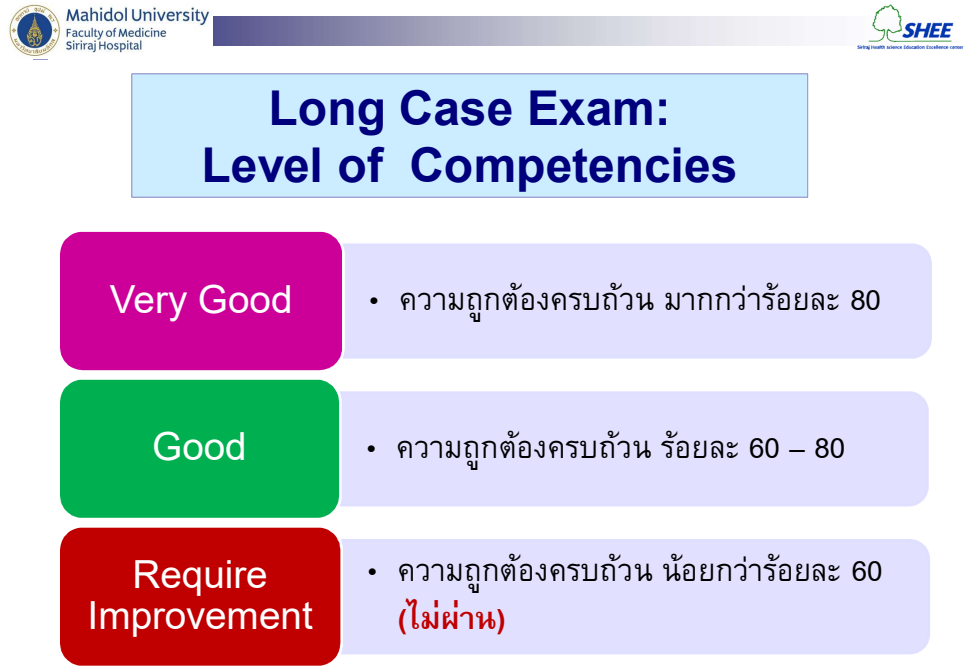
Mahidol University Faculty of Medicine Siriraj Hospital

SHEE

ข้อกำหนดของ ศรว. ในการสอบ long case

1. จำนวนผู้ป่วยอย่างน้อย 2 ราย
2. โรค หรือ ปัญหาสอดคล้องกับเกณฑ์มาตรฐานฯแพทยสภา
3. ผู้ป่วยใน หรือ ผู้ป่วยนอก
4. รูปแบบการสอบ 3 ขั้นตอน
   1) Patient encounter under direct observation 30 นาที
   2) Case discussion 20 – 30 นาที
   3) Patient encounter 10 นาที


Mahidol University
Faculty of Medicine
Siriraj Hospital
SHEE
Long Case Exam:
Clinical Competencies
History taking (15)
Physical examination (15)
Data organization and presentation (10)
Case discussion: reasoning and analysis (15)
Decision making and problem solving (15)
Communication skills (15)
Professional attitudes and etiquette (15)


Mahidol University
Faculty of Medicine
Siriraj Hospital
SHEE
Long Case Exam:
Level of Competencies
Very Good
• ความถูกต้องครบถ้วน มากกว่าร้อยละ 80
Good
• ความถูกต้องครบถ้วน ร้อยละ 60 – 80
Require Improvement
• ความถูกต้องครบถ้วน น้อยกว่าร้อยละ 60
(ไม่ผ่าน)

รศ. พญ.กษณา รักษมณี

หัวข้อ : Portfolio

เวชบันทึกศิริราช บทความทั่วไป

การใช้แฟ้มสะสมงานอย่างมีประสิทธิภาพในทางแพทยศาสตรศึกษา (Effective Uses of Portfolio in Medical Education)

ผู้ช่วยศาสตราจารย์ แพทย์หญิงกษณา รักษมณี
ภาควิชาวิสัญญีวิทยา, คณะแพทยศาสตร์ศิริราชพยาบาล, มหาวิทยาลัยมหิดล, กรุงเทพมหานคร 10700.

● แพทยศาสตรศึกษาในปัจจุบันนั้นกำลังพัฒนาไปในทิศทางที่ส่งเสริมให้ผู้เรียนมี competency ในด้านต่าง ๆ ตามคุณลักษณะที่พึงมี และพัฒนาตนเองให้เป็น lifelong learners จึงมุ่งเน้นการเรียนรู้แบบผู้ใหญ่ ซึ่งมีลักษณะสำคัญคือผู้เรียนเป็นคนสำคัญที่จะกำหนดทิศทางการเรียนรู้โดยผู้วางหลักสูตรจะต้องกำหนด outcome และ competency ต่าง ๆ ที่ต้องการ และผู้เรียนเป็นผู้รับผิดชอบต่อพัฒนาการของตนเอง การใช้แฟ้มสะสมงาน (portfolio) นั้นเป็นวิธีสำคัญที่ช่วยสนับสนุนให้มีการติดตามระยะยาวและวางแผนพัฒนา competency ต่าง ๆ จนลุล่วงและได้ตาม outcome ของหลักสูตรที่วางไว้

● Portfolio นั้นเป็นสิ่งที่ใช้กันอย่างแพร่หลายในวงการการศึกษาทั่ว ๆ ไป ทั้งระดับประถม มัธยม จนถึงระดับอุดมศึกษา และได้รับความนิยมและมีผลอย่างมากในวิชาชีพอื่น เช่นนักดนตรี ศิลปิน จิตรกร สำหรับในทางแพทยศาสตรศึกษานั้นยังจัดว่าค่อนข้างใหม่ และบางครั้งอาจใช้ประโยชน์ได้ไม่เต็มที่นัก แต่จริง ๆ แล้วการใช้ portfolio ในทางแพทยศาสตรศึกษานั้นสามารถใช้ได้ทั้งระดับก่อนปริญญาและหลังปริญญา เพื่อติดตามและเพิ่มพูนพัฒนาการของนักศึกษาในระยะยาว

Portfolio คืออะไร

● Portfolio เป็นการรวบรวมผลงานของนักศึกษา ซึ่งขึ้นกับผู้บริหารหลักสูตร ว่ามีวัตถุประสงค์ใดในการนำ portfolios มาใช้ และจะเลือกองค์ประกอบใดมาไว้ในแฟ้ม เพื่อแสดงถึงพัฒนาการในระยะยาวในด้านใดบ้าง ทั้งทางด้านความรู้ ทักษะในด้านต่างๆ ทัศนคติ รวมไปถึงพัฒนาการความเป็นวิชาชีพแพทย์

● Portfolio แตกต่างจาก logbooks ที่ใช้กันอย่างแพร่หลายในการเก็บผลงานและติดตามพัฒนาการทางด้านทักษะบางประการของนักศึกษา เนื่องจากการสะสมผลงานใน logbooks มักเน้นที่การเก็บจำนวนผู้ป่วยหรือกรณีศึกษาเท่านั้น ในขณะที่ portfolio ต้องประกอบไปด้วยกระบวนการ reflection คือการคิดทบทวนถึงผลงานนั้น ๆ ว่ามีสิ่งใดที่ทำได้ดีแล้ว และมีสิ่งใดที่มีโอกาสพัฒนาบ้าง ซึ่งเพิ่มโอกาสในการเรียนรู้โดยการสะสมประสบการณ์

● การทำ reflection นั้น เกิดขึ้นได้เมื่อนักศึกษาได้ทบทวนผลงานของตนเองที่ได้สะสมมาเป็นระยะยาว ซึ่งเป็นกระบวนการที่ทำให้นักศึกษาได้ทบทวนถึงประสบการณ์การเรียนรู้ที่ผ่านมา ทำให้เข้าใจว่าสิ่งใดได้เรียนรู้ไปแล้วและสิ่งใดที่ยังขาดอยู่ อีกทั้งเข้าใจถึงพัฒนาการทางการเรียนของตนเอง ทำให้สามารถวางแผนการเรียนอย่างมีประสิทธิภาพต่อไปได้

● การใช้ portfolio จะมีประสิทธิภาพมากขึ้นเมื่อใช้ร่วมกับระบบอาจารย์ที่ปรึกษา โดยให้อาจารย์ร่วมดูแลพัฒนาการและร่วมประเมิน portfolio เป็นระยะ ๆ สามารถช่วยนักศึกษาหาจุดแข็ง และข้อควรพัฒนา และร่วมกันวางแผนพัฒนาในระยะยาวได้

ประโยชน์ของ portfolio

● Portfolio สามารถใช้ประโยชน์ทั้งเพื่อการเรียนการสอนและการประเมินผล การเรียนการสอนนั้นเกิดโดยการที่ทำให้การเรียนรู้เห็นเป็นรูปธรรมจับต้องได้โดยการนำมาใส่แฟ้ม และนำมาเปรียบเทียบให้เห็นพัฒนาการของผู้เรียนเอง อีกทั้งกลไกการ reflection นั้นช่วยทำให้ประสบการณ์ที่ผ่านมา เช่น การได้เรียนรู้ในเคสใด ๆ ผ่านการกลั่นกรองและเชื่อมโยงความรู้ ทำให้สามารถต่อยอดความรู้ขึ้นไปอีกได้ การได้ทบทวนประสบการณ์ของตนเอง ยังช่วยให้นักศึกษาเห็นข้อดีและจุดบกพร่อง สามารถวางแผนพัฒนาตนเองในระยะยาวได้

● การใช้ portfolio เพื่อการประเมินนั้นใช้เสริมจากเพื่อการเรียนรู้ได้เช่นกัน เพียงเพิ่มการวัดผลเพื่อให้ผ่านเกณฑ์มาตรฐานซึ่งการประเมินโดย portfolio มีจุดเด่นคือมีความเชื่อมโยงที่ใกล้ชิดของการเรียนรู้และการประเมินเนื่องจากต้องมีการติดตามผลระยะยาว อีกทั้งยังสามารถประเมินในเรื่องที่ทำได้ยากโดยการประเมินอื่น ๆ (สอบในกระดาษ ประเมินจากการปฏิบัติงาน) เช่นการประเมินทัศนคติของวิชาชีพแพทย์ การประเมินความรับผิดชอบต่อการเรียนรู้ของตนเอง และความสามารถในการพัฒนาตนเองอย่างต่อเนื่อง โดยการจัดตั้ง portfolio ที่ใช้เพื่อประเมินผล ต้องวางโครงสร้างของ portfolio ให้ดีและสอดคล้องกับ outcome ในด้านต่าง ๆ ของหลักสูตรที่ต้องการให้นักศึกษามีความสามารถ จึงจะนำมาช่วยในการประเมินได้อย่างมีประสิทธิภาพ

องค์ประกอบของ portfolio

● Portfolio ของแต่ละสถาบัน หรือแต่ละหลักสูตรนั้นอาจมีความหลากหลายในแง่ของรูปแบบและองค์ประกอบ แต่ส่วนสำคัญที่เป็นองค์ประกอบที่เหมือนกันคือ ผลงานที่สะสมเป็นระยะยาว feedback ที่ได้รับ ความก้าวหน้าและพัฒนาการของนักศึกษา และแผนการเพื่อพัฒนา competency ของนักศึกษา

● องค์ประกอบของ portfolio นั้นขึ้นกับวัตถุประสงค์ เช่นหากมีวัตถุประสงค์หลักคือเพื่อการเรียนรู้ ก็ควรจัดองค์ประกอบให้แสดงถึงพัฒนาการของนักศึกษา เช่น ผลการประเมินทักษะหัตถการต่างๆ ความก้าวหน้าของงานวิจัย เคสสำคัญที่ได้พบได้ศึกษา และ reflective note ที่แสดงถึงการเรียนรู้ในช่วงต่างๆ แต่หากมีวัตถุประสงค์เพื่อการประเมินผล ก็ควรจัดองค์ประกอบที่แสดงให้เห็นถึงความสามารถที่ผ่านตามเกณฑ์ เช่น ผลการประเมินหลายรูปแบบทั้งระหว่างภาคเรียนและปลายภาค หลักฐานที่แสดงว่าผ่านการอบรมในหลักสูตรที่จำเป็นเช่น การอบรมการช่วยคืนชีพ รวมถึงผลงานที่ได้นำเสนอเช่น นำเสนองานวิจัยในงานประชุมวิชาการ เป็นต้น

● การเลือกรูปแบบโครงสร้างของ portfolio หากทำได้อย่างเหมาะสมและสอดคล้องกับวัตถุประสงค์ก็จะช่วยส่งเสริมการเรียนรู้ และเพิ่ม competency ของนักศึกษาได้ โดยโครงสร้างของ portfolio นั้นมีอยู่ 4 แบบหลัก ๆ คือ

1. Shopping trolley: เป็นการจัด portfolio ที่มีรูปแบบน้อยที่สุด คือการสะสมผลงานและกิจกรรมทุกอย่างของนักศึกษา ที่ผ่านไปในช่วงเวลาของการเรียน ซึ่งอาจรวมไปถึงวารสารหรือ guideline ที่นักศึกษาสนใจหรือใช้ในการศึกษาหรือทำโครงการใด ๆ ไปจนถึงจดหมายขอบคุณจากผู้ป่วย โดยผู้ที่จะเลือกว่าจะนำสิ่งใดใส่หรือไม่ใส่ไว้ในแฟ้มคือนักศึกษาเอง และแฟ้มประเภทนี้มักไม่ได้รับการประเมินจากอาจารย์ที่ปรึกษา การใช้แฟ้มชนิดนี้เพื่อให้นักศึกษาได้ทบทวนพัฒนาการของตนเอง โดยเลือกสิ่งที่สนใจจะเก็บสะสมเอง

2. Toast rack: เป็นรูปแบบของ portfolio ที่มีการจัดเป็นช่องๆ ตามหมวดหมู่ โดยแต่ละหมวดหมู่แยกจากกันชัดเจนและแสดงถึงพัฒนาการของแต่ละหมวด ซึ่งอาจแบ่งตาม competency ที่ต้องการติดตาม โดยจะต้องมีการติดตามอย่างมีแบบแผน เช่น การประเมินทักษะการทำหัตถการในการปฏิบัติงานจริง (clinical evaluation exercise) โดยที่แต่ละหมวดแยกออกจากกันอย่างชัดเจนและไม่มีการรวมผลการประเมินเข้าด้วยกัน

3. Cake mix: เป็นรูปแบบที่มีการแบ่งเป็นหมวดหมู่เช่นกัน แต่ต่างจากแบบ toast rack ตรงที่มีการผสมผสานของแต่ละหมวดหมู่เหมือนส่วนผสมของเค้ก โดยส่วนผสมของเค้กนั้นก็ประกอบไปด้วย competency ต่าง ๆ และต้องมีการผสมผสานเพื่อประเมินผลลัพธ์รวมให้ได้มาตรฐานตาม outcome ของแต่ละหลักสูตร เพื่อให้มั่นใจ

ว่าได้ผลลัพธ์ที่ตามแต่ละหลักสูตรได้วางไว้ (learning outcome = cake)

4. Spinal column: การจัด portfolio รูปแบบนี้เปรียบเทียบ competency ต่าง ๆ เรียงตัวเหมือนกระดูกสันหลัง (vertebrae) เมื่อมีหลาย ๆ competency ก็เรียงต่อกันจนเกิด central column ของเป้าหมายเป็นคุณลักษณะต่าง ๆ ที่นักศึกษาต้องมี และการประเมินว่ามีคุณสมบัติตาม competency ต่าง ๆ นั้นเปรียบเสมือน nerve root ที่เข้ามาสู่กระดูกสันหลัง เมื่อมีครบทุก competency จึงประกอบกันเป็น spine การประเมินแต่ละชนิดนั้นอาจมีความซ้ำซ้อนกันหรือแตกต่างกันไปเลยก็ได้ ขึ้นกับการเชื่อมโยงโดยnerve root ไปสู่แต่ละ competency

สิ่งที่ต้องคำนึงถึงเมื่อจะใช้ portfolio ในหลักสูตร

● การจะเลือกใช้ portfolio รูปแบบใด และประกอบไปด้วยอะไรบ้าง ขึ้นกับ

1. ชนิดและระดับของผู้เรียนว่าเป็นระดับใดต้องเก็บสะสมเป็นเวลานานแค่ไหน

2. เป้าหมายของ portfolio ว่าใช้เพื่อการเรียนการสอนหรือการประเมิน

3. โครงสร้างของหลักสูตร ว่าสามารถจะสอดแทรกการใช้ portfolio เข้าไปได้ในช่วงไหน และมีระยะเวลาการ monitor โดยอาจารย์ที่ปรึกษาได้ถี่แค่ไหน

4. วิธีการใช้ว่าจะใช้แบบกระดาษ electronic หรือผสมผสานทั้งสองวิธี

ปัญหาที่พบได้บ่อยของการใช้ portfolio

● เนื่องจากเป็นกระบวนการที่ใช้อาจารย์ผู้ประเมินหลายคน ความน่าเชื่อถือของผู้ประเมินจึงเป็นปัจจัยที่อาจส่งผลให้วิธีนี้ลดความน่าเชื่อถือลงได้ การฝึกฝนซักซ้อมทำความเข้าใจกับอาจารย์ที่ปรึกษาผู้ประเมิน สามารถลดความแตกต่างของการประเมินลงได้

ตัวอย่างแสดงองค์ประกอบของ portfolio ในหลักสูตรแพทย์ประจำบ้านวิสัญญีวิทยา

Evaluation	**การนำเสนอผลงาน**
การประเมินการปฏิบัติงานจากหน่วยงานต่างๆ	การนำเสนอ journal club
การประเมินรอบด้าน	การนำเสนอ case presentation
การประเมินจากเพื่อนร่วมงาน	การนำเสนอผลงานวิจัยในงานประชุมต่าง ๆ
การประเมินตนเอง	**Project ต่าง ๆ**
หัตถการ	CSR project
แบบประเมินการทำหัตถการต่าง ๆ	QA project
ใบประกาศ	Individual learning plan
ใบประกาศผ่านการอบรมช่วยคืนชีพ	Reflective note:
ใบประกาศผ่านการอบรมช่วยคืนชีพเด็กและทารก	Elective นอกหน่วยงาน
คะแนน	กรณีศึกษา
คะแนนสอบระหว่างภาค	เหตุการณ์ในภาวะวิกฤต
คะแนนสอบปลายภาคแต่ละชั้นปี	Ethical challenge
	อื่น ๆ
	โดยแพทย์ประจำบ้านเลือกมาใส่แฟ้มสะสมงานเอง

● ทัศนคติของนักศึกษาต่อการใช้ portfolio เนื่องจากเป็นวิธีที่ต้องอาศัยความร่วมมือและความเอาใจใส่ของนักศึกษาจึงจะได้ผลดี หากนักศึกษามีทัศนคติที่ไม่เห็นความสำคัญของการใช้ portfolio ก็อาจทำให้ประสิทธิภาพลดลงได้ ซึ่งปัญหานี้แก้ไขได้โดยการปรับทัศนคติ รณรงค์ให้เห็นความสำคัญ และการสร้างบรรยากาศที่ส่งเสริมการใช้ portfolio

เอกสารอ้างอิง

1. Challis M. AMEE Medical Education Guide No. 11 (revised): Portfolio-based learning and assessment in medical education. Medical Teacher 1999;21:370-86.
2. David MFB, Davis MH, Harden RM, Howie PW, Ker J, Pippard MJ. AMEE Medical Education Guide No. 24: Portfolios as a method of student assessment. Medical Teacher 2001;23:535-51.
3. Dent J, Harden RM. A practical guide for medical teachers. Elsevier Health Sciences, 2013.
4. Van Tartwijk J, Driessen EW. Portfolios for assessment and learning: AMEE Guide no. 45. Medical Teacher 2009;31:790-801.

รศ. ดร.นพ.เชิดศักดิ์ ไอรมณีรัตน์

หัวข้อ : Clinical performance ratings

Clinical Performance Ratings

เชิดศักดิ์ ไอรมณีรัตน์
ภาควิชาศัลยศาสตร์ คณะแพทยศาสตร์ศิริราชพยาบาล
มหาวิทยาลัย มหิดล

1

Competence and Performance

- Competence = The capacity of a person to perform a defined task (Maximal ability)
- Performance = The actual act in carrying out or execute the duty (Typical ability)

Clinical Performance Ratings

Ratings of clinical performance based on observing real-life clinical practice by attending faculty members

3

Objectives

- เมื่อสิ้นสุดการอบรมแล้ว อาจารย์ผู้เข้าอบรมสามารถ
  - บอกหลักการพื้นฐานของการประเมิน performance ratings ได้
  - พัฒนาแบบประเมิน clinical performance ratings ที่มีคุณภาพดี ซึ่งนำไปสู่การประเมินที่ถูกต้อง และเที่ยงตรง

Outline

- Clinical performance ratings
  - Advantages and disadvantages
  - Improving the rating quality
    - Raters
    - Rating instrument

Clinical Performance Ratings

- Advantages
  - Typical performance assessment
  - Motivation for clinical learning
  - Inexpensive

Clinical Performance Ratings

- Disadvantages
  - Subjective ratings
  - Unstructured settings
  - Adequacy of observation
  - Low reliability

Rater Errors

- Construct-irrelevance variance in performance ratings that is associated with raters' behavior, not with the actual performance of ratees
- Valid use of clinical performance assessment requires monitoring and controlling of rater errors.

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. Journal of Applied Measurement, 4, 386–422.

8

Reducing Rater Errors

- Improving raters
- Improving a rating instrument

Improving Raters

1. Rater training
2. Rater monitoring
3. Rater feedback

Rating Instrument

- Item
- Scale

Instrument A

1. How much time do you spend on homework?
   A. 1 hour/day B. 2 hours/day
   C. 3 hours/day D. 4 hours/day
2. The amount of homework for this course was …
   A. too little B. reasonable C. too much

Writing Effective Items

- Remember your purpose
- Keep it simple
- Focused: include only one topic per item
- Start with easy-to-respond items
- Group items into sections, position these sections in a logical order

Characteristics of A Good Scale

1. Well-defined category
2. Appropriate number of categories
3. Proper handling of middle category
4. Ordered
5. Research-based

แบบประเมินการปฏิบัติงานของนักศึกษาแพทย์ปี 6
คณะแพทยศาสตร์ศิริราชพยาบาล

นศ.พ. รหัส
โรงพยาบาล ภาควิชา/แผนก
หอผู้ป่วย ช่วงเวลาปฏิบัติงาน

หัวข้อการประเมิน	%	ดีเด่น (10)	ดี (8-9)	ผ่าน (6-7)	ไม่ผ่าน (<6)	NA
1. ความรู้		มีความรู้พื้นฐานที่สำคัญอย่างดีและสามารถนำมาประยุกต์ใช้ในการดูแลผู้ป่วยเป็นอย่างดี	มีความรู้พื้นฐานที่สำคัญอย่างดีแต่ยังนำมาประยุกต์ใช้ในการดูแลผู้ป่วยไม่ดีนัก	มีความรู้พื้นฐานที่สำคัญแต่ไม่สามารถนำมาประยุกต์ใช้ในการดูแลผู้ป่วย	ขาดความรู้พื้นฐานที่สำคัญ	
2. ทักษะ						
2.1 การแก้ปัญหาทางคลินิก		รวบรวมข้อมูลปัญหาได้สมบูรณ์ คิดวิเคราะห์แก้ปัญหาได้ด้วยตนเองดี	รวบรวมข้อมูลปัญหาได้สมบูรณ์ คิดวิเคราะห์แก้ปัญหาได้ด้วยตนเอง	รวบรวมข้อมูลปัญหาได้สมบูรณ์ แต่ต้องชี้แนะการคิดวิเคราะห์แก้ปัญหา	การรวบรวมข้อมูลปัญหาและการคิดวิเคราะห์แก้ปัญหามีหลงรอง	
2.2 ความสามารถในการดูแลผู้ป่วยและการตัดสินใจ		เลือกการสืบค้นและการรักษาได้ถูกต้อง สามารถบอกเหตุผล และคำนึงถึงปัจจัยรอบด้าน	เลือกการสืบค้นและการรักษาได้ถูกต้อง สามารถบอกเหตุผล แต่ยังขาดการคำนึงถึงปัจจัยบางด้าน	เลือกการสืบค้นและการรักษาได้ถูกต้อง แต่ไม่สามารถบอกเหตุผลได้ชัดเจน	ไม่สามารถเลือกการสืบค้นและการรักษาได้ถูกต้อง	
2.3 การบันทึกเวชระเบียน		มีข้อมูลสำคัญครบถ้วน เป็นระเบียบ อ่านง่าย ลงลายมือชื่อและรหัส	มีข้อมูลสำคัญครบถ้วน แต่ ไม่เป็นระเบียบ อ่านยาก หรือ ไม่ลงลายมือชื่อ/รหัส	ขาดข้อมูลสำคัญบางอย่าง เช่น ประวัติส่วนตัวและสังคม ประวัติยา progress note, procedure/surgical note, etc.	ขาดข้อมูลที่สำคัญหลายอย่าง ไม่เขียน progress note	
2.4 การทำหัตถการ		ทำหัตถการที่สำคัญได้เองอย่างแคล่วคล่อง ขั้นตอนการทำถูกต้อง มีความชำนาญในการใช้เครื่องมือ และติดตามดูแลผู้ป่วยหลังทำหัตถการอย่างเหมาะสม	สามารถทำหัตถการที่สำคัญได้แต่ไม่แคล่วคล่องมาก ต้องการความช่วยเหลือในบางขั้นตอน มีการติดตามดูแลผู้ป่วยหลังทำหัตถการอย่างเหมาะสม	สามารถทำหัตถการที่สำคัญได้ แต่ต้องการความช่วยเหลือค่อนข้างมาก หรือขาดการติดตามดูแลผู้ป่วยหลังผ่าตัด	ไม่สามารถทำหัตถการที่สำคัญได้ แม้จะได้รับการชี้แนะแล้ว ไม่รู้ขั้นตอนการทำหัตถการ และ/หรือขาดทักษะพื้นฐานในการทำหัตถการ	
2.5 ทักษะการนำเสนอ		เป็นขั้นตอนดีมาก เข้าใจง่าย	เป็นขั้นตอน ทั้งเข้าใจ โดยอาจต้องถามเพิ่มเติมเล็กน้อย	ไม่เป็นขั้นตอน ต้องถามเพิ่มเติมค่อนข้างมากจึงจะเข้าใจ	สับสนมาก นักเรียนไม่มีความเข้าใจในเรื่องที่นำเสนอ	
2.6 การสื่อสารกับผู้ป่วย/ญาติ		ดีมาก ผู้ป่วยและญาติพึงพอใจมาก	ดี ผู้ป่วยและญาติเข้าใจโรคที่เป็น	ผู้ป่วยและญาติบางคนไม่เข้าใจโรค	ใช้ภาษาไม่เหมาะสม หรือ สร้างความสับสนให้แก่ผู้ป่วยและญาติ	
3. ความเป็นวิชาชีพแพทย์						
3.1 ความสามารถในการเรียนรู้ด้วยตนเอง		แสดงถึงความใฝ่รู้ ค้นคว้าเพิ่มเติมได้ดีด้วยตนเอง	แสดงถึงความใฝ่รู้ ค้นคว้าเพิ่มเติมได้โดยต้องชี้แนะวิธีการ	ต้องกระตุ้นและชี้แนะวิธีการจึงจะค้นคว้าเพิ่มเติม	ขาดความใฝ่รู้ แม้จะได้รับการกระตุ้นและชี้แนะ	
3.2 การวางตัวที่เหมาะสม		ตรงต่อเวลา บุคลิกภาพ ความประพฤติ การแต่งกายถูกระเบียบ ทุกกาลเทศะ	ตรงต่อเวลา บุคลิกภาพ ความประพฤติ การแต่งกายถูกระเบียบ เป็นส่วนใหญ่	ไม่ตรงต่อเวลา บุคลิกภาพ การแต่งกายถูกระเบียบเป็นส่วนใหญ่	มีพฤติกรรมที่ไม่เหมาะสม และไม่ปรับปรุงหลังจากได้รับคำตักเตือน	
3.3 ความรับผิดชอบ		รับผิดชอบดีมาก หรือได้รับคำชมในการดูแลผู้ป่วยและการอยู่เวรอย่างมาก	รับผิดชอบดีในการดูแลผู้ป่วยและการอยู่เวร	ไม่มีข้อร้องเรียนเรื่องความรับผิดชอบในการดูแลผู้ป่วยและการอยู่เวร	ไม่รับผิดชอบ หรือ มีข้อร้องเรียนในการดูแลผู้ป่วย และการอยู่เวร	
3.4 เจตคติและจริยธรรม		ดูแลผู้ป่วยทั้งร่างกายและจิตใจอย่างดี เคารพสิทธิของผู้ป่วย	ดูแลผู้ป่วยทั้งร่างกายและจิตใจ เคารพสิทธิของผู้ป่วย	การดูแลผู้ป่วยขาดมิติด้านจิตใจแต่ยังเคารพสิทธิของผู้ป่วย	การดูแลผู้ป่วยขาดมิติด้านจิตใจ และไม่เคารพสิทธิของผู้ป่วย	
3.5 มนุษยสัมพันธ์กับผู้ร่วมงาน		มีมนุษยสัมพันธ์ที่ดีมาก และการทำงานเป็นทีมดีมาก	มีมนุษยสัมพันธ์ที่ดี ทำงานร่วมกับผู้อื่นได้	ขาดมนุษยสัมพันธ์ หรือมีปัญหาในการทำงานร่วมกับผู้อื่น	มนุษยสัมพันธ์ไม่ดี และ ไม่สามารถทำงานร่วมกับผู้อื่นได้	
เวลาปฏิบัติงาน		ครบ	ป่วย.....วัน	ลา.....วัน	ขาด.....วัน	

ความคิดเห็นเพิ่มเติม

ผู้ประเมิน .....
(.....)
วันที่ .....
ตำแหน่ง ❒ หัวหน้าแผนก/หอผู้ป่วย ❒ อาจารย์อำนวยการ

หมายเหตุ กรุณาใส่คะแนนในช่องสีเหลี่ยมหน้าช่องที่ต้องการ (ไม่มีจุดทศนิยม); NA = ไม่สามารถประเมินได้; % = น้ำหนักของแต่ละหัวข้ออาจแตกต่างกันในแต่ละภาควิชา

Group Work

- ให้อาจารย์ออกแบบใบประเมิน performance ในบริบทใดก็ได้ที่อาจารย์มีส่วนเกี่ยวข้อง

1. Item: กำหนดหัวข้อที่อาจารย์จะประเมินทั้งหมดในแบบประเมิน
2. Scale: ให้เลือกหนึ่งหัวข้อและสร้าง scale สำหรับหัวข้อนั้น

(เวลา 10 นาที)

Key Points: Performance Ratings

- Remember what to observe
- Rate when you still remember the students
- Multiple ratings: multiple raters, time points
- Rate when you are in a stable emotional state
- Be consistent in your rating standards (within and across groups)
- Rate each item independently: avoid halo effect
- Use the full range of scores: avoid restriction of range

แบบประเมินการปฏิบัติงานของนักศึกษาแพทย์ปี 6
คณะแพทยศาสตร์ศิริราชพยาบาล

นศ.พ. รหัส กลุ่ม
ฝึกปฏิบัติงานที่ ภาควิชา/แผนก
หอผู้ป่วย ช่วงเวลาปฏิบัติงาน ถึง

หัวข้อการประเมิน	%	ดีเด่น (10)	ดี (8-9)	ผ่าน (6-7)	ไม่ผ่าน (<6)	หมายเหตุ
1. ความรู้		มีความรู้พื้นฐานที่สำคัญอย่างดีและสามารถนำมาประยุกต์ใช้ในการดูแลผู้ป่วยเป็นอย่างดี	มีความรู้พื้นฐานที่สำคัญอย่างดีแต่ยังนำมาประยุกต์ใช้ในการดูแลผู้ป่วยไม่ดีนัก	มีความรู้พื้นฐานที่สำคัญแต่ไม่สามารถนำมาประยุกต์ใช้ในการดูแลผู้ป่วย	ขาดความรู้พื้นฐานที่สำคัญ	
2. ทักษะ						
2.1 การแก้ปัญหาทางคลินิก		รวบรวมข้อมูลปัญหาได้สมบูรณ์ คิดวิเคราะห์แก้ปัญหาได้ด้วยตนเองดี	รวบรวมข้อมูลปัญหาได้สมบูรณ์ คิดวิเคราะห์แก้ปัญหาได้ด้วยตนเอง	รวบรวมข้อมูลปัญหาได้สมบูรณ์ แต่ต้องชี้แนะการคิดวิเคราะห์แก้ปัญหา	การรวบรวมข้อมูลปัญหาและการคิดวิเคราะห์แก้ปัญหาบกพร่อง	
2.2 ความสามารถในการดูแลผู้ป่วยและการตัดสินใจ		เลือกการสืบค้นและการรักษาได้ถูกต้อง สามารถบอกเหตุผล และคำนึงถึงปัจจัยรอบด้าน	เลือกการสืบค้นและการรักษาได้ถูกต้อง สามารถบอกเหตุผล แต่ยังขาดการคำนึงถึงปัจจัยบางด้าน	เลือกการสืบค้นและการรักษาได้ถูกต้อง แต่ไม่สามารถบอกเหตุผลได้ชัดเจน	ไม่สามารถเลือกการสืบค้นและการรักษาได้ถูกต้อง	
2.3 การบันทึกเวชระเบียน		มีข้อมูลสำคัญครบถ้วน เป็นระเบียบ อ่านง่าย ลงลายมือชื่อและรหัส	มีข้อมูลสำคัญครบถ้วน แต่ ไม่เป็นระเบียบ อ่านยาก หรือ ไม่ลงลายมือชื่อ/รหัส	ขาดข้อมูลสำคัญบางอย่าง เช่น ประวัติส่วนตัวและสังคม ประวัติยา progress note, procedure/surgical note, etc.	ขาดข้อมูลที่สำคัญหลายอย่าง ไม่เขียน progress note	
2.4 การทำหัตถการ		ทำหัตถการที่สำคัญได้เองอย่างแคล่วคล่อง ขั้นตอนการทำถูกต้องมีความชำนาญในการใช้เครื่องมือและติดตามดูแลผู้ป่วยหลังทำหัตถการอย่างเหมาะสม	สามารถทำหัตถการที่สำคัญได้แต่ไม่แคล่วคล่องมาก ต้องการความช่วยเหลือในบางขั้นตอน มีการติดตามดูแลผู้ป่วยหลังทำหัตถการอย่างเหมาะสม	สามารถทำหัตถการที่สำคัญได้ แต่ต้องการความช่วยเหลือค่อนข้างมาก หรือขาดการติดตามดูแลผู้ป่วยหลังผ่าตัด	ไม่สามารถทำหัตถการที่สำคัญได้ แม้จะได้รับการชี้แนะแล้ว ไม่รู้ขั้นตอนการทำหัตถการ และ/หรือขาดทักษะพื้นฐานในการทำหัตถการ	
2.5 ทักษะการนำเสนอ		เป็นขั้นตอนดีมาก เข้าใจง่าย	เป็นขั้นตอน ฟังเข้าใจ โดยอาจต้องถามเพิ่มเติมเล็กน้อย	ไม่เป็นขั้นตอน ต้องถามเพิ่มเติมค่อนข้างมากจึงจะเข้าใจ	สับสนมาก นักเรียนไม่มีความเข้าใจในเรื่องที่นำเสนอ	
2.6 การสื่อสารกับผู้ป่วย/ญาติ		ดีมาก ผู้ป่วยและญาติพึงพอใจมาก	ดี ผู้ป่วยและญาติเข้าใจโรคที่เป็น	ผู้ป่วยและญาติบางคนไม่เข้าใจโรค	ใช้วาจาไม่เหมาะสม หรือสร้างความสับสนให้แก่ผู้ป่วยและญาติ	
3. ความเป็นวิชาชีพแพทย์						
3.1 ความสามารถในการเรียนรู้ด้วยตนเอง		แสดงถึงความใฝ่รู้ ค้นคว้าเพิ่มเติมได้ด้วยตนเอง	แสดงถึงความใฝ่รู้ ค้นคว้าเพิ่มเติมได้โดยต้องชี้แนะวิธีการ	ต้องกระตุ้นและชี้แนะวิธีการจึงจะค้นคว้าเพิ่มเติม	ขาดความใฝ่รู้ แม้จะได้รับการกระตุ้นและชี้แนะ	
3.2 การวางตัวที่เหมาะสม		ตรงต่อเวลา บุคลิกภาพ ความประพฤติ การแต่งกายถูกระเบียบทุกกาลเทศะ	ตรงต่อเวลา บุคลิกภาพ ความประพฤติ การแต่งกายถูกระเบียบเป็นส่วนใหญ่	ไม่ตรงต่อเวลา บุคลิกภาพ การแต่งกายถูกระเบียบเป็นส่วนใหญ่	มีพฤติกรรมที่ไม่เหมาะสมและไม่ปรับปรุงหลังจากได้รับคำตักเตือน	
3.3 ความรับผิดชอบ		รับผิดชอบดีมาก หรือได้รับคำชมในการดูแลผู้ป่วยและการอยู่เวรอย่างมาก	รับผิดชอบดีในการดูแลผู้ป่วยและการอยู่เวร	ไม่มีข้อร้องเรียนเรื่องความรับผิดชอบในการดูแลผู้ป่วยและการอยู่เวร	ไม่รับผิดชอบ หรือมีข้อร้องเรียนในการดูแลผู้ป่วยและการอยู่เวร	
3.4 เจตคติและจริยธรรม		ดูแลผู้ป่วยทั้งร่างกายและจิตใจอย่างดี เคารพสิทธิของผู้ป่วย	ดูแลผู้ป่วยทั้งร่างกายและจิตใจ เคารพสิทธิของผู้ป่วย	การดูแลผู้ป่วยขาดมิติด้านจิตใจ แต่ยังเคารพสิทธิของผู้ป่วย	การดูแลผู้ป่วยขาดมิติด้านจิตใจและไม่เคารพสิทธิของผู้ป่วย	
3.5 มนุษยสัมพันธ์กับผู้ร่วมงาน		มีมนุษยสัมพันธ์ที่ดีมาก การทำงานเป็นทีมดีมาก	มีมนุษยสัมพันธ์ที่ดี ทำงานร่วมกับผู้อื่นได้	ขาดมนุษยสัมพันธ์ หรือมีปัญหาในการทำงานร่วมกับผู้อื่น	มนุษยสัมพันธ์ไม่ดี และ ไม่สามารถทำงานร่วมกับผู้อื่นได้	
เวลาปฏิบัติงาน		ครบ	ป่วย.....วัน	ลา.....วัน	ขาด.....วัน	

ความคิดเห็นเพิ่มเติม .....
.....
.....
.....
.....

ผู้ประเมิน .....
(.....)
วันที่ .....

<u>หมายเหตุ</u> กรุณาให้คะแนนในช่องสีเหลี่ยมหน้าช่องที่ต้องการ (ไม่มีจุดทศนิยม), NA = ไม่สามารถประเมินได้

การประเมินการปฏิบัติงานทางคลินิก (Clinical performance assessment)

เชิดศักดิ์ ไอรมณีรัตน์

Purposeful assessment drives instruction and affects learning.

Wisconsin's principles for teaching and learning

บทบาทหน้าที่ของอาจารย์แพทย์ระดับคลินิกนั้นนอกจากจะต้องทำการสอนแล้ว การประเมินผลการปฏิบัติงานของนักศึกษาแพทย์ หรือ แพทย์ประจำบ้านก็เป็นสิ่งที่อาจารย์ต้องทำควบคู่กันไปด้วย ในบทความนี้ผู้นิพนธ์จะได้นำเสนอหลักการ และแนวปฏิบัติเพื่อให้อาจารย์แพทย์สามารถทำการประเมินการปฏิบัติงานของนักศึกษาแพทย์หรือแพทย์ประจำบ้านได้อย่างถูกต้อง เที่ยงตรง และเป็นธรรม

คำจำกัดความ

การประเมินผล (Assessment) หมายถึงกระบวนการที่ใช้เพื่อบันทึกระดับของความรู้ ทักษะ และเจตคติของผู้เรียน ซึ่งมักบันทึกเป็นระดับคะแนนที่สามารถเปรียบเทียบกันได้ระหว่างผู้เรียน

วิธีการที่อาจารย์ใช้ทำการประเมินผู้เรียนในระดับชั้นคลินิกนั้น สามารถแบ่งออกเป็นสองกลุ่มการประเมินได้แก่

1. <u>การประเมินความสามารถในห้องสอบ (competence)</u> การประเมินในกลุ่มนี้เป็นการประเมินที่อาจารย์จัดขึ้นในสถานการณ์ที่มีการควบคุมตัวแปรที่อาจส่งผลต่อความสามารถของนักศึกษาโดยมุ่งหวังจะวัดระดับความรู้ความสามารถสูงสุดที่ผู้เรียนมี โดยไม่มีปัจจัยอื่นมารบกวน มักเป็นการจัดสอบในห้องสอบ โดยมีการแจ้งผู้สอบให้มีการเตรียมตัวมาสอบในหัวข้อที่กำหนด ในวันและเวลาที่กำหนด สิ่งที่วัดได้จัดเป็นระดับความสามารถสูงสุดที่ผู้สอบสามารถแสดงออกมาได้ การสอบส่วนใหญ่ในโรงเรียนแพทย์จะเป็นการประเมินผลในกลุ่มนี้ เช่น การสอบข้อสอบข้อเขียน (ปรนัย หรือ อัตนัย), การสอบปากเปล่า (oral examination), การสอบ Objective Structured Clinical Examination (OSCE) เป็นต้น

2. <u>การประเมินความสามารถในการปฏิบัติงานจริง (performance)</u> การประเมินในกลุ่มนี้เป็นการประเมินจากการสังเกตการปฏิบัติงานของผู้เรียนในสถานการณ์จริง ซึ่งระดับความรู้ ความสามารถที่ผู้สอบแสดงออกมาให้อาจารย์เห็นนั้นอาจมีปัจจัยรบกวนอื่นๆมาเกี่ยวข้องด้วย เช่น ระบบการทำงาน สภาพแวดล้อม สภาพความสัมพันธ์ระหว่างผู้สอบกับคนรอบข้าง สภาพจิตใจของผู้เข้าสอบ ฯลฯ สิ่งที่วัดได้นั้นอาจขาดความเป็นมาตรฐานเดียวกันระหว่างผู้สอบแต่ละคนไปบ้าง แต่สิ่งที่ประเมินได้จากการประเมินความสามารถในกลุ่มนี้น่าจะสอดคล้องกับระดับความรู้ ความสามารถที่นักศึกษาหรือแพทย์ประจำบ้านใช้ทำงานจริงในชีวิตประจำวันมากกว่า

ในบทความนี้ผู้นิพนธ์มุ่งประเด็นการอภิปรายไปที่การประเมินความสามารถในการปฏิบัติงานจริง (performance) เป็นหลัก เนื่องจากเป็นการประเมินที่อาจารย์แพทย์ทำควบคู่ไปกับการสอนรูปแบบต่างๆที่มีการกล่าวถึงในตำรานี้ การประเมินการปฏิบัติงานทางคลินิกที่มีใช้กันอย่างแพร่หลายในวงการแพทยศาสตรศึกษาในประเทศไทยคือ

การจัดทำแบบฟอร์มให้อาจารย์สังเกตการปฏิบัติงานของนักศึกษาหรือแพทย์ประจำบ้านในหลากหลายหัวข้อ ตลอดช่วงระยะเวลาที่อยู่ภายใต้การดูแลของอาจารย์ ซึ่งจะเป็นรูปแบบการประเมินผลที่บทความนี้กล่าวถึงเป็นหลัก

ข้อพิจารณาในการประเมินผล

โดยทั่วไปแล้วเมื่ออาจารย์วางแผนจะทำการประเมินผลการเรียนรู้ ของนักศึกษา มีปัจจัยที่ต้องพิจารณาอยู่ 7 ประการด้วยกัน ได้แก่

1. ความถูกต้อง (Validity)

ความถูกต้องของผลการประเมินหมายถึงระดับคะแนนที่ได้นั้นแสดงถึงระดับความรู้ ความสามารถของนักศึกษาที่อาจารย์ต้องการวัดผลจริงๆ กล่าวคือผู้ที่ได้คะแนนสูง แสดงถึงระดับความรู้ ความสามารถที่สูง ในทางกลับกันผู้ที่ได้คะแนนต่ำ คือผู้ที่มีระดับความรู้ ความสามารถที่ต่ำ หากมีปัจจัยอื่นใดที่มีผลรบกวนการแปลผลดังกล่าว (validity threats) ก็จะลดระดับความถูกต้องของผลการประเมินลง ตัวอย่างปัจจัยรบกวนความถูกต้องของการประเมินการปฏิบัติงาน เช่น ความแตกต่างในมาตรฐานการให้คะแนนของอาจารย์ ความแตกต่างกันของลักษณะผู้ป่วยที่นักศึกษาแต่ละคนดูแล เป็นต้น

2. ความเที่ยงตรง (Reliability)

ความเที่ยงตรงของคะแนนหมายถึงหากนำนักศึกษาคนเดิมที่มีระดับความรู้ ความสามารถเท่าเดิม มาทำการประเมินผลซ้ำ คะแนนที่ได้จะมีค่าใกล้เคียงกันไหม ผลการสอบที่มีความเที่ยงสูง คือผลการสอบที่เมื่อสอบซ้ำ คะแนนก็จะเท่าเดิมหรือใกล้เคียงเดิม โดยทั่วไปแล้วเรารายงานความเที่ยงของคะแนนสอบเป็นตัวเลขมีค่า 0 – 1 โดยค่าดัชนีความเที่ยงที่ใกล้ศูนย์บอกถึงคะแนนสอบไม่ค่อยเที่ยง แต่หากค่าดัชนีความเที่ยงใกล้หนึ่ง แสดงว่าคะแนนสอบมีความเที่ยงสูง การประเมินการปฏิบัติงานทางคลินิกโดยทั่วไปจัดเป็นการประเมินผลที่มีความสำคัญปานกลาง มักต้องการระดับความเที่ยงตั้งแต่ 0.8 ขึ้นไป

3. ความเสมอภาค (Equivalence)

ความเสมอภาคของการประเมินผลหมายถึงผลการประเมินนักศึกษาในความรู้ หรือทักษะเดียวกันที่ทำในต่างวัน เวลา หรือสถานที่กัน สามารถนำมาเปรียบเทียบกันได้โดยไม่มีการได้เปรียบหรือเสียเปรียบกันเกิดขึ้น เช่นการสอบข้อเขียนวิชาเดียวกันของนักศึกษาที่ปฏิบัติงานกันคนละกลุ่ม สอบกันคนละวัน ก็ต้องมีมาตรการในการควบคุมให้ข้อสอบดังกล่าวมีระดับความยากง่ายใกล้เคียงกัน ในการประเมินการปฏิบัติงานของนักศึกษาแพทย์ อาจารย์แพทย์ก็ควรวางระบบให้นักศึกษาที่ปฏิบัติงานคนละกลุ่มเกิดความมั่นใจได้ว่ามาตรฐานการประเมินมีความยุติธรรม ไม่มีกลุ่มใดได้เปรียบ

4. ความเป็นไปได้ (Feasibility)

อาจารย์ผู้วางแผนการประเมินจำต้องศึกษาความเป็นไปได้ของการจัดประเมินผลด้วย ไม่ว่าจะเป็นในแง่เวลา สถานที่ งบประมาณ บุคลากร ฯลฯ เนื่องจากการพัฒนาการประเมินผลให้มีคุณภาพดีตามปัจจัยสามข้อแรก มักต้องการการลงทุน ลงแรงเพิ่มขึ้น แต่หากขาดงบประมาณ ก็จำต้องมีการลดหย่อนมาตรการต่างๆที่วางแผนไว้บ้าง เพื่อให้สามารถดำเนินการได้

5. ผลกระทบทางการศึกษา (Educational effect)

การประเมินผลที่ดีนั้นจะช่วยส่งเสริมให้ผู้เรียนกระตือรือร้นที่จะทำการศึกษา พัฒนาความรู้ และทักษะของตนเอง มีพฤติกรรมการเรียนรู้ที่เหมาะสม ตัวอย่างของการประเมินผลที่มีผลกระทบทางการศึกษาที่ไม่ดีนักเช่นการออกข้อสอบปรนัยที่เน้นการท่องจำเป็นการประเมินผลหลัก โดยไม่มีการประเมินรูปแบบอื่นมาเสริม ผลกระทบที่เกิดขึ้นก็คือนักศึกษาจะมุ่งเน้นท่องเนื้อหาในตำรา โดยไม่ใส่ใจการดูแลผู้ป่วยมากเท่าที่ควร ในทางตรงข้าม การประเมินการปฏิบัติงานบนหอผู้ป่วย เป็นสิ่งที่ช่วยส่งเสริมให้นักศึกษาสนใจผู้ป่วย ให้เวลากับผู้ป่วยมากขึ้น เป็นการส่งเสริมพฤติกรรมการเรียนรู้ที่ต้องการ

6. ผลเร่งการเรียนรู้ (Catalytic effect)

การประเมินผลที่ดีนั้นควรมีการนำเอาข้อมูลผลการประเมินนั้นมาให้ feedback ให้แก่ผู้เรียน เพื่อหวังผลให้ผู้เรียนนำไปพัฒนาปรับปรุงตัวให้มีความรู้ ความสามารถดีขึ้น ในการประเมินการปฏิบัติงานของนักศึกษา หรือแพทย์ประจำบ้านนั้น อาจารย์ได้มีโอกาสสังเกตผู้เรียนในหลายแง่มุมทั้งในด้านความรู้ ทักษะ และเจตคติ ข้อมูลที่อาจารย์ใช้เป็นพื้นฐานของการให้คะแนนในใบประเมินการปฏิบัติงานนับว่าเป็นข้อมูลที่เป็นประโยชน์ต่อตัวผู้เรียนเองด้วยซึ่งหากอาจารย์สามารถจัดหาเวลาพูดคุยกับตัวนักศึกษาหรือแพทย์ประจำบ้านผู้ได้รับการประเมินเพื่อให้ข้อมูลย้อนกลับ (feedback) ได้จะทำให้ได้ผลเร่งการเรียนรู้ด้วย

7. การยอมรับได้ของทุกฝ่ายที่เกี่ยวข้อง (Acceptability)

การประเมินผลที่ดีนั้นควรนำไปสู่ผลการประเมินที่เป็นที่ยอมรับได้ของทุกฝ่ายที่เกี่ยวข้อง ไม่ว่าจะเป็นนักศึกษาผู้สอบ อาจารย์ผู้ให้คะแนน เจ้าหน้าที่ เป็นต้น

ข้อดีและข้อจำกัดของการประเมินการปฏิบัติงานคลินิก

ใบประเมินการปฏิบัติงานคลินิกมีใช้กันอย่างแพร่หลาย และอาจารย์ผู้เกี่ยวข้องกับการสอนนักศึกษา หรือแพทย์ประจำบ้านในระดับคลินิก ต้องใช้เป็นประจำ เหตุที่อาจารย์ต้องทำการประเมินด้วยใบประเมินดังกล่าวเป็นเพราะการประเมินในรูปแบบนี้มีข้อดีอยู่หลายประการ อย่างไรก็ตามอาจารย์ก็ต้องตระหนักด้วยว่าการประเมินนี้ก็มีข้อจำกัดอยู่พอสมควร การทราบถึงข้อดี และ ข้อจำกัดของการประเมินผู้เรียนด้วยวิธีนี้น่าจะนำไปสู่การใช้ข้อมูลที่ได้มาจากแบบประเมินอย่างเหมาะสม

1. ข้อดี

การประเมินรูปแบบนี้มีข้อดีหลายประการ ได้แก่

I. ผลการประเมินสามารถสะท้อนระดับความรู้ ความสามารถของนักศึกษาที่ใช้ปฏิบัติงานในชีวิตจริง ซึ่งอาจแตกต่างไปจากผลการประเมินในห้องสอบ

II. ส่งเสริมให้นักศึกษาสนใจการเรียนรู้บนหอผู้ป่วย

III. เป็นการประเมินผลที่ราคาถูก ไม่ต้องมีการจัดสอบ ไม่ต้องมีการเตรียมอุปกรณ์พิเศษใดๆ เพียงแค่สังเกตการปฏิบัติงาน แล้วบันทึกคะแนน

2. ข้อจำกัด

การประเมินรูปแบบนี้มีข้อจำกัดอยู่หลายประการ

i. คะแนนที่ให้อาศัยการตัดสินด้วยดุลยพินิจของอาจารย์ซึ่งอาจมีมาตรฐานในการให้คะแนนแตกต่างกัน

ii. อาจารย์ผู้ให้คะแนนอาจมีโอกาสสังเกตพฤติกรรมของนักศึกษาหรือแพทย์ประจำบ้านไม่มากพอ

iii. สภาพแวดล้อมต่างๆ รวมทั้งผู้ป่วยที่ดูแล มีความแตกต่างกัน นักศึกษาหรือแพทย์ประจำบ้านบางคนอาจถูกประเมินในบริบทที่การดูแลรักษาผู้ป่วยทำได้อย่างมีประสิทธิภาพ ในขณะที่คนอื่นอาจถูกประเมินในบริบทที่การทำงานยุ่งยากซับซ้อนกว่า เปรียบเสมือนทำข้อสอบที่มีความยากง่ายต่างกัน

iv. ความเที่ยงของคะแนนที่ได้มักค่อนข้างต่ำ

ความคลาดเคลื่อนของคะแนนอันเนื่องมาจากผู้ให้คะแนน (Rater errors)

ปัญหาที่สำคัญของการให้คะแนนแบบประเมินการปฏิบัติงานทางคลินิกคือความคลาดเคลื่อนของคะแนนอันเนื่องมาจากอาจารย์ผู้ให้คะแนน กล่าวคืออาจารย์สองท่านสังเกตพฤติกรรมการปฏิบัติงานของผู้เรียนคนเดียวกัน อาจารย์อาจให้คะแนนแตกต่างกันได้ ลักษณะความคลาดเคลื่อนของคะแนนนี้มีจากหลายสาเหตุ เช่น ความแตกต่างกันของมาตรฐานในการให้คะแนน (leniency or severity error), การใช้มาตรฐานที่ไม่สม่ำเสมอ มีการเปลี่ยนแนวทางในการให้คะแนนตามอารมณ์ (rater inconsistency), การใช้แบบประเมินที่ไม่ถูกวิธี โดยอาจารย์ใช้ผลการตัดสินคะแนนในข้อหนึ่งเป็นตัวกำหนดคะแนนของข้ออื่นๆ (halo effect), การที่อาจารย์บางท่านจำกัดช่วงของคะแนนที่ให้ในแบบประเมิน (restriction of range) เป็นต้น ซึ่งความคลาดเคลื่อนของคะแนนเหล่านี้ส่งผลให้เกิดความไม่เป็นธรรมในการประเมิน และทำให้คะแนนมีความเที่ยงต่ำ การใช้แบบประเมินการปฏิบัติงานทางคลินิกจึงต้องมีมาตรการในการควบคุมความคลาดเคลื่อนของคะแนนจากเหตุเหล่านี้ควบคู่ไปด้วย

โดยทั่วไปแล้วเราสามารถลดความคลาดเคลื่อนของคะแนนได้ด้วยสองมาตรการใหญ่ๆ ได้แก่ (1) การพัฒนาอาจารย์ผู้ให้คะแนน และ (2) การพัฒนาแบบประเมิน

1. การพัฒนาอาจารย์ผู้ให้คะแนน

สาเหตุสำคัญประการหนึ่งของความคลาดเคลื่อนของคะแนนคืออาจารย์ผู้ให้คะแนนมีความเข้าใจเกณฑ์การให้คะแนนแตกต่างไปจากผู้พัฒนาแบบประเมิน การจัดให้มีการชี้แจงวิธีการใช้แบบประเมินให้อาจารย์ผู้เกี่ยวข้องทราบ รวมทั้งเปิดโอกาสให้อาจารย์ได้ทดลองใช้แบบประเมินแล้วอภิปรายแลกเปลี่ยนความเห็นกันจะทำให้อาจารย์ผู้เกี่ยวข้องกับการประเมินนี้มีความเข้าใจที่ตรงกันมากขึ้น หลังจากที่มีการชี้แจงแล้ว ก็ควรให้มีการตรวจสอบคะแนนที่ได้จากใบประเมินของอาจารย์แต่ละท่านว่ามีอาจารย์ท่านใดที่น่าจะใช้เกณฑ์การประเมินที่แตกต่างจากอาจารย์ท่านอื่นบ้าง หาก

พบว่ามีผลการประเมินของอาจารย์ท่านใดท่านหนึ่งที่มีความคลาดเคลื่อนของคะแนนมาก การให้ข้อมูลย้อนกลับ (feedback) แก่อาจารย์ท่านนั้นเพื่อให้เกิดการปรับเปลี่ยนแนวทางในการให้คะแนนก็จะช่วยให้ความคลาดเคลื่อนของคะแนนมีน้อยลงเรื่อยๆ

2. การพัฒนาแบบประเมิน

การสร้างแบบประเมินที่ดีนั้นควรปฏิบัติตามหลักการพื้นฐานต่างๆดังต่อไปนี้

2.1 เริ่มต้นสร้างแบบประเมินโดยมีความชัดเจนในวัตถุประสงค์ว่าต้องการประเมินความรู้ ทักษะ หรือเจตคติในด้านใดบ้าง ควรทำการค้นคว้าเพิ่มเติมว่ามีผู้รู้ท่านอื่นได้สร้างเครื่องมือเพื่อประเมินสิ่งเดียวกันนี้มาก่อนหรือไม่ มีองค์กรวิชาชีพ หรือสถาบันฝึกอบรมอื่นที่ได้พัฒนาแบบประเมินในเรื่องที่คล้ายคลึงกันมาก่อนหรือไม่ การได้ข้อมูลเพิ่มเติมเหล่านี้จะทำให้หัวข้อต่างๆที่จะทำการประเมินครบถ้วน

2.2 ข้อความในแต่ละข้อเขียนด้วยภาษาที่อ่านเข้าใจง่าย สั้น และ กระชับ ควรให้อาจารย์ท่านอื่น หรือ นักศึกษาช่วยอ่านและแสดงความเห็นว่ามีส่วนใดของแบบประเมินที่อ่านไม่เข้าใจบ้าง และทำการปรับแก้ตามความเหมาะสม

2.3 ในแต่ละข้อให้ทำการประเมินความรู้ หรือทักษะ หรือเจตคติ เพียงด้านใดด้านหนึ่งเท่านั้น

2.4 พยายามจัดกลุ่มหัวข้อที่ทำการประเมินให้ประเด็นที่มีความคล้ายคลึงกันอยู่ข้อใกล้ๆกัน จะทำให้อาจารย์ผู้ประเมินทำการกรอกใบให้คะแนนได้สะดวกกว่า

2.5 ตัวเลือกระดับคะแนน สามารถสร้างได้หลายรูปแบบ แต่รูปแบบที่สามารถลดความคลาดเคลื่อนของคะแนนได้มากที่สุดคือ behavioral-anchored rating scale (BARS) ซึ่งมีการแบ่งคะแนนที่จะให้เป็นระดับจากน้อยไปมาก โดยในแต่ละระดับคะแนนนั้นมีการเขียนบรรยายลักษณะพฤติกรรมของผู้ถูกประเมินอย่างชัดเจนว่าต้องมีพฤติกรรมอย่างไร จึงจะเหมาะสมกับการได้คะแนนในระดับดังกล่าว

2.6 ควรจำกัดระดับของคะแนนที่อาจารย์ผู้ประเมินสามารถให้ได้ อย่าให้มีจำนวนระดับมากจนเกินไป โดยทั่วไปแล้วระดับคะแนนที่อาจารย์สามารถแยกแยะความรู้ ความสามารถของผู้เรียนได้ควรอยู่ในช่วง 3 – 6 ระดับ การมีจำนวนระดับที่มากเกินไปมักสร้างความลำบากใจแก่อาจารย์ผู้ประเมินว่าแยกคะแนนระดับที่ใกล้เคียงกันได้ยาก

2.7 หากจะจัดให้มีระดับคะแนนที่อยู่กึ่งกลาง (เช่น มี 5 ระดับคะแนน จาก 1 – 5 ระดับคะแนนกึ่งกลางคือ 3) ต้องระมัดระวังว่าบางครั้งอาจารย์ผู้ให้คะแนนอาจให้คะแนนกึ่งกลางดังกล่าวโดยที่นักศึกษาหรือแพทย์ประจำบ้านไม่ได้มีระดับความรู้ ความสามารถอยู่ที่ระดับกึ่งกลางจริง แต่อาจารย์ให้คะแนนดังกล่าวด้วยเหตุผลอื่น เช่น ไม่ทันได้สังเกตพฤติกรรมดังกล่าว ไม่แน่ใจ ไม่มีโอกาสให้ผู้เรียนได้แสดงความรู้ หรือทักษะในด้านดังกล่าว ฯลฯ วิธีการแก้ปัญหาคือการจัดทำช่องประเมินขึ้นมาอีกช่องหนึ่งที่ชื่อ “ไม่สามารถประเมินได้” ขึ้นมาเพื่อให้อาจารย์ที่ไม่สามารถประเมินความรู้ หรือทักษะของผู้เรียนในเรื่องดังกล่าวได้ไม่ต้องเลือกระดับคะแนนกึ่งกลางโดยความจำใจ

ข้อแนะนำในการประเมินการปฏิบัติงาน

เพื่อให้การประเมินความสามารถจากการปฏิบัติงานจริงเป็นไปอย่างถูกต้อง ได้ผลการประเมินที่เที่ยงตรงและเป็นธรรม ตรงตามหลักการต่างๆที่กล่าวข้างต้น ผู้นิพนธ์มีข้อแนะนำดังต่อไปนี้

1. ให้อาจารย์ผู้ประเมินทุกท่านศึกษาแบบประเมินก่อนเริ่มสังเกตพฤติกรรมการทำงานของนักศึกษาหรือแพทย์ประจำบ้าน ให้อาจารย์จดจำให้ได้ก่อนว่ามีหัวข้อใดต้องทำการประเมินบ้าง เนื่องจากการประเมินในบางหัวข้ออาจารย์ต้องกำหนดบทบาท หรือสร้างสถานการณ์ให้ผู้เรียนแสดงความรู้ ความสามารถออกมา จึงจะประเมินได้ เช่น หากแบบประเมินกำหนดให้ประเมินความสามารถในการนำเสนอประวัติผู้ป่วย อาจารย์ก็ต้องจัดสถานการณ์ในการทำงานให้นักศึกษาที่จะถูกประเมินได้นำเสนอประวัติผู้ป่วย เป็นต้น
2. ให้อาจารย์ทำการให้คะแนนในใบประเมินในขณะที่ยังจดจำนักศึกษาหรือแพทย์ประจำบ้านผู้ถูกประเมินได้ โดยทั่วไปแล้วคือวันสุดท้ายของการปฏิบัติงานของนักศึกษาหรือแพทย์ประจำบ้าน เนื่องจากในปัจจุบันมีนักศึกษาและแพทย์ประจำบ้านหมุนเวียนปฏิบัติงานในหอผู้ป่วย หรือ หน่วยงานต่างๆจำนวนมาก มีโอกาสที่อาจารย์จะลืมว่านักศึกษาหรือแพทย์ประจำบ้านแต่ละคนนั้นมีระดับความรู้ ความสามารถเป็นอย่างไรเมื่อทิ้งเวลาให้เนิ่นนานออกไป ดังนั้นอาจารย์ควรจัดเป็นกิจวัตรในการทำงานที่ทุกๆช่วงที่มีการหมุนเวียนนักศึกษาหรือแพทย์ประจำบ้าน ต้องทำการกรอกใบประเมินทันที อย่าปล่อยจนถึงเวลาที่มีเจ้าหน้าที่มาตามแล้วซึ่งอาจเป็นเวลา 2 – 3 เดือนผ่านไปแล้ว
3. ในการปฏิบัติงานแต่ละช่วงเวลาของนักศึกษาหรือแพทย์ประจำบ้าน ให้อาจารย์จัดให้นักศึกษาได้รับการสังเกตและประเมินโดยอาจารย์หลายท่าน ในหลายบริบท และหลายครั้ง ยิ่งมีการประเมินมากครั้ง ในมากบริบท และมากผู้ประเมิน ผลการประเมินที่ได้มาจะช่วยยืนยันกันเองได้ ทำให้ความน่าเชื่อถือของคะแนนมีมากขึ้น หากระยะเวลาปฏิบัติงาน 4 สัปดาห์ของนักศึกษามีอาจารย์เพียงท่านเดียวทำการประเมินหนึ่งครั้งแล้วนักศึกษาได้คะแนนต่ำ อาจมีการร้องเรียนว่าเป็นเพราะอาจารย์ท่านที่ประเมินไปสังเกตพฤติกรรมเขาในวันที่เขามีปัญหาขึ้นมาพอดี และไม่ใช่พฤติกรรมปกติที่เขาทำในวันอื่นๆ แต่หากมีอาจารย์หลายท่าน ประเมินหลายครั้ง และทุกครั้งนั้นก็ได้ผลคะแนนที่ต่ำเช่นเดียวกันหมด ความน่าเชื่อถือของผลประเมินก็มากขึ้นว่านักศึกษาคนดังกล่าวมีระดับความรู้ ความสามารถหรือเจตคติที่ไม่ดีจริงๆ
4. ให้อาจารย์บันทึกระดับคะแนนในใบประเมินในช่วงเวลาที่อาจารย์ไม่อยู่ในสภาวะอารมณ์หงุดหงิด หิว หรือเหนื่อยล้า เนื่องจากอารมณ์ที่ปรวนแปรมีผลต่อการตัดสินใจของผู้ประเมินได้ ดังนั้นหากอาจารย์ตระหนักดีว่ากำลังไม่พอใจนักศึกษาหรือแพทย์ประจำบ้านคนใดซึ่งได้มีพฤติกรรมไม่เหมาะสมในการทำงาน ขอให้อาจารย์ชะลอการบันทึกคะแนนไว้ก่อน รอให้อารมณ์ และความรู้สึกของเรานั้นกลับสู่สภาวะปกติก่อน การตัดสินใจต่างๆในการให้คะแนนจะได้ทำได้อย่างปราศจากอคติ

5. ให้อาจารย์อ่านหัวข้อในใบประเมินและตัดสินคะแนนทีละข้อ เนื่องจากหัวข้อต่างๆที่อยู่บนแบบประเมินแต่ละหัวข้อนั้นจะได้รับการออกแบบให้วัดผลความรู้ ทักษะ หรือเจตคติที่แตกต่างกันไป จึงไม่มีความจำเป็นที่คะแนนในแต่ละหัวข้อต้องสอดคล้องกัน อาจารย์สามารถให้คะแนนความรู้สูง แต่ มนุษยสัมพันธ์กับเพื่อนร่วมงานต่ำก็ได้ ขอให้อาจารย์หลีกเลี่ยงวิธีการให้คะแนนแบบที่ทุกข้อได้คะแนนเท่ากันโดยไม่ได้พิจารณารายละเอียด
6. ให้อาจารย์ใช้มาตรฐานเดียวกันในการตัดสินคะแนนของนักศึกษาหรือแพทย์ประจำบ้านในทุกกลุ่มที่ปฏิบัติงาน พยายามอย่าให้ปัจจัยอื่นนอกเหนือไปจากเกณฑ์ที่ระบุไว้ในแบบประเมินมีอิทธิพลทำให้เกิดความยืดหยุ่นในเกณฑ์การพิจารณาคะแนน ไม่ว่าจะเป็นความสนิทสนมส่วนตัว หรือ ความสามารถในมิติอื่นๆนอกเหนือไปจากหัวข้อที่กำหนดในแบบประเมิน
7. ให้อาจารย์ตัดสินคะแนนโดยไม่จำกัดช่วงคะแนน แต่ให้ใช้เกณฑ์ประเมินเป็นหลัก หากนักศึกษามีระดับความรู้ความสามารถไม่ผ่านเกณฑ์ประเมิน ก็ควรประเมินคะแนนอยู่ในระดับไม่ผ่าน การประเมินผลตามจริงจะทำให้ได้คะแนนที่มีความเที่ยงสูง และแยกแยะระดับความรู้ ความสามารถของนักศึกษาได้ดีกว่าคะแนนที่มีค่าพอๆกันในนักศึกษาทุกคนไม่ว่าจะปฏิบัติงานดีหรือไม่ก็ตาม

เอกสารอ่านเพิ่มเติม

1. Amin Z, Eng KH. *Basics in medical education*. Singapore: World Scientific Publishing; 2003.
2. Myford CM, Wolfe EW. Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *J Appl Meas.* 2003;4:386 - 422.
3. Norcini J, Anderson B, Bollela V, et al. Criteria for good assessment: consensus statement and recommendations from the Ottawa 2010 Conference. *Med Teach.* 2011;33(3):206-214.
4. Rethans JJ, Norcini JJ, Baron-Maldonado M, et al. The relationship between competence and performance: implications for assessing practice performance. *Med Educ.* 2002;36(10):901-909.
5. Norcini J, Holmboe E. Work-based assessment. In: Cantillon P, Wood D, eds. *ABC of learning and teaching in medicine, 2nd ed.* Oxford: Wiley-Blackwell; 2010.
6. Norcini J. Workplace assessment. In: Swanwick T, ed. *Understanding medical education: Evidence, theory and practice*. Oxford: Wiley-Blackwell; 2010.
7. Turnbull J, Van Barneveld C. Assessment of clinical performance: In-training evaluation. In: Norman GR, van der Vleuten CPM, Newble DI, eds. *International handbook of research in medical education*. Dordrecht: Kluwer academic publishers; 2002.
8. Mavis B. Assessing student performance. In: Jeffries WB, Huggett KN, eds. *An introduction to medical teaching*. Dordrecht: Springer; 2010.

อ.นพ.ภูมิ์ ตริตระการ

CLINICAL TEACHING MADE EASY

Workplace-based assessment

Workplace-based assessment is now widespread throughout medicine. If carried out well, such assessments reconnect teaching and testing to the benefit of the learner. But workplace-based assessment brings a unique set of challenges to medical education and requires fresh thinking about how we consider and construct assessment programmes.

This article outlines some of the principles underpinning the design of workplace-based assessment and considers some of the tools that have been adopted for use within assessment programmes. The unique challenges of workplace-based assessment are considered, in particular the thorny issue of 'reliability'.

What is workplace-based assessment?

Workplace-based assessment refers to the assessment of what doctors actually do in practice and is predominantly carried out in the workplace itself. Workplace-based assessment in the training context relies on the use of tools for gathering information about aspects of trainees' work which are then used as vehicles for offering direct, timely and relevant feedback. The collection of workplace-based assessment data is learner-led and brought together, usually in a portfolio of evidence, to inform judgments about the trainee's overall progress.

So how does workplace-based assessment fit with traditional forms of testing in medicine?

Miller (1990) provides a useful pyramidal model (*Figure 1*) for mapping assessment methods currently available in medical education and illustrates how workplace-based assessment relates to the assessment of clinical competence.

'Knows' forms the base of Miller's pyramid, the entry point in the development of expertise. This tier is best assessed using simple knowledge tests such as multiple choice questions. The next tier up 'knows how' seeks to measure understanding or application of knowledge and is assessed using instruments such as unfolding patient management problems, extended matching or short essay questions. Higher up, objective structured clinical examinations assess at the 'shows how' level where students are required to demonstrate not only knowledge and understanding, but that they can bring together and manipulate relevant knowledge, skills and attitudes in a controlled situation.

The problem is that what doctors do in controlled assessment situations correlates poorly with their actual performance in professional practice (Rethans et al, 2002). Assessment of competence in a contextual vacuum is all very well but how can we know what happens in the messiness of real professional practice – what the doctor actually 'does'? This is where workplace-based assessment comes into its own.

Is it useful?

The utility, or usefulness, of an assessment has been defined as a product of its reliability, validity, cost-effectiveness, acceptability and educational impact (van der Vleuten, 1996). Utility can be applied to an entire assessment system or to an individual assessment method or component of the system. The concept is important in that no single element should be regarded as predominant. Assessment design then inevitably leads to a trade off between individual elements. Thus, traditional approaches to maximize the reliability or reproducibility of assessments can have a negative educational impact on the learner by reducing the opportunity for meaningful developmental feedback. Workplace-based assessments offer high educational impact but might not be as reliable as other highly structured tests such as multiple choice questions.

Historically, the seductiveness of standardized testing led medical education to rely on externally administered assessments delivered at the end of programmes of training. Workplace-based assessment offers an opportunity to re-evaluate this situation and reintegrate teaching, learning and assessment (*Figure 2*), in other words, providing assessment that is 'built in' and not 'bolt on'.

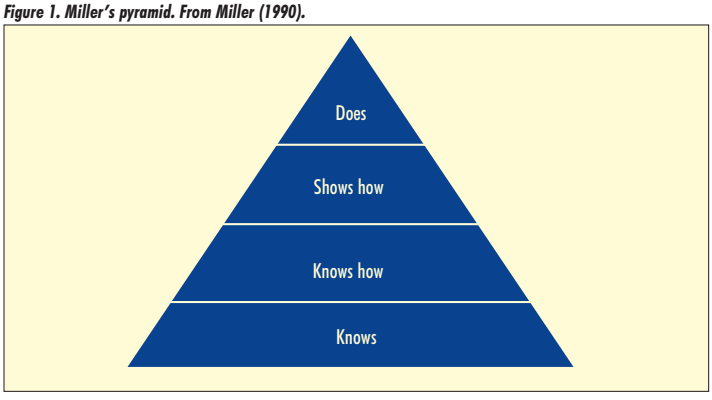
From methods to programmes

Traditional approaches to medical assessment have been founded on the notion that domains of competence (e.g. problem solving, communication skills) are stable and generic. It was considered possible to design tests that assessed these domains separately and reliably leading to a 'one trait, one instrument' approach (Schuwirth and van der Vleuten, 2004). However,

Dr Tim Swanwick *is Faculty Development Lead, London Deanery, London WC1B 5DN, Visiting Fellow, Institute of Education, London University, and Visiting Professor, University of Bedfordshire and* **Dr Nav Chana** *is Senior Lecturer in the Faculty of Medicine and Biomedical Sciences, St George's University of London, and Associate Director of General Practice, London Deanery*

Correspondence to: Dr T Swanwick

Figure 1. Miller's pyramid. From Miller (1990).



there has been a growing realization that competence is specific to particular clinical situations or contexts. In order to overcome this problem, it is vital to sample widely across both the content of the curriculum and the contexts in clinical care is delivered.

Given the complexity of assessing professional competence it is now recognized that assessment should be construed as a programme of activity requiring the acquisition of quantitative and qualitative information from different sources. As a major contribution to such programmes, assessing doctors in their actual working environment offers the opportunity to gather information using a variety of different tools, so building a 'rich picture' of their working practices.

Workplace-based assessments will not replace standardized assessments. There are issues in relation to reliability as a result of inconsistent application of tools by different raters or assessors. There is potential conflict in the role of the trainer who is supervising the learner, but also involved in the assessment process. And there are problems of attribution when routinely collected clinical practice data are assessed. So in order to gain the benefits while mitigating the risks, a number of key issues should be considered in the design and implementation of such assessment programmes.

What to assess?

The areas chosen to assess in workplace-based assessment are usually expressed as a series of competencies. These should be blueprinted against the curriculum and, in the way they are expressed, should encourage learner development. Let us look at those three issues in a little more detail:

Competency-based

Workplace-based assessment is usually competency-based. Despite criticisms of competency-based education as a whole (Talbot, 2004), concerns have usually been voiced where competencies are viewed as narrow, reductionist and overly simplistic. Competencies used for designing workplace-based assessments are best written as holistic statements which are framed as 'a complex structuring of attributes needed for intelligent performance in specific situations' (Gonczi, 1994).

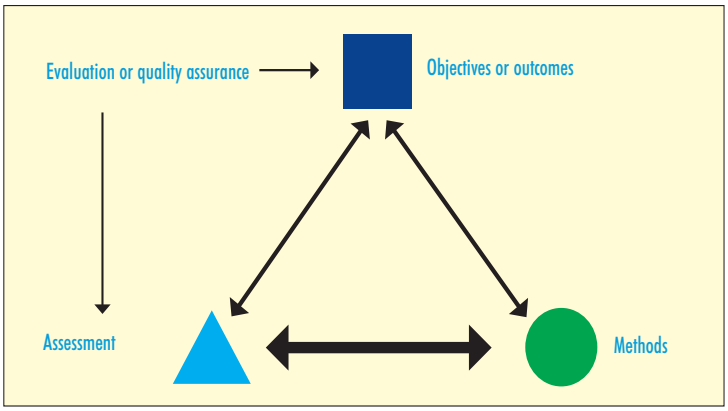


Figure 2. The educational paradigm: integrating teaching, learning and assessment.

CHECK ORIGINAL

Blueprinted

To ensure that assessments are integrated with the curriculum, competencies chosen for assessment should map directly onto the curriculum to ensure that there is both adequate coverage and widespread sampling. Some aspects of a curriculum will be more efficiently assessed through other means, clinical knowledge being an obvious case in point, however, some will be best assessed in the workplace. Indeed many aspects of professional performance such as team working, leadership and commitment to continuing professional development, are virtually impossible to assess in any other way.

Developmental

As already discussed, workplace-based assessment offers the opportunity to connect teaching, learning and assessment, and the developmental aspect of the assessment should therefore be a key feature. Developmental progressions in the literature, such as the novice to expert progression described by Dreyfus and Dreyfus (1986), may be helpful in constructing a developmental continuum of competence. Such a continuum has the advantage of explicitly illustrating the direction of travel for trainees, rather than merely pointing out the level below which they should not fall. This supports the concept of ongoing evidence collection throughout the training period, but with regular, well-circumscribed staging reviews at which the developmental framework is reviewed and the learner's progress through it judged.

So, workplace-based assessment provides useful formative and developmental feedback but it also has a summative role and informs judgments about overall progress. This raises the tension of potentially mixing formative and summative elements, but it is possible to address this through the careful design of the assessment system. Separating the interpretation of evidence from its elicitation is one way around the problem (Wiliam and Black, 1996). In other words, when it is assessment time, the learner needs to know, and be adequately prepared for it.

How much evidence is enough?

Collecting 'sufficient' evidence is essential in making a judgment about the attainment of competence. As we have seen, sampling widely across a number of clinical and contextual situations is important to overcome the problem of case specificity. In the assessment of 'work' there is no single method that will do it all and a variety of sources of information will be needed. This gives rise to the notion of a 'tool-box' of assessment methods.

In considering individual tools it is worth recognizing that, even unstandardized, they can be made sufficiently reliable, provided the tools are used sensibly and expertly, and enough sampling occurs (van der Vleuten and Schuwirth, 2005). But it is important to remember that the tools themselves only form a small part of an overall assessment programme and attention should focus on the utility of the entire programme of assessment, not just the individual tools themselves.

Confidence in the reproducibility of judgments made on the basis of work-

place-based assessment can be improved through triangulation. This involves using a range of different methods to collect evidence using multiple raters over a sustained period of time. Triangulation with other assessments external to the workplace is also important and an overarching assessment strategy for each training programme, in which workplace-based assessment is supported by other test methods – such as those of 'knowledge' and 'skills for clinical method', is essential.

Which methods?

The methods for used for providing feedback and gathering workplace evidence in current use tend to be variations on one of four themes; observations of clinical activities, discussion of clinical cases, analysis of performance data and multi-source feedback.

Observations of clinical activities

Traditionally, clinical skills have been assessed by the 'long case' presentation. The problem of case specificity using this technique, limiting the potential to sample widely, has given rise to the mini-clinical evaluation exercise or mini-CEX (Norcini et al, 1995). This tool has been developed to assess the clinical skills that trainees most often use in real patient encounters. It is based on assessment of multiple complete or partial clinical encounters observed by an educational supervisor or other clinician.

The direct observation of procedural skills (DOPS) is another widely used tool, and one of a number of similar instruments based around the assessment of real-life activities where the focus is on the skill with which the activity was performed. 'The consistent feature is that one or more assessors, who are trained in the assessment of that skill, make a judgment about a real life performance' (Postgraduate Medical Education and Training Board, 2007).

A raft of other observational tools encompassing a wide range of workplace activities are in also current use including the procedure-based assessment of the Intercollegiate Surgical Curriculum, the mini-imaging interpretation exercise of the Royal College of Radiologists and the assessment of teaching of the Royal College of Psychiatrists.

Discussion of clinical cases

The origin of the use of case-based discussion in UK training assessment systems stemmed from their use in the General Medical Council's performance procedures (Southgate et al, 2001) deriving originally from chart-stimulated recall oral assessments used in the USA and Canada. Case-based discussion is one of the evidence gathering tools used in workplace-based assessment in the UK foundation programme and is also being used in specialty training programmes such as in medicine, paediatrics and general practice.

Analysis of performance data

Norcini (2003) describes the basis for making a judgment on clinical performance data as having three potential sources; outcomes, process and volume. Outcomes of care, while being the most desirable measure, are limited by problems of attribution (to the individual), complexity, case mix and numbers. This is a particular problem in the assessment of trainee performance.

The process of care is more directly attributable to the individual doctor but effective processes do not necessarily mirror the best patient outcomes. The use of volumes of activity is premised on the basis that the more of a given activity that a doctor performs, the better their quality of care is likely to be. This basis for judgment is typified by the log books of the craft specialties such as surgery.

Multi-source feedback

The aim of using multi-source feedback to assess doctors in the workplace is to view a person's work from a variety of perspectives. In medical settings, physician colleagues (peers), co-workers and patients can be asked to complete surveys about the doctor. The person being assessed receives feedback based on his/her own aggregate ratings, usually along with average ratings of others being assessed at the same time. There is also a clear opportunity for comparing self-assessment data with those provided by raters.

Multi-source feedback tools can be subdivided into peer-rating tools, such as the mini-PAT (mini peer-rating assessment tool) used in foundation training, and patient satisfaction questionnaires, a significant number of which are in use in the UK (Chisholm and Askham, 2006).

Portfolios

Workplace-based assessments are usually collected within a structured portfolio. A portfolio comprises a dossier of evidence collected over time, which demonstrates a doctor's education and practice achievements (Wilkinson et al, 2002). There are many portfolio models (Webb et al, 2002) but in essence, if well constructed, a portfolio should chronicle the journey of a learner towards the attainment of professional expertise. A portfolio:

- Aims to serve as the reflective learning log of the learner, available to be shared with his/her educational supervisor
- Demonstrates the learner's progress towards covering the breadth and depth of the curriculum
- Acts as a repository for assessments
- Provides a framework for learning agreements between learners and teachers
- Charts a learner's progression and can help in making career choices and decisions.

The majority of portfolios used in medical education are web-based although with significant differences in structure and design between specialties and stage of training.

Quality assurance

Returning to the concept of utility, workplace-based assessment has huge strengths in the area of validity by virtue of its assessment of real or authentic material. Potentially it may have significant educational impact because of the reconnection of teaching and learning. Acceptability and cost-effectiveness are also potential winners but depend largely on how programmes are implemented. There are, however, significant issues with reliability as understood by traditional psychometric approaches. As Southgate et al (2001) point out, 'establishing the reliability of assessments of performance in the workplace is difficult because they rely on expert judgements of unstandardised material'.

In workplace-based assessment there are several specific threats to reliability:

- Inter-observer variation: the tendency for one observer to mark consistently higher or lower than another
- Intra-observer variation: variation in an observer's performance for no apparent reason (the 'good day/bad day' phenomenon)

- Case specificity: variation in the candidate's performance from one challenge to another, even when they seem to test the same attribute.

In the context of workplace-based assessment it is therefore helpful to reframe reliability as an attempt to maximize 'consistency and comparability'. Baker et al (1992) propose a number of activities that can help to do this, namely:

- Specification of standards, criteria, scoring guides
- Calibration of assessors and moderators
- Moderation of results, particularly those on the borderline
- Training of assessors, with retraining where necessary
- Verification and audit through the collection of assessment data.

It is clear, then, that the implementation of a successful workplace-based assessment programme will require training for assessors, arrangements for calibration, a procedure for the moderation of results and a raft of quality control checks. The more that teachers can be engaged in assessment, for example in selecting methodologies, generating standards and discussing criteria, the more the educational benefits of this powerful form of assessment can be realized.

Conclusions

Workplace-based assessment offers the opportunity to connect teaching, learning and assessment, provides a means for assessment of problematic areas that require evaluation of real performance in practice and is a useful component of an overall assessment programme. In order for its benefits to be realized there needs to be: clarity about what is being assessed through the identification of holistically described professional competencies; attention given to the developmental nature of the assessment; a variety of assessment tools used to gather evidence from multiple clinical contexts using multiple raters; and processes in place by which evidence can be collated, synthesized and judged at regular intervals by an educational supervisor to assess the learner's progress with consistency and comparability across assessment programmes maximized through a robust programme of quality assurance. BJHM

Conflict of interest: none.

Baker E, O'Neil H, Linn R (1992) Policy and validity prospects for performance-based assessment. *Am Psychol* **48**(12): 1210–18

Chisholm A, Askham J (2006) *What Do You Think of Your Doctor? A review of questionnaires for gathering patients' feedback about their doctor.* Picker Institute, Europe

Dreyfus H, Dreyfus S (1986) *Mind over machine. The Power of Human Intuition Expertise in the Era of the Computer.* Basil Blackwell, Oxford

Gonczi A (1994) Competency based assessment in the professions in Australia. *Assessment in Education* **1**(1): 27–44

Miller G (1990) The assessment of clinical skills/competence/performance. *Acad Med* **65**(Suppl): S63–7

Norcini J (2003) ABC of learning and teaching in medicine. Work based assessment. *BMJ* **326:** 753–5

Norcini J, Blank L, Arnold G, Kimball H (1995) The mini-CEX: a preliminary investigation. *Ann Intern Med* **125:** 795–9

Postgraduate Medical Education and Training Board (2007) *Developing and Maintaining an Assessment System – a guide to good practice.* Postgraduate Medical Education and Training Board, London

Rethans J, Norcini J, Baron-Maldonado M, Blackmore D, Jolly B, La Duca T (2002) The relationship between competence and performance: implications for assessing practice performance. *Med Educ* **36:** 901–9

Southgate L, Cox J, David T et al (2001) The assessment of poorly performing doctors: the development of the assessment programmes for the General Medical Council's Performance Procedures. *Med Educ* **35**(Suppl 1): 2–8

Schuwirth L, van der Vleuten C (2004) Changing education, changing assessment, changing research. *Med Educ* **38:** 805–12

Talbot M (2004) Monkey see, monkey do: a critique of the competency model in graduate medical education. *Med Educ* **38:** 1–7

van der Vleuten C (1996) The assessment of professional competence: developments, research and practical implications. *Advances in Health Science Education* **1:** 41–67

van der Vleuten C, Schuwirth L (2005) Assessing professional competence: from methods to programmes. *Med Educ* **39:** 309–17

Webb C, Gray M, Jasper M, Miller C, McMullan M, Scholes J (2002) Models of portfolios. *Med Educ* **36**(10): 897–8

Wilkinson TJ, Challis M, Hobma SO, Newble DI, Parboosingh JT, Sibbald JG, Wakeford R (2002) The use of portfolios for assessment of the competence and performance of doctors in practice. *Med Educ* **36:** 918–24

Wiliam D, Black P (1996) Meanings and consequences: a basis for distinguishing formative and summative functions of assessment? *Br Educ Res J* **22:** 537–48

KEY POINTS

- Workplace-based assessment is now widespread across all specialities and all stages of training.
- Workplace-based assessment offers the opportunity to connect teaching, learning and assessment.
- Workplace-based assessment has a dual function of offering focussed and timely feedback to trainees as well as providing data to support more long range judgments about trainee progress.
- Workplace-based assessment requires new ways of thinking about reliability based on maximizing consistency and comparability.

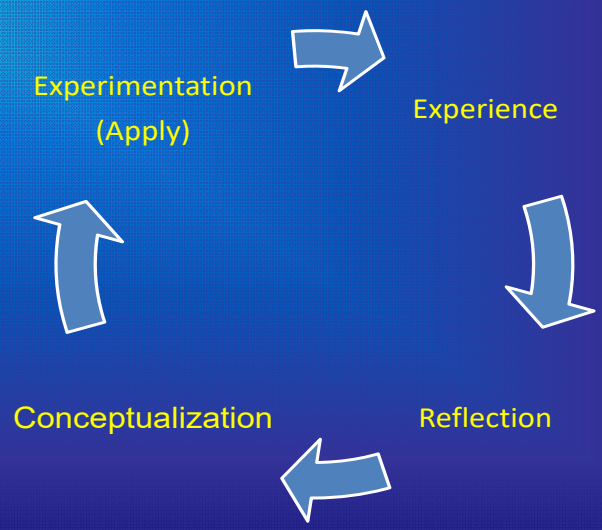
รศ. ดร.นพ.เชิดศักดิ์ ไอรมณีรัตน์

หัวข้อ :Summary

Summary

นพ. เชิดศักดิ์ ไอรมณีรัตน์
ภาควิชาศัลยศาสตร์ คณะแพทยศาสตร์ศิริราชพยาบาล
มหาวิทยาลัยมหิดล

Experiential Learning Theory



Kolb DA. Experiential learning. Englewood cliffs, NJ: Prentice-Hall, 1984.
Schön, D. The Reflective Practitioner, New York: Basic Books, 1983.

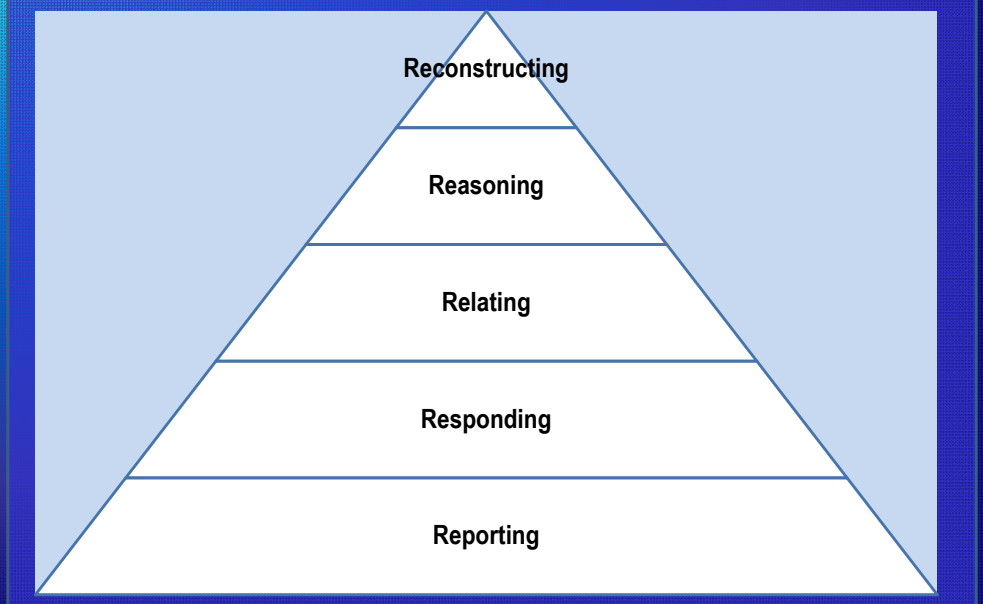
A complex and deliberate process of thinking about and interpreting experience in order to learn from it.

This is a conscious process which does not occur automatically, but is in response to experience and with a definite purpose.

Reflection is a highly personal process, and the outcome is a changed perspective, or learning.

Atkins and Murphy (1995)

Five Levels of Reflection



Bain JD, et al. Reflecting on practice: Student teachers' perspectives, Flaxton, 2002..

Examples

- Reporting: *วันนี้ได้เรียนเรื่อง...*
- Responding: *ฉันรู้สึกชอบแนวคิดเรื่อง...*
- Relating: *...ฉันมีปัญหาในการสร้างข้อสอบ* OSCE *คือ...*
- Reasoning: *เหตุที่ข้อสอบ* OSCE *ที่ฉันออกผู้ป่วยมักแสดงไม่สมบทบาทเป็นเพราะ... ฉันควรจะแก้ไขโดย...*
- Reconstructing: *ฉันออกแบบการประเมินผลวิธีใหม่โดย...*

Bain JD, et al. Reflecting on practice: Student teachers' perspectives, Flaxton, 2002..

Discussion

- ให้แต่ละกลุ่มอภิปรายว่าจากประสบการณ์การอบรม มีสิ่งใดที่สมาชิกในกลุ่มได้เรียนรู้ และสามารถนำไปใช้ประโยชน์ได้สูงสุด

Summary of the Workshop

- Thursday
  - MCQ
  - Constructed response exam
  - OSCE
- Friday
  - Long case exam
  - Portfolio
  - Clinical performance ratings
  - Workplace-based assessment

Questions & Comments

Cherdsak.ira@mahidol.ac.th


SHEE
Siriraj Health science Education Excellence center
Shee.si.mahidol.ac.th


True Success is not in the learning, but in its application to the benefit of mankind
HRH Prince Mahidol of Songkla

กระดาษบันทึก

Question & Comments

ศูนย์ความเป็นเลิศด้านการศึกษาวิทยาศาสตร์สุขภาพ (ศศว)
Siriraj Health science Education Excellence center (SHEE)

ฝ่ายการศึกษาก่อนปริญญา คณะแพทยศาสตร์ศิริราชพยาบาล
สำนักงาน: ตึกอดุลยเดชวิกรม ชั้น 6 (ห้อง 656)
Tel. 02 419 9978, 02 419 96637 Fax. 02 412 3901

shee.si.mahidol.ac.th
shee.mahidol@gmail.com
mahidol.shee
SHEE FC

SHEE
Siriraj Health science Education Excellence center