



มหาวิทยาลัยมหิดล
คณะแพทยศาสตร์
ศิริราชพยาบาล



เอกสารประกอบการอบรม

Assessment workshop for clinical teachers

สำหรับอาจารย์แพทย์โรงพยาบาลสมเด็จพระปิ่นเกล้า

15 – 17 กรกฎาคม 2563

รูปแบบ SHEE live training



<http://shee.si.mahidol.ac.th>

สารบัญ

	หน้า
กำหนดการ.....	1
รายชื่อผู้ร่วมอบรม (แบบแบ่งกลุ่ม)	2
เอกสารประกอบการอบรม	
15 กรกฎาคม 2563	3
หัวข้อ : What is good assessment?	4
หัวข้อ : How to choose assessment methods?	7
หัวข้อ : Validity and reliability	8
หัวข้อ : EPA: Entrustable professional activities	12
หัวข้อ : Standard setting	34
หัวข้อ : Grading	41
16 กรกฎาคม 2563	53
หัวข้อ : Multiple-choice questions item development	54
หัวข้อ : Multiple-choice questions item analysis	70
หัวข้อ : Constructed response item development	87
17 กรกฎาคม 2563	105
หัวข้อ : OSCE item development	106
หัวข้อ : Long case examination	109
หัวข้อ : Portfolio	112
หัวข้อ : Rating scale development	124
หัวข้อ : Workplace-based assessment	128

(ร่าง) กำหนดการอบรมเชิงปฏิบัติ เรื่อง Assessment workshop for clinical teachers
สำหรับอาจารย์แพทย์โรงพยาบาลสมเด็จพระปิ่นเกล้า
รุ่นที่ 1 ระหว่างวันที่ 15 - 17 กรกฎาคม พ.ศ. 2563 (รูปแบบ online)

Join Zoom Meeting

<https://thairen.zoom.us/j/65791717761?pwd=TmVMc3hZSXRsYzZuS3E2c2xYTnpqdz09>

Meeting ID: 657 9171 7761

Password: ass1517

วันที่ 15 กรกฎาคม 2563		
08.30 – 09.30 น.	What is good assessment?	รศ.ดร. นพ.เชิดศักดิ์ ไอรณณรัตน์ ผศ. นพ.สุประพัฒน์ สนใจพาณิชย์
09.30 – 10.00 น.	How to choose assessment methods?	
10.15 – 11.30 น.	Validity and reliability	
11.30 – 12.00 น.	EPA: Entrustable professional activities	
13.00 – 14.30 น.	Standard setting	
14.45 – 15.45 น.	Grading	
15.45 – 16.00 น.	Summary	
วันที่ 16 กรกฎาคม 2563		
08.30 – 10.00 น.	Multiple-choice questions item development	รศ.ดร. นพ.เชิดศักดิ์ ไอรณณรัตน์ ผศ. นพ.ทศ หาญรุ่งโรจน์
10.15 – 11.00 น.	Multiple-choice questions item review	
11.00 – 12.00 น.	Multiple-choice questions item analysis	
13.00 – 14.30 น.	Constructed response item development	
14.45 – 15.45 น.	Constructed response item review	
15.45 – 16.00 น.	Summary	
วันที่ 17 กรกฎาคม 2563		
08.30 – 10.00 น.	OSCE item development	รศ.ดร. นพ.เชิดศักดิ์ ไอรณณรัตน์ ผศ. นพ.ทศ หาญรุ่งโรจน์
10.15 – 11.30 น.	OSCE item review	
11.30 – 12.00 น.	Long case examination	
13.00 – 14.00 น.	Portfolio	
14.00 – 14.45 น.	Rating scale development	
15.00 – 15.45 น.	Workplace-based assessment	
15.45 – 16.00 น.	Summary	

หมายเหตุ: กำหนดการอาจมีการเปลี่ยนแปลงตามความเหมาะสม

หน้า 1/1

รายชื่อผู้ร่วมอบรม

รูปแบบ SHEE Live training ผ่าน Zoom meeting

กลุ่ม	ลำดับ	ยศ ชื่อ สกุล	หน่วยงาน/กลุ่มงาน
1	1	น.อ.อานัน นิมมวณ	กลุ่มงานศัลยกรรม
1	2	น.อ.นพดล เหนระกุล	กลุ่มงานศัลยกรรม
1	3	น.อ.อภิรัฐ แสงเพชรส่อง	กลุ่มงานศัลยกรรม
1	4	น.อ.อิทธิพล ประสิทธิ์ดำรง	กลุ่มงานศัลยกรรมกระดูก
1	5	น.อ.จักรพล จันทร์ประสิทธิ์	กลุ่มงานศัลยกรรมกระดูก
1	6	ร.ท.ศุภะโชค วัฒนกิจไกรเลิศ	กลุ่มงานศัลยกรรมกระดูก
1	7	น.อ.หญิง ศิริพรรณ โกมลประเสริฐ	งานบริหาร ศูนย์แพทยศาสตรศึกษา
1	8	น.ท.หญิง อภิวรรณิ แหวนทอง	งานวิชาการ ศูนย์แพทยศาสตรศึกษา
1	9	น.ต.หญิง ผุสดี ศิริวัฒนา	งานประกันคุณภาพฯ ศูนย์แพทยศาสตรศึกษา
1	10	น.อ.ปราโมทย์ กาญจนกิจสกุล	กลุ่มงานสูตินรีเวชกรรม
1	11	น.อ.จลนัยน์ ทิศาปราโมทย์กุล	กลุ่มงานสูตินรีเวชกรรม
1	12	น.อ.ชรินทร์ มิตินันท์วงศ์	กลุ่มงานสูตินรีเวชกรรม
1	13	พญ.จิตตรินทร์ ศิริระอัมพูช	กลุ่มงานสูตินรีเวชกรรม
1	14	พญ.นิธินันท์ บุญยสนธิกุล	กลุ่มงานสูตินรีเวชกรรม
1	15	น.อ.หญิง วรรษญา อุดมศักดิ์	กลุ่มงานกุมารเวชกรรม
1	16	น.ท.หญิง วิวรรณ สุจริต	กลุ่มงานกุมารเวชกรรม
1	17	ร.อ.หญิง กนกกาญจน์ นาคะสุวรรณ	กลุ่มงานกุมารเวชกรรม
1	18	น.อ.หญิง จริยา สันตติอนันต์	กลุ่มงานวิสัญญีกรรม
1	19	น.อ.หญิง ภิญญาดา เจริญผล	กลุ่มงานวิสัญญีกรรม
1	20	น.อ.หญิง สุมัทนา ณ บ่อมเพชร	กลุ่มงานวิสัญญีกรรม
1	21	น.ท.หญิง ปิยะภัทร สมานพิบูลย์ผล	กลุ่มงานวิสัญญีกรรม
1	22	น.ท.หญิง ชุตินาศ ไชยรักษ์	กลุ่มงานวิสัญญีกรรม
2	23	น.อ.พิทักษ์ พงศ์นนทชัย	ศูนย์หัวใจ
2	24	น.อ.บริพันธ์ สุวชิรัตน์	กลุ่มงานโสต ศอ นาสิกกรรม
2	25	น.ท.หญิง นงาณศ์ เกษโกวิท	กลุ่มงานจักษุกรรม
2	26	น.ท.หญิง พัชรพร หวังวรวิทย์	กลุ่มงานจักษุกรรม
2	27	น.อ.หญิง ภาวิกา ธรรมโน	ศูนย์แพทยศาสตรศึกษา/กลุ่มงานจักษุกรรม
2	28	น.ท.บัณฑิต นวนพรัตน์สกุล	กลุ่มงานรังสีวิทยา
2	29	น.อ.หญิง พรพิมล รัตนาวีวัฒน์พงศ์	กลุ่มงานเวชศาสตร์ฟื้นฟู
2	30	น.ท.หญิง อภิพร กาญจนบุญชร	กลุ่มงานเวชศาสตร์ฟื้นฟู
2	31	น.อ.อดิพงษ์ สุจิรัตน์	กลุ่มงานอายุรเวชกรรม
2	32	น.อ.จตุรงค์ ตันติมงคลสุข	กลุ่มงานอายุรเวชกรรม
2	33	น.อ.หญิง ชนกานาถ วัชรากกร	กลุ่มงานอายุรเวชกรรม
2	34	น.อ.หญิง ดิราภรณ์ บุญยรัตน์	กลุ่มงานอายุรเวชกรรม
2	35	น.อ.สรภพ ภัคดีวงศ์	กลุ่มงานอายุรเวชกรรม
2	36	น.ท.ธีรพล ปัญชัยพรพล	กลุ่มงานอายุรเวชกรรม
2	37	น.ท.หญิง ธนาวดี สิริธนต์พันธ์	กลุ่มงานอายุรเวชกรรม
2	38	น.ท.หญิง อรภัทรา คงประยูร	กลุ่มงานเวชศาสตร์ฉุกเฉิน
2	39	น.ท.หญิง ชัชชชา จรรย์ยานนท์	กลุ่มงานเวชศาสตร์ฉุกเฉิน
2	40	น.ต.หญิง ยุกานต์ ไพบูลย์วงษ์	กลุ่มงานเวชศาสตร์ฉุกเฉิน
2	41	ร.อ.กฤษฎา ชุมวณิชย์	กลุ่มงานเวชศาสตร์ฉุกเฉิน
2	42	นพ.สมประสงค์ เกียรติวัฒนชัย	กลุ่มงานเวชศาสตร์ฉุกเฉิน

เอกสารประกอบการอบรม



15 กรกฎาคม 2563

15 กรกฎาคม 2563

What is good assessment?

What is Good Assessment?

นพ. เชิดศักดิ์ โอรมเจริญรัตน์
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

Outline

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

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

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

2. Reliability

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

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

10

3. Equivalence

- การทดสอบหัวข้อเดียวกันกับนักศึกษาระดับชั้นเรียนเดียวกัน ที่ทดสอบกันต่างเวลา 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

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)

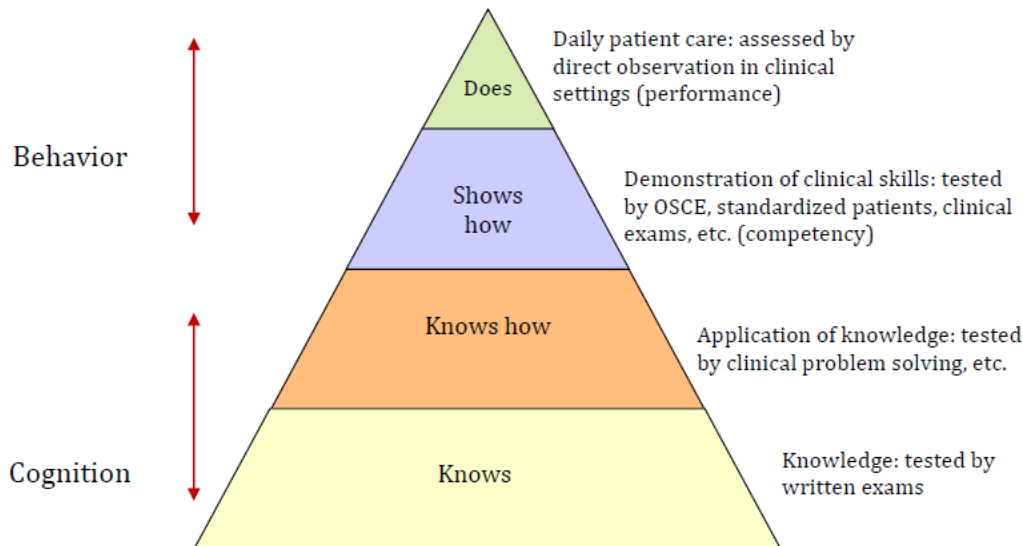
7. Effective assessment requires feedback to students that emphasizes strengths of performance and weaknesses to be corrected.
8. Effective assessment must be supported by a comprehensive grading and reporting system

15 กรกฎาคม 2563

How to choose assessment methods?

Miller's Pyramid of Assessment

Miller's Pyramid of Assessment provides a framework for assessing clinical competence in medical education and can assist clinical teachers in matching learning outcomes (clinical competencies) with expectations of what the learner should be able to do at any stage.



Adapted from: Ramani S, Leinster S, AMEE Guide no 34: Teaching in the clinical environment. Medical Teacher, 2008;30(4):347-364.

KNOWS forms the base of the pyramid and the foundation for building clinical competence.

Example: Learner is assessed his/her knowledge of the principles/content of basic science knowledge through a multiple choice exam/similar assessment tools.

KNOWS HOW uses knowledge in the acquisition, analysis, and interpretation of data.

Example: Learner knows how to, given a patient scenario, utilise the history and physical examination and diagnostic test data to identify the scientific basis of the patient's condition or initial management plan.

SHOWS HOW requires the learner to demonstrate the integration of knowledge and skills into successful clinical performance.

Example: Learner shows how to diagnose, develop and implement a treatment plan and effectively explain it to the patient and/or family.

DOES focuses on assessment of clinical performance in actual practice settings.

Example: Learner demonstrates the ability to evaluate the patient's condition and to revise the management plan as warranted and counsel the patient and/or family.

15 กรกฎาคม 2563

Validity and reliability

Validity and Reliability

เชิดศักดิ์ ไอรมนิรัตน์
ภาควิชาศัลยศาสตร์ คณะแพทยศาสตร์ศิริราชพยาบาล
มหาวิทยาลัย มหิดล

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

Historical Concepts

- Three types of validity
 - Content validity
 - Construct validity
 - Criterion validity

Contemporary Concepts

- AERA, APA, NCME. Standards for educational and psychological testing 2014.
- Assessments are not valid or invalid, rather assessment scores have more (or less) validity evidence to support the proposed interpretations.
- Validity requires multiple sources of evidence to support or refute meaningful score interpretation.

Validity of faculty ratings

5

Sources of Validity Evidence

- Content
- Response processes
- Internal structure
- Relationship to other variables
- Consequences

Validity of faculty ratings

6

Reliability

- Consistency of test scores
 - If we test the students/residents again, will they get the same scores?
- High values: highly consistent test 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

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

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

15

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

16

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
- $r = 2(0.67)/[1+(2-1)(0.67)] = 0.80$

True Score Theory

- Each student has a true score, a hypothetical value representing a score free of error.
- If we test a student repeatedly, the average of the obtained scores would approximate the true score, with a standard deviation of SEM.

SEM

$$SEM = SD\sqrt{1-r}$$

SD = standard deviation
r = internal consistency reliability

- ↑SD (more spread of score): higher SEM
- ↑r (more accurate measures): smaller SEM

What should we do with students with an SEM around cut score?

- False positive: Passing students who should have fail the examination
- False negative: Failing students who should have pass the examination

Reliability of Mastery Tests

- Consistency of decisions on two test forms

		Form B	
		Pass	Fail
Form A	Pass	a	b
	Fail	c	d

$$\% \text{ consistency} = 100 \times (a + d)/(a+b+c+d)$$

Performance Assessment

- Inter-rater agreement
 - Percentage of agreement between the two
 - Correlation between the two
 - Intraclass correlation

Summary

- Validity
 - Validity threats
 - Five sources of validity evidence
- Reliability
 - Reliability of standard written exam
 - Reliability of mastery tests
 - Reliability of performance assessment

15 กรกฎาคม 2563

EPA: Entrustable professional activities

Entrustable Professional Activity (EPA)

เชิดศักดิ์ ไอร่มเจริญรัตน์
ศูนย์ความเป็นเลิศด้านการศึกษาระดับสหสาขา
คณะแพทยศาสตร์ศิริราชพยาบาล

Assessment Approaches

Does
Shows how
Knows how
Knows

Miller's Pyramid

2 2

Assessment at "Does" level

- Does => Professional task

EPA

Outline

- EPA
 - Definitions: EPA, competencies, milestones
 - Key concepts
 - How to proceed with EPA?
 - Assessment in EPA framework

EPA

- Entrustable Professional Activity
 - A unit of professional practice, defined as tasks or responsibilities that trainees are entrusted to perform unsupervised once they have attained sufficient specific competence

AAMC. Core entrustable professional activities for entering residency: Faculty and learners' guide, Washington DC, 2014.

Competency

- Competency: An observable ability of a professional, integrating multiple components such as knowledge, skills, values, and attitudes

AAMC. Core entrustable professional activities for entering residency: Faculty and learners' guide, Washington DC, 2014.

Key Concepts

- EPAs are not an alternative for competencies, but a means to translate competencies into clinical practice.
- Competencies are descriptors of physicians.
- EPAs are descriptors of work.
- An EPA usually requires multiple competencies in an integrative, holistic nature.

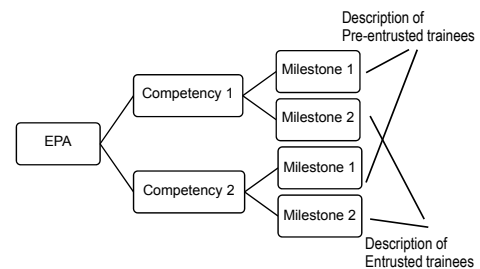
EPAs and Competencies

EPA	Med knowledge	Patient care	Interpersonal skills	Professionalism	Practice-based learning	Systems-based practice
Performing appendectomy	x	x				
Executing a patient handover	x	x	x			x
Designing therapy protocol	x				x	
Chairing multidisciplinary meeting		x	x	x		x
Request organ donation			x	x		
Manage CRF		x	x	x		x

Milestone

- Stages in the development of specific competencies
- Milestones may link to a supervisor's EPA decisions

How They Related?



How to Proceed?

- How many EPAs are useful?
 - GME < 20
- Describe EPA
- Link EPA with competencies
- Describe milestones

Assessing Trainees

1. Observation but no execution
2. Execution with direct, proactive supervision
3. Execution with reactive supervision (i.e., on request)
4. Supervision at a distance
5. Supervision provided by the trainee to more junior colleagues

Cate OT. Nuts and bolts of entrustable professional activities. JGME 2013.



Tomorrow's Doctors, Tomorrow's Cures®



Core Entrustable Professional Activities for Entering Residency

EPA 1 Toolkit: Gather a History and Perform a Physical Examination

Learn

Serve

Lead

Association of
American Medical Colleges



Core Entrustable Professional Activities for Entering Residency



EPA 1 Toolkit: Gather a History and Perform a Physical Examination

Association of American Medical Colleges
Washington, D.C.



EPA 1 Workgroup

Beth Barron, MD, Columbia University
Philip Orlander, MD, The University of Texas Health Science Center at Houston
Michael L. Schwartz, PhD, Yale University

Senior Editors

Vivian Obeso, MD, Florida International University
David Brown, MD, Florida International University
Carrie Phillipi, MD, PhD, Oregon Health & Science University

Editors

Meenakshy Aiyer, MD, University of Illinois
Beth Barron, MD, Columbia University
Jan Bull, MA, Association of American Medical Colleges
Teresa J. Carter, EdD, Virginia Commonwealth University
Matthew Emery, MD, MSc, Michigan State University
Colleen Gillespie, PhD, New York University
Mark Hormann, MD, The University of Texas Health Science Center at Houston
Abbas Hyderi, MD, MPH, University of Illinois
Carla Lupi, MD, Florida International University
Michael L. Schwartz, PhD, Yale University
Margaret Uthman, MD, The University of Texas Health Science Center at Houston
Eduard E. Vasilevskis, MD, MPH, Vanderbilt University
Sandra Yingling, PhD, University of Illinois

AAMC Staff

Alison Whelan, MD
Chief Medical Education Officer

Chris Hanley, MBA
Project Manager

Lynn Shaul, MA
Senior Research Specialist

For inquiries and correspondence, contact Dr. Vivian Obeso at vobeso@fiu.edu, Carrie Phillipi at phillica@ohsu.edu, or Dr. Alison Whelan at awhelan@aamc.org.

This is a publication of the Association of American Medical Colleges. The AAMC serves and leads the academic medicine community to improve the health of all. aamc.org

© 2017 Association of American Medical Colleges. May be reproduced and distributed with attribution for educational or noncommercial purposes only.

Suggested Toolkit Citation:

Obeso V, Brown D, Aiyer M, Barron B, Bull J, Carter T, Emery M, Gillespie C, Hormann M, Hyderi A, Lupi C, Schwartz ML, Uthman M, Vasilevskis EE, Yingling S, Phillipi C, eds.; for Core EPAs for Entering Residency Pilot Program. *Toolkits for the 13 Core Entrustable Professional Activities for Entering Residency*. Washington, DC: Association of American Medical Colleges; 2017. aamc.org/initiatives/coreepas/publicationsandpresentations.

Suggested One-Page Schematic Citation:

Barron B, Orlander P, Schwartz ML. *Core Entrustable Professional Activities for Entering Residency—EPA 1 Schematic: Gather a History and Perform a Physical Examination*. Obeso V, Brown D, Phillipi C, eds. Washington, DC: Association of American Medical Colleges; 2017. aamc.org/initiatives/coreepas/publicationsandpresentations.



Core Entrustable Professional Activities for Entering Residency



Contents

User Guide	2
One-Page Schematics	3
Frequently Asked Questions	5
EPA 1 Schematic	7
Appendix 1: Core EPA Pilot Supervision and Coactivity Scales	8
Appendix 2: Resources Related to EPA 1	10
Appendix 3: Behaviors and Vignettes	12
Appendix 4: The Physician Competency Reference Set (PCRS)	13
References	17
Publications From the Core EPA Pilot	17
Other Related Publications	17



Core Entrustable Professional Activities for Entering Residency



User Guide

This toolkit is for medical schools interested in implementing the Core Entrustable Professional Activities (EPAs) for Entering Residency. Written by the AAMC Core EPA Pilot Group, the toolkit expands on the EPA framework outlined in the *EPA Developer's Guide* (AAMC 2014). The Pilot Group identified progressive sequences of student behavior that medical educators may encounter as students engage in the medical school curriculum and became proficient in integrating their clinical skills. These sequences of behavior are articulated for each of the 13 EPAs in one-page schematics to provide a framework for understanding EPAs; additional resources follow.

This toolkit includes:

- One-page schematic of each EPA
- Core EPA Pilot supervision and coactivity scales
- List of resources associated with each EPA
- Reference to EPA bulleted behaviors and vignettes from the *Core EPA Guide*
- The Physician Competency Reference Set
- Opportunities for engagement with the Core EPA Pilot



Core Entrustable Professional Activities for Entering Residency



One-Page Schematics

In 2014, the AAMC launched a pilot project with 10 institutions to address the feasibility of implementing 13 EPAs for entering residency in undergraduate medical education. To standardize our approach as a pilot and promote a shared mental model, the Core EPA Pilot Group developed one-page schematics for each of the 13 EPAs.

These schematics were developed to translate the rich and detailed content within *The Core Entrustable Professional Activities for Entering Residency Curriculum Developers' Guide* published in 2014 by the AAMC into a one-page, easy-to-use format (AAMC 2014). These one-page schematics of developmental progression to entrustment provide user-friendly descriptions of each EPA. We sought fidelity to the original ideas and concepts created by the expert drafting panel that developed the *Core EPA Guide*.

We envision the one-page schematics as a resource for:

- Development of curriculum and assessment tools
- Faculty development
- Student understanding
- Entrustment committees, portfolio advisors, and others tracking longitudinal student progress

Understanding the One-Page Schematic

Performance of an EPA requires integration of multiple competencies (Englander and Carraccio 2014). Each EPA schematic begins with its list of key functions and related competencies. The functions are followed by observable behaviors of increasing ability describing a medical student's development toward readiness for indirect supervision. The column following the functions lists those behaviors requiring immediate correction or remediation. The last column lists expected behaviors of an entrustable learner.

The members of the Curriculum and Assessment Team of the Core EPA Pilot Group led this initiative. Thirteen EPA groups, each comprising representatives from four to five institutions, were tasked with creating each EPA schematic. Development of the schematics involved an explicit, standardized process to reduce variation and ensure consistency with functions, competencies, and the behaviors explicit in the *Core EPA Guide*. Behaviors listed were carefully gathered from the *Core EPA Guide* and reorganized by function and competency and listed in a developmental progression. The Curriculum and Assessment Team promoted content validity by carrying out iterative reviews by telephone conference call with the members of the Core EPA Pilot Group assigned to each EPA.

EPA Curriculum and Assessment

Multiple methods of teaching and assessing EPAs throughout the curriculum will be required to make a summative entrustment decision about residency readiness. The schematics can help to systematically identify and map curricular elements required to prepare students to perform EPAs. Specific prerequisite curricula may be needed to develop knowledge, skills, and attitudes before the learner engages in practice of the EPA.

To implement EPAs, medical schools should identify where in the curriculum EPAs will be taught, practiced, and assessed. Among other modalities, simulation, reflection, and standardized and structured experiences will all provide data about student competence. However, central to the concept of entrustment is the global performance of EPAs in authentic clinical settings, where the EPA is taught and assessed holistically, not as the sum of its parts.



Workplace-Based Assessments: Supervision and Coactivity Scales

On a day-to-day basis, clinical supervisors make and communicate judgments about how much help (coactivity) or supervision a student or resident needs. “Will I let the student go in the room without me? How much will I let the student do versus observe? Because I wasn’t present to observe, how much do I need to double-check?” Scales for clinical supervisors to determine how much help or supervision a student needs for a specific activity have been proposed (Chen et al 2015; Rekman et al 2016). There is limited validity evidence for these scales, and no published data comparing them. Given our initial experience, the Core EPA Pilot Group has agreed on a trial using modified versions of these scales (Appendix 1).

Resources

The Pilot Group compiled a list of resources, including relevant Critical Synthesis Packages from MedEdPORTAL®, a review of current existing literature, teaching methods, and assessment tools related to each EPA (Appendix 2). This collection of products may help schools with implementation. For example, schools may find the teaching methods and assessment tools useful when considering multiple sources of data about student performance that may eventually contribute to a summative entrustment decision. The Pilot Group concluded that new teaching methods and assessment tools will be needed to complement these resources. This need is particularly relevant for workplace-based assessments where the synthetic performance of an EPA is linked to a level of supervision. We envision the one-page schematics as a resource for the development of new teaching and assessment methods.



Core Entrustable Professional Activities for Entering Residency



Frequently Asked Questions

Why are EPAs important?

In many cases, medical school graduates are perceived by residency program directors as insufficiently prepared at the beginning of their residency training for indirect supervision in clinical skills and for exhibiting professional behaviors. The EPAs define a shared set of clinical activities that residents are expected to perform on day one of residency. This is an important opportunity for undergraduate medical education to develop a new construct toward preparedness and, as an end goal, improvements in patient safety. Ideally, students will perform the Core EPAs consistently in situations of varying complexity as they practice and receive actionable feedback, formulating learning goals for future demonstrations of competence.

What does “entrustment” mean in the context of the EPAs?

Entrustment is defined as trustworthiness in applying knowledge, skills, and attitudes in performance of an EPA. To be “trustworthy,” students must consistently demonstrate attributes such as conscientiousness, knowledge of their own limits and help-seeking behavior (discernment), and truthfulness (Kennedy et al 2008). Throughout medical education, students should be assessed on trustworthiness—though this may occur implicitly or explicitly. The EPA framework makes this assessment explicit and transparent.

EPA entrustment is defined as a judgment by a supervisor or collection of supervisors signaling a student has met specific, defined expectations for needing limited supervision. The Core EPA Pilot Group recommends the formation of an entrustment committee to make evidence-based summative entrustment decisions about each student’s readiness for residency (Brown et al 2017).

What is the relationship between competencies and EPAs?

The EPA framework reorganizes competencies into observable units of clinical work by function. Each function is a subunit of work required to perform an EPA. The functions and related competencies are the parts, and the EPA is the whole. The Toolkit’s one-page schematics highlight an EPA’s specific functions with underlying competencies into observable behaviors within a developmental progression toward entrustment.

Although tracking progression within individual functions can help learners develop appropriate skills, monitoring learner progress toward entrustability for that EPA requires synthesis: At some point the learner must apply each of the functions in execution of the EPA task. *To this end, we emphasize the importance of the holistic nature of the EPA and prioritize assessment for entrustment in these activities in workplace settings as a whole, not as the sum of their parts.*

Is the one-page schematic designed as a rubric for student assessment?

No, the one-page schematics are not intended to serve as assessment tools. They can serve as guides for development of instructional, feedback, and assessment tools for EPAs. We share them as a framework for understanding the developmental progression that graduating medical students should demonstrate as a reflection of their readiness for residency.



Core Entrustable Professional Activities for Entering Residency



How can I or my institution become more involved?

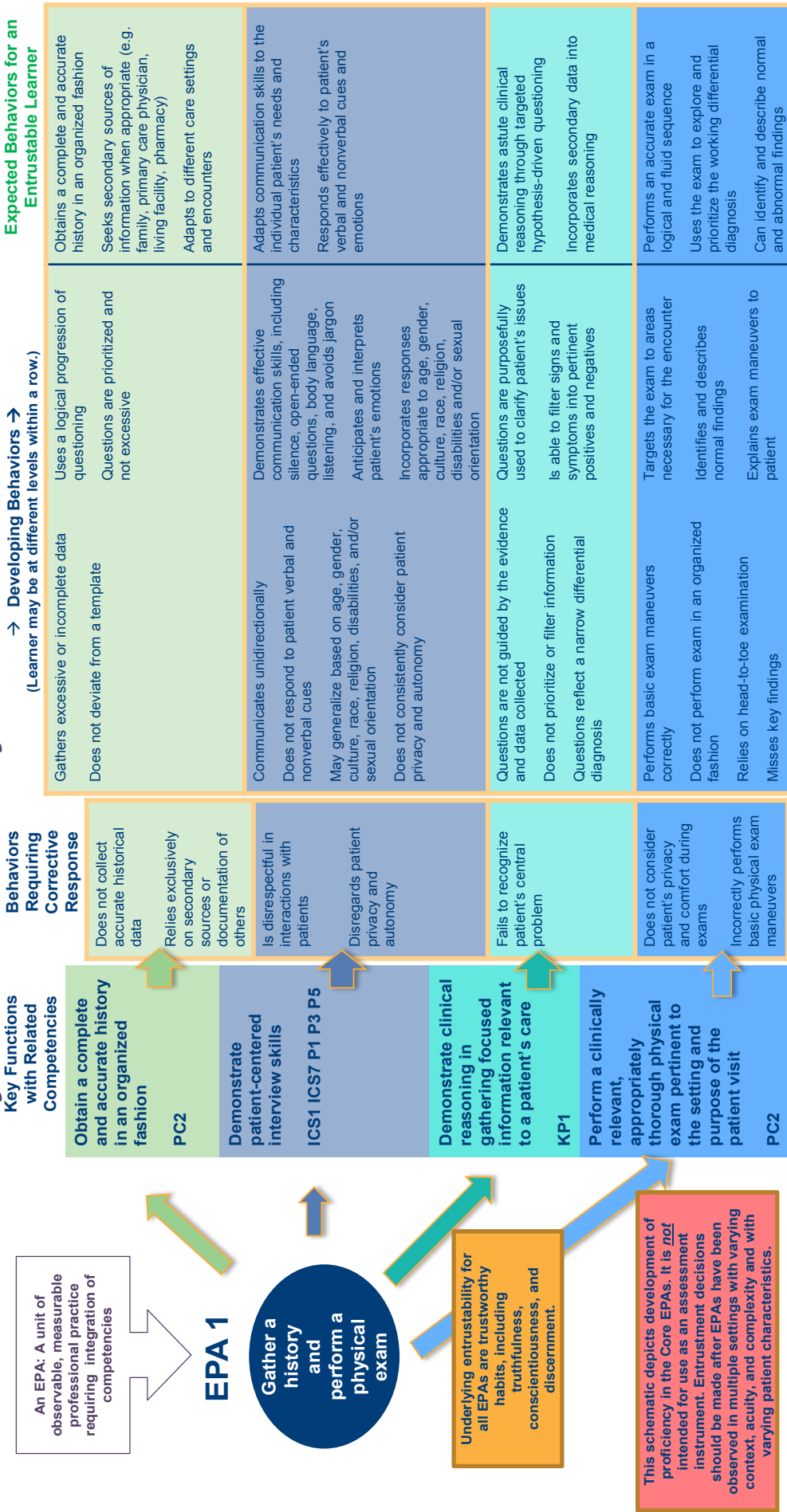
Medical schools in the AAMC pilot, those interested in implementing EPAs, and those wondering about the faculty resources needed to teach and assess EPAs are already part of a dynamic learning community. Opportunities for engaging with others exist through the AAMC Core EPA listserv, conference presentations, collaborative projects, and in informal medical education networks. Your contributions help shape the work of the Core EPA Pilot project and are a source of new ideas, feedback, and suggestions for implementation. We invite you to continue your conversations with us by sharing the decisions you face within the unique culture of your institution.

- To subscribe to the Core EPAs listserv, send a blank email to subscribe-coreepas@lists.aamc.org. To post a comment to the listserv, simply send an email to coreepas@lists.aamc.org.
- Core EPA Pilot Website: <https://www.aamc.org/initiatives/coreepas/>
- Publications from the Core EPA Pilot Group:
<https://www.aamc.org/initiatives/coreepas/publicationsandpresentations/>
- Core EPA Pilot Group email for queries and observations: coreepas@aamc.org

Core Entrustable Professional Activities for Entering Residency



EPA 1: Gather a History and Perform a Physical Exam



Barron, B, Ohlander, P, Schwartz, ML, Obeso V, Brown D, Phillip C, eds.; for Core EPAs for Entering Residency Pilot Program Adapted from the Association of American Medical Colleges (AAMC): Core entrustable professional activities for entering residency. 2014.

Association of American Medical Colleges



Appendix 1: Core EPA Pilot Supervision and Coactivity Scales

Scales for clinical supervisors to determine how much help (coactivity) or supervision they judge a student needs for a specific activity have been proposed—the Chen entrustment scale and the Ottawa scale (Chen et al 2015; Rekman et al 2016). There is limited validity evidence for these scales and no published data comparing them. We include these published tools here for your reference. The Core EPA Pilot Group has agreed on a trial using modified versions of these scales (described below). A description of how the pilot is working with these scales is available on the [Core EPA website](#).

Modified Chen entrustment scale: If you were to supervise this student again in a similar situation, which of the following statements aligns with how you would assign the task?	Corresponding excerpt from original Chen entrustment scale (Chen et al 2015)
1b. “Watch me do this.”	1b. Not allowed to practice EPA; allowed to observe
2a. “Let’s do this together.”	2a. Allowed to practice EPA only under proactive, full supervision as coactivity with supervisor
2b. “I’ll watch you.”	2b. Allowed to practice EPA only under proactive, full supervision with supervisor in room ready to step in as needed
3a. “You go ahead, and I’ll double-check all of your findings.”	3a. Allowed to practice EPA only under reactive/on-demand supervision with supervisor immediately available, all findings double-checked
3b. “You go ahead, and I’ll double-check key findings.”	3b. Allowed to practice EPA only under reactive/on demand supervision with supervisor immediately available, key findings double-checked



Core Entrustable Professional Activities for Entering Residency



Modified Ottawa scale: In supervising this student, how much did you participate in the task?	Original Ottawa scale (Rekman et al 2016)
1. “I did it.” Student required complete guidance or was unprepared; I had to do most of the work myself.	1. “I had to do.” (i.e., requires complete hands-on guidance, did not do, or was not given the opportunity to do)
2. “I talked them through it.” Student was able to perform some tasks but required repeated directions.	2. “I had to talk them through.” (i.e., able to perform tasks but requires constant direction)
3. “I directed them from time to time.” Student demonstrated some independence and only required intermittent prompting.	3. “I had to prompt them from time to time.” (i.e., demonstrates some independence, but requires intermittent direction)
4. “I was available just in case.” Student functioned fairly independently and only needed assistance with nuances or complex situations.	4. “I needed to be there in the room just in case.” (i.e., independence but unaware of risks and still requires supervision for safe practice)
5. (No level 5: Students are ineligible for complete independence in our systems.)	5. “I did not need to be there.” (i.e., complete independence, understands risks and performs safely, practice ready)



Core Entrustable Professional Activities for Entering Residency



Appendix 2: Resources Related to EPA 1

Hypothesis-Driven Physical Examination (HDPE)

Uchida T, Heiman H. Critical synthesis package: hypothesis-driven physical examination (HDPE). MedEdPORTAL Publications. 2013;9:9435. doi.org/10.15766/mep.2374-8265.9435.

Mini-Clinical Evaluation Exercise

Perkowski L. Critical synthesis package: mini-clinical evaluation exercise (mCEX). MedEdPORTAL Publications. 2014;10:9793. doi.org/10.15766/mep.2374-8265.9793.

Faculty Observer Rating Scale (FORS)

Nadir N. Critical synthesis package: faculty observer rating scale (FORS). MedEdPORTAL Publications. 2014;10:9853. doi.org/10.15766/mep.2374-8265.9853.

Interpreter Scale (IS)

Pelts M, Albright D. Critical synthesis package: interpreter scale (IS). MedEdPORTAL Publications. 2014;10:9845. doi.org/10.15766/mep.2374-8265.9845.

Patient-Practitioner Orientation Scale (PPOS)

Trapp S, Stern M. Critical synthesis package: patient-practitioner orientation scale (PPOS). MedEdPORTAL Publications. 2013;9:9501. doi.org/10.15766/mep.2374-8265.9501.

Assessment of Professional Behaviors (APB)

Fornari A, Akbar S, Tyler S. Critical synthesis package: assessment of professional behaviors (APB). MedEdPORTAL Publications. 2014;10:9902. doi.org/10.15766/mep.2374-8265.9902.

MAAS-Global Manual 2000

Lacy N. Critical synthesis package: MAAS-global. MedEdPORTAL Publications. 2015;11:10028. dx.doi.org/10.15766/mep.2374-8265.10028.

Cross-Cultural Counseling Inventory–Revised (CCCI-R)

Young K. Critical synthesis package: cross-cultural counseling inventory–revised (CCCI-R). MedEdPORTAL Publications. 2014;10:9950. doi.org/10.15766/mep.2374-8265.9950.

CAM Health Belief Questionnaire (CHBQ)

Nicolais C, Stern M. Critical synthesis package: CAM health belief questionnaire (CHBQ). MedEdPORTAL Publications. 2014;10:9882. doi.org/10.15766/mep.2374-8265.9882.

Relational Communication Scale (RCS)



Core Entrustable Professional Activities for Entering Residency



Hartmark-Hill J. Critical synthesis package: relational communication scale (RCS). MedEdPORTAL Publications. 2013;9:9454. doi.org/10.15766/mep.2374-8265.9454.

Communication Assessment Tool (CAT)

Ibrahim H. Critical synthesis package: communication assessment tool (CAT). MedEdPORTAL Publications. 2014;10:9806. dx.doi.org/10.15766/mep.2374-8265.9806.

Liverpool Communication Skills Assessment Scale (LCSAS)

Islam L, Dorflinger L. Critical synthesis package: Liverpool communication skills assessment scale (LCSAS). MedEdPORTAL Publications. 2015;11:10126. dx.doi.org/10.15766/mep.2374-8265.10126.

Communication Curriculum Package

Hofert S, Burke M, Balighian E, Serwint J. Improving provider-patient communication: a verbal and non-verbal communication skills curriculum. MedEdPORTAL Publications. 2015;11:10087. dx.doi.org/10.15766/mep.2374-8265.10087.

Professionalism Mini-Evaluation Exercise (P-MEX)

Gathright M. Critical synthesis package: professionalism mini-evaluation exercise (P-MEX). MedEdPORTAL Publications. 2014;10:9929. doi.org/10.15766/mep.2374-8265.9929.

Rochester Communication Rating Scale

Stalburg C. Critical synthesis package: Rochester communication rating scale. MedEdPORTAL Publications. 2015;11:9969. doi.org/10.15766/mep.2374-8265.9969.

Evidence in the Literature

Gowda D, Blatt B, Fink MJ, Kosowicz LY, Baecker A, Silvestri RC. A core physical exam for medical students: results of a national survey. *Acad Med*. 2014;89(3):436-442. doi: 10.1097/acm.000000000000137.



Core Entrustable Professional Activities for Entering Residency



Appendix 3: Behaviors and Vignettes

The [Core EPA Guide](#) produced by the AAMC contains additional detailed information that may be useful for curriculum designers.

1. For a convenient list of behaviors for this EPA that were used to develop a developmental progression, we refer you to the [Core EPA Guide](#).
2. For exemplars of learner vignettes that highlight pre-entrustable and entrustable scenarios, please see the [Core EPA Guide](#).



Core Entrustable Professional Activities for Entering Residency



Appendix 4: The Physician Competency Reference Set (PCRS)

The Physician Competency Reference Set (Englander et al 2013) is provided for cross-referencing with the one-page schematic.

1. PATIENT CARE (PC): Provide patient-centered care that is compassionate, appropriate, and effective for the treatment of health problems and the promotion of health

- 1.1 Perform all medical, diagnostic, and surgical procedures considered essential for the area of practice
- 1.2 Gather essential and accurate information about patients and their condition through history-taking, physical examination, and the use of laboratory data, imaging, and other tests
- 1.3 Organize and prioritize responsibilities to provide care that is safe, effective, and efficient
- 1.4 Interpret laboratory data, imaging studies, and other tests required for the area of practice
- 1.5 Make informed decisions about diagnostic and therapeutic interventions based on patient information and preferences, up-to-date scientific evidence, and clinical judgment
- 1.6 Develop and carry out patient management plans
- 1.7 Counsel and educate patients and their families to empower them to participate in their care and enable shared decision making
- 1.8 Provide appropriate referral of patients, including ensuring continuity of care throughout transitions between providers or settings and following up on patient progress and outcomes
- 1.9 Provide health care services to patients, families, and communities aimed at preventing health problems or maintaining health
- 1.10 Provide appropriate role modeling
- 1.11 Perform supervisory responsibilities commensurate with one's roles, abilities, and qualifications

2. KNOWLEDGE FOR PRACTICE (KP): Demonstrate knowledge of established and evolving biomedical, clinical, epidemiological, and social-behavioral sciences, as well as the application of this knowledge to patient care

- 2.1 Demonstrate an investigatory and analytic approach to clinical situations
- 2.2 Apply established and emerging biophysical scientific principles fundamental to health care for patients and populations
- 2.3 Apply established and emerging principles of clinical sciences to diagnostic and therapeutic decision making, clinical problem solving, and other aspects of evidence-based health care
- 2.4 Apply principles of epidemiological sciences to the identification of health problems, risk factors, treatment strategies, resources, and disease prevention/health promotion efforts for patients and populations
- 2.5 Apply principles of social-behavioral sciences to provision of patient care, including assessment of the impact of psychosocial-cultural influences on health, disease, care-seeking, care compliance, and barriers to and attitudes toward care
- 2.6 Contribute to the creation, dissemination, application, and translation of new health care knowledge and practices



Core Entrustable Professional Activities for Entering Residency



- 3. PRACTICE-BASED LEARNING AND IMPROVEMENT (PBLI): Demonstrate the ability to investigate and evaluate their care of patients, to appraise and assimilate scientific evidence, and to continuously improve patient care based on constant self-evaluation and lifelong learning**

 - 3.1 Identify strengths, deficiencies, and limits in one's knowledge and expertise
 - 3.2 Set learning and improvement goals
 - 3.3 Identify and perform learning activities that address one's gaps in knowledge, skills, or attitudes
 - 3.4 Systematically analyze practice using quality-improvement methods, and implement changes with the goal of practice improvement
 - 3.5 Incorporate feedback into daily practice
 - 3.6 Locate, appraise, and assimilate evidence from scientific studies related to patients' health problems
 - 3.7 Use information technology to optimize learning
 - 3.8 Participate in the education of patients, families, students, trainees, peers, and other health professionals
 - 3.9 Obtain and utilize information about individual patients, populations of patients, or communities from which patients are drawn to improve care
 - 3.10 Continually identify, analyze, and implement new knowledge, guidelines, standards, technologies, products, or services that have been demonstrated to improve outcomes
- 4. INTERPERSONAL AND COMMUNICATION SKILLS (ICS): Demonstrate interpersonal and communication skills that result in the effective exchange of information and collaboration with patients, their families, and health professionals**

 - 4.1 Communicate effectively with patients, families, and the public, as appropriate, across a broad range of socioeconomic and cultural backgrounds
 - 4.2 Communicate effectively with colleagues within one's profession or specialty, other health professionals, and health-related agencies (see also interprofessional collaboration competency, IPC 7.3)
 - 4.3 Work effectively with others as a member or leader of a health care team or other professional group (see also IPC 7.4)
 - 4.4 Act in a consultative role to other health professionals
 - 4.5 Maintain comprehensive, timely, and legible medical records
 - 4.6 Demonstrate sensitivity, honesty, and compassion in difficult conversations (e.g., about issues such as death, end-of-life issues, adverse events, bad news, disclosure of errors, and other sensitive topics)
 - 4.7 Demonstrate insight and understanding about emotions and human responses to emotions that allow one to develop and manage interpersonal interactions
- 5. PROFESSIONALISM (P): Demonstrate a commitment to carrying out professional responsibilities and an adherence to ethical principles**

 - 5.1 Demonstrate compassion, integrity, and respect for others
 - 5.2 Demonstrate responsiveness to patient needs that supersedes self-interest
 - 5.3 Demonstrate respect for patient privacy and autonomy



Core Entrustable Professional Activities for Entering Residency



- 5.4 Demonstrate accountability to patients, society, and the profession
- 5.5 Demonstrate sensitivity and responsiveness to a diverse patient population, including but not limited to diversity in gender, age, culture, race, religion, disabilities, and sexual orientation
- 5.6 Demonstrate a commitment to ethical principles pertaining to provision or withholding of care, confidentiality, informed consent, and business practices, including compliance with relevant laws, policies, and regulations

6. SYSTEMS-BASED PRACTICE (SBP): Demonstrate an awareness of and responsiveness to the larger context and system of health care, as well as the ability to call effectively on other resources in the system to provide optimal health care

- 6.1 Work effectively in various health care delivery settings and systems relevant to one's clinical specialty
- 6.2 Coordinate patient care within the health care system relevant to one's clinical specialty
- 6.3 Incorporate considerations of cost awareness and risk–benefit analysis in patient and/or population-based care
- 6.4 Advocate for quality patient care and optimal patient care systems
- 6.5 Participate in identifying system errors and implementing potential systems solutions
- 6.6 Perform administrative and practice management responsibilities commensurate with one's role, abilities, and qualifications

7. INTERPROFESSIONAL COLLABORATION (IPC): Demonstrate the ability to engage in an interprofessional team in a manner that optimizes safe, effective patient- and population-centered care

- 7.1 Work with other health professionals to establish and maintain a climate of mutual respect, dignity, diversity, ethical integrity, and trust
- 7.2 Use the knowledge of one's own role and those of other professions to appropriately assess and address the health care needs of the patients and populations served
- 7.3 Communicate with other health professionals in a responsive and responsible manner that supports the maintenance of health and the treatment of disease in individual patients and populations
- 7.4 Participate in different team roles to establish, develop, and continuously enhance interprofessional teams to provide patient- and population-centered care that is safe, timely, efficient, effective, and equitable

8. PERSONAL AND PROFESSIONAL DEVELOPMENT (PPD): Demonstrate the qualities required to sustain lifelong personal and professional growth

- 8.1 Develop the ability to use self-awareness of knowledge, skills, and emotional limitations to engage in appropriate help-seeking behaviors
- 8.2 Demonstrate healthy coping mechanisms to respond to stress
- 8.3 Manage conflict between personal and professional responsibilities
- 8.4 Practice flexibility and maturity in adjusting to change with the capacity to alter behavior
- 8.5 Demonstrate trustworthiness that makes colleagues feel secure when one is responsible for the care of patients
- 8.6 Provide leadership skills that enhance team functioning, the learning environment, and/or the health care delivery system



Core Entrustable Professional Activities for Entering Residency



- 8.7 Demonstrate self-confidence that puts patients, families, and members of the health care team at ease
- 8.8 Recognize that ambiguity is part of clinical health care and respond by using appropriate resources in dealing with uncertainty



Core Entrustable Professional Activities for Entering Residency



References

Publications From the Core EPA Pilot Group

Brown DR, Gillespie CC, Warren JB. [EPA 9—Collaborate as a member of an interprofessional team: a short communication from the AAMC Core EPAs for Entering Residency Pilot Schools](#). *Med Sci Educ*. 2016;26(3):457-461.

Brown DR, Warren JB, Hyderi A, Drusin RE, Moeller J, Rosenfeld M, et al. [Finding a path to entrustment in undergraduate medical education: a progress report from the AAMC Core Entrustable Professional Activities for Entering Residency Entrustment Concept Group](#). *Acad Med*. 2017;92(6):774-779.

Englander R, Cameron T, Ballard AJ, Dodge J, Bull J, Aschenbrenner CA. [Toward a common taxonomy of competency domains for the health professions and competencies for physicians](#). *Acad Med*. 2013;88(8):1088-1094.

Englander R, Carraccio C. [From theory to practice: making entrustable professional activities come to life in the context of milestones](#). *Acad Med*. 2014;89(10):1321-1323.

Favreau MA, Tewksbury L, Lupi C, Cutrer WB, Jokela JA, Yarris LM. [Constructing a shared mental model for faculty development for the Core Entrustable Professional Activities for Entering Residency](#). *Acad Med*. 2017;92(6):759-764.

Lomis KD, Ryan MS, Amiel JM, Cocks PM, Uthman MO, Esposito KF. [Core entrustable professional activities for entering residency pilot group update: considerations for medical science educators](#). *Med Sci Educ*. 2016;26(4):797-800.

Lomis K, Amiel JM, Ryan MS, et al. [Implementing an entrustable professional activities framework in undergraduate medical education: early lessons from the AAMC core entrustable professional activities for entering residency pilot](#). *Acad Med*. 2017;92(6):765-770.

Other Related Publications

Association of American Medical Colleges (AAMC). Core entrustable professional activities for entering residency. mededportal.org/icollaborative/resource/887. May 28, 2014. Accessed March 1, 2016.

Chen HC, van den Broek WS, ten Cate O. [The case for use of entrustable professional activities in undergraduate medical education](#). *Acad Med*. 2015;90(4):431-436.

Englander R, Cameron T, Ballard AJ, Dodge J, Bull J, Aschenbrenner CA. [Toward a common taxonomy of competency domains for the health professions and competencies for physicians](#). *Acad Med*. 2013;88(8):1088-1094.

Kennedy TJT, Regehr G, Baker GR, Lingard L. [Point-of-care assessment of medical trainee competence for independent clinical work](#). *Acad Med*. 2008;83(10):S89-S92.

Peters H, Holzhausen Y, Boscardin C, ten Cate O, Chen HC. [Twelve tips for the implementation of EPAs for assessment and entrustment decisions](#). *Med Teach*. 2017;39(8):802-807.

Rekman J, Hamstra SJ, Dudek N, Wood T, Seabrook C, Gofton W. [A new instrument for assessing resident competence in surgical clinic: the Ottawa clinic assessment tool](#). *J Surg Educ*. 2016;73(4):575-582.

15 กรกฎาคม 2563

Standard setting

Chapter

36

Section 6:
Assessment

Standard setting

J. Norcini, D. W. McKinley

Introduction

In medical education, it is common to need to identify knowledge or performance that is 'just good enough' for a particular purpose. One example is a pass or fail multiple-choice examination, where a single score is chosen as the cutoff. Passing examinees achieve that cutoff score or higher, while failing examinees do not. Passing implies sufficient knowledge or skill given the purpose of the test, while failing implies insufficient knowledge or skill. Standard setting is the process of demarcating the level of knowledge and skill indicating proficiency and identifying a score on the examination that corresponds to it.

Unlike many medical tests, educational assessments only rarely have a gold standard against which to establish the validity of a cutoff score. The nature of a 'competent' physician or 'unsatisfactory' medical student varies over time, place and many other factors. Consequently, standards on educational tests are the expression of judgement in the context of a particular assessment, its purpose and the wider social-professional environment.

Because standards are based on judgement, methods for selecting them are not intended to discover an underlying truth. Instead, they are a means for gathering a variety of perspectives, blending them together and expressing them as a single score on a particular assessment. Consequently, the methods do not differ in the correctness of the standards they yield, but in their credibility and defensibility. This chapter describes the types of standards, specifies the important characteristics of the standard setters and the methods, reviews some of the common methods for setting standards and provides a framework for evaluating their credibility (Norcini & Shea 1997, Norcini 2003, Norcini & Guille 2002).

Types of standards

There are two types of standards:

- relative
- absolute.

Relative standards are expressed in terms of the performance of a group of examinees. For instance, a relative standard may be that the 120 examinees with the highest scores are admitted to medical school. This type of standard is appropriate for assessments intended to select a certain number or percentage of examinees, such as tests for admissions or placement.

Absolute standards are expressed in terms of the performance of examinees against the test material. For instance, a passing score may be that any examinee who correctly answers 75% or more of the questions knows enough anatomy to pass. This type of standard is appropriate for assessments intended to determine whether examinees have the necessary knowledge or skills for a particular purpose, such as course completion or graduation from medical school.

Important characteristics of the standard setters and standard setting methods

The characteristics of the standard setters are likely to have the biggest impact on the credibility of a standard. The standard setters must understand the purpose of the test and the reason for establishing the cut score, know the content and be familiar with the examinees. In a low-stakes setting like a course, a single faculty member is credible, but standards will vary over time, and he or she has a conflict of interest in being both the teacher and assessor. In a high-stakes setting like licensure, a significant number of standard setters need to be involved because this increases the

reproducibility of standards and reduces the effects of 'hawks' and 'doves'. Ideally, the group would be free of conflicts of interest, include a mix of educators and practitioners and be balanced with regard to gender, race, geography and the like.

The specific method chosen to set standards is not as important as whether it produces results that are fit for the purpose of the test, relies on informed expert judgement, demonstrates due diligence, is supported by a body of research and is easy to explain and implement.

FIT FOR PURPOSE

The method must produce standards that are consistent with the purpose of the assessment. Methods that turn out relative standards are to be used when the purpose is to select a specific number of examinees. Methods that turn out absolute standards are to be used when the purpose is to judge competence.

BASED ON INFORMED JUDGEMENT

Methods for setting standards can be based entirely on empirical results (e.g. consequences, performance on criteria), entirely on expert judgement or on a blend of the two. There are only rarely instances in which it is possible to base a standard entirely on empirical results in medical education, with the exception of a few admissions testing situations (where outcome data, like successful completion of a course, are available and relative standards are being used).

Instead, most of the methods allow a standard to be based solely on the judgement of experts, without reference to performance data (e.g. the difficulty of the questions, the pass rate). Moreover, standard setters sometimes become uncomfortable when data are presented, thinking that it 'biases' their judgements.

In fact, methods for setting standards are not intended to discover an essential truth but to create a credible standard out of the judgements of experts. Such credibility derives from decisions that are based on all of the available information. Consequently, methods that permit and encourage expert judgement in the presence of performance data are preferable.

DEMONSTRATES DUE DILIGENCE

Methods that require the standard setters to expend thoughtful effort will demonstrate due diligence and this lends credibility to the final result. In contrast, methods that require quick, global judgements are less credible, and methods requiring several days of effort are unnecessary.

SUPPORTED BY RESEARCH

Methods supported by a research literature will produce more credible results. Ideally, studies should

show that standards are reasonable compared to those produced by other methods, reproducible over groups of judges, insensitive to potentially biasing effects and sensitive to differences in test difficulty and content.

EASILY EXPLAINED AND IMPLEMENTED

Credibility is enhanced if the method is easy to explain and implement. This decreases the amount of training required for the judges, increases the likelihood of their compliance and consistency and assures examinees that they are being treated fairly.

Methods for setting standards

There is a host of methods for setting standards, and many have variations. Reviews and descriptions are available elsewhere (Berk 1986, Cusimano 1996), but according to Livingston and Zieky (1982) they fall into four categories:

- relative methods
- absolute methods based on judgements about assessment content (assessment-centred)
- absolute methods based on judgements about individual examinees (examinee-centred)
- compromise methods.

All of the methods require that several standard setters be selected and that they meet as a group. As the name implies, relative methods produce relative standards and thus judgements are made about what proportion of the examinees should pass. The two groups of methods for setting absolute standards differ in the type of judgements that are being collected. In one group, the standard setters consider whether individual examinees should pass, and these judgements are aggregated to derive the cutoff. In the other group, the standard setters consider individual test questions, and these judgements are combined to calculate the cutting score. The compromise methods require judgements about both what proportion of the examinees should pass and what score they need to achieve to do so. The final result is a compromise between these two types of judgements.

RELATIVE METHODS

In the fixed-percentage method, each standard setter announces what percentage (or number) of examinees is qualified to pass. Their judgements are recorded for all to see, and they then engage in a discussion, often led off by those with the highest and lowest estimates. All are free to change, and when the discussions are over the estimates are averaged. The standard is that score which passes the average percentage (or number) of examinees.

In the reference group method, the process is exactly the same except that the standard setters have a particular group of examinees in mind (e.g. graduates of a certain set of schools or examinees with specific educational experiences). The selection of this reference group is based on the fact that the standard setters are most familiar with them and able to make good judgements about them. The cutting score established for this reference group is applied without modification to all other examinees.

These methods are quick and easy to use, they only have to be repeated occasionally, the standard setters are comfortable making the required judgements and they apply equally well to all different test formats. However, the standards will vary over time with the ability of the examinees, and they are independent of how much examinees know and the content of the test.

ABSOLUTE METHODS BASED ON JUDGEMENTS ABOUT TEST QUESTIONS (TEST-CENTRED)

The two most popular methods in this category have been proposed by Angoff and Ebel. Both methods require that the standard setters specify the characteristics of a borderline group of examinees. The borderline group excludes examinees who would clearly pass or fail and is composed of those about whom the standard setters are uncertain.

In Angoff's method, the standard setters estimate the proportion of the borderline group that would

respond correctly to an item. These are discussed with all being free to change their estimates, and the process is repeated for all items on the test. To calculate the standard, the estimates for each item are averaged and the averages are summed (see Table 36.1). Often, as a 'reality check', examinee performance is provided as well. In this example, the percentage of all examinees choosing the correct option (p value) is provided.

In Ebel's method, the standard setters build a classification table for the items in the test. For example, they might decide to classify items by difficulty (easy, medium and hard) and frequency with which encountered in practice (common and uncommon). The standard setters then assign each item to one of the categories. After all items are assigned, they estimate the proportion of items in a category that borderline examinees will answer correctly (see Table 36.2). As with Angoff's method, a discussion ensues, and estimates can be changed. To determine the standard, the estimates for each category are averaged, multiplied by the number of items in the category, and summed.

These methods are widely used in high-stakes testing situations and there is a considerable body of research supporting this method. The standard setters review every item on the test, resulting in more informed judgements. However, the standard setters sometimes have difficulty envisioning the performance of a borderline group and so feel that they are simply making up numbers. These methods can also be time consuming for long tests.

Table 36.1 Application of Angoff's method to an eight-item test

Question	Standard setter					Judges' mean	Percent choosing correct option
	1	2	3	4	5		
1	.90	.85	.80	.75	.85	.83	0.90
2	.60	.55	.40	.35	.50	.48	0.50
3	.70	.60	.65	.50	.55	.60	0.70
4	.85	.75	.80	.65	.70	.75	0.70
5	.95	.90	.85	.75	.80	.85	0.80
6	.50	.50	.45	.40	.50	.47	0.50
7	.65	.55	.45	.45	.60	.54	0.45
8	.85	.70	.80	.65	.75	.75	0.70
Standard (cut score)						5.27	

The meeting of five standard setters begins with a discussion of the characteristics of a borderline group of students. When the standard setters reach consensus, they turn to a consideration of the first item. The standard setters each estimate aloud what proportion of the hypothetical borderline group would respond correctly to the question. Their estimates are written on a board for all to see and a discussion ensues, led by the standard setters with the highest and lowest estimates. All standard setters are free to change their estimates. The standard setters proceed in this manner through all of the items on the test. The cut score is taken as the sum of the standard setters' mean estimates for each question.



1. If the test is very long and security is not a major issue, have the standard setters meet and judge 30–40 items. Then ask them to do the remainder at home.
2. A reliable standard will result even when subsets of standard setters make judgements about subsets of the items on a test, as long as there are enough doctors involved.
3. If the standard setters have not taken the assessment, they should take it before the meeting because it prevents overconfidence and unrealistically high standards.
4. Give the standard setters the correct answers during the meeting unless they are overconfident. It prevents embarrassment.

In addition to establishing standards for individuals, the Angoff method has been used to set school-level standards (Stern et al 2006). Standards could be set to evaluate the relative strengths and weaknesses of a medical school's programme in providing adequate training and experiences to students based on established competencies (e.g. Tomorrow's Doctors, UK; CANMeds, Canada; ACGME Core Competencies, United States). To set school-level standards, standard setters developed a profile of the borderline school, reviewed assessment materials involved in setting student-level standards, and then were asked to estimate the percentage of students who would

receive passing scores on each of the measures at a borderline school. Variations in judgements were discussed, and then standard setters were provided with consequential data for applying their initial standard to eight unidentified schools, and were permitted to change their judgements. In this study, each of the assessments addressed one of six competency domains but the method could easily apply to standard setting for other outcomes (e.g. course objectives).

To simplify use of the Angoff with OSCEs, a variation is to have judges estimate the station score that would be attained by the borderline examinee. These scores would be averaged for each station and summed to determine the standard. Methods focusing on global judgements of examinee performance have been increasingly studied for OSCEs, and an overview to those methods is provided in the next section.

ABSOLUTE METHODS BASED ON JUDGEMENTS ABOUT INDIVIDUAL EXAMINEES (EXAMINEE-CENTRED)

In the contrasting groups method, a random sample of examinees is drawn before the meeting. The entire test paper of each member of this randomly selected group is taken to the meeting and given, one at a time, to the standard setters. They decide as a group whether each performance is passing or failing. To calculate the standard, the scores of the passers and failers are graphed separately but on the same piece

Table 36.2 Application of Ebel's method to a 100-item test

Category	Average proportion correct	No. of questions	Expected score
Common			
Easy	.95	20	19
Medium	.80	40	32
Hard	.70	10	7
Uncommon			
Easy	.70	15	10.5
Medium	.50	10	5
Hard	.30	5	1.5
Standard (cut score)			75

The standard setters begin with a discussion of how the test questions should be classified; they choose two dimensions, frequency and difficulty. As a group, they go through all of the questions on the test one by one and place them into one of the categories (e.g. easy and common). The standard setters then discuss the characteristics of a borderline group of students. When the group reaches consensus on these characteristics, they turn to a consideration of the first category of questions. The standard setters each estimate aloud what proportion of the hypothetical borderline group would respond correctly to the questions in the category. Their estimates are written on a board for all to see and a discussion ensues, led by the standard setters with the highest and lowest estimates. All standard setters are free to change their estimates. When done, the group proceeds through the rest of the categories. To derive the cut score, the average proportion correct is multiplied by the number of questions in each category to produce expected scores. These expected scores are summed to calculate the standard.

of paper. The cutting score can then be derived in a variety of ways. For example, the point where the curve of the passers overlaps least with the curve of the failers could be taken as the standard. Figure 36.1 provides an illustration.

Standard setters are usually comfortable making the judgements required by this method, and it has the advantage of informing them with the actual test performance of examinees. Further, by shifting where the standard is set, it is possible to maximize or minimize false-positive and false-negative decisions. However, it is difficult for the standard setters to produce a balanced judgement when the test is relatively long. In addition, a large number of examinees need to be judged to produce precise results.

In a variation that makes the contrasting groups method more efficient for an OSCE, the standard setters consider the performances of a sample of examinees on each case. They sort them into 'acceptable' and 'unacceptable' groups, identify the score separating them, and then sum these scores across all of the stations to arrive at the standard for the test.

The contrasting groups method was applied to teacher evaluations (Shea et al 2009), and the standard setters in this study were familiar with teaching responsibilities and qualities needed for promotion and appointment. They reviewed data obtained from

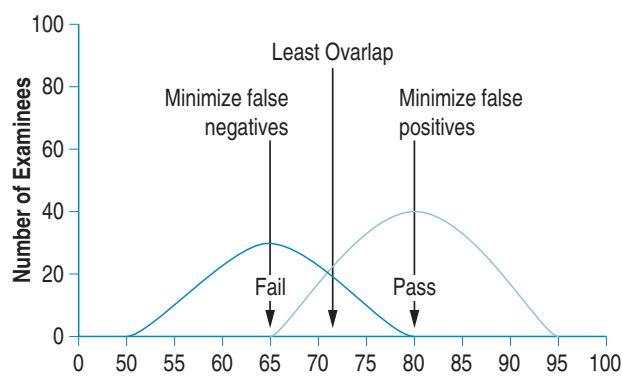


Fig. 36.1 Application of the contrasting groups method to a 100-item test. A sample of students is drawn at random, and the standard setters review the first student's answers to the entire test. After making a decision as a group about whether the performance merits a pass or a fail, the standard setters make similar decisions about the remaining students one by one. After judgements have been made about all the students, the scores of the failing group and passing group are graphed separately. The cut score is usually set at the point of least overlap between the two distributions, 72 in this example. If there is a need to minimize false negatives, the cut score is set at 65, and if there is a need to minimize false positives, it is set at 80.

learners by department, faculty rank and setting (i.e. classroom, clinical). Their task was to sort dossiers of teacher performance into four groups: 'superior', 'excellent', 'satisfactory' and 'unsatisfactory'. Cut scores for each category were established by calculating the means of all judges and identifying the mean between the average scores of two adjacent groups. For example, if the average score for 'superior' dossiers was 85% and the average score for 'excellent' dossiers was 80%, the cut score for the 'superior' category would be 82.5%. For those concerned with promotion (or remediation) of faculty, standard setting may aid data interpretation.

A technique proposed by Dauphinee et al (1997) combines elements of both the Angoff and contrasting groups methods. When physicians are used to observe and score OSCE stations, they can be asked to rate each examinee in such a way that borderline performances are identified. The scores of the examinees with these borderline performances are averaged and then combined over all the stations in the assessment. One potential disadvantage is that for small-scale OSCEs, it may be impossible to secure enough judgements of borderline performance to define a valid standard. To avoid this potential shortcoming, a method using the entire score range has been studied. In the 'borderline regression method', checklist scores are regressed on the global ratings (Kramer et al 2003, Wood et al 2006). This approach has the advantage of using ratings for all examinees, and the midpoint of the rating scale was used to predict the cut score for each station and then averaged to derive the test standard.

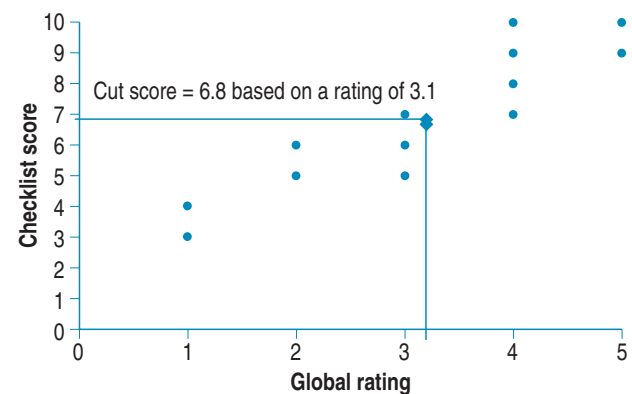


Fig. 36.2 Application of the borderline regression method to a single station. As part of scoring, judges (faculty members) provide a global rating of performance on the station (in this example, unacceptable, borderline, acceptable, good, superior). A regression analysis is run with ratings as the independent variable and checklist score as the dependent variable. The resulting equation is used to find the checklist score corresponding to a particular rating. In this case, the mean rating (3.1) was used, resulting in a station cut score of 6.81 out of 10.

A similar approach that has been studied involved having judges rate each performance as 'adequate' or 'inadequate'. A regression analysis was used to identify the station score at which 50% of the judges rated the performance as 'adequate' and 50% rated the performance as 'inadequate' (McKinley et al 2005). An illustration of the borderline regression method for a single station is provided in Fig. 36.2. To the extent that the judgements are collected as part of the assessment, these methods are all simple to implement, provide consistent ratings, and have been shown to produce acceptable outcomes.

COMPROMISE METHODS

The Hofstee method is the most popular exemplar of this class of methods. The standard setters are asked to produce four judgements: the maximum and minimum acceptable pass rates and maximum and minimum acceptable cutoff scores. These judgements are discussed and changed, as with the other methods, and the final results for the four estimates are obtained by averaging across standard setters. The percentage of examinees who would pass for every possible value of the cut score on the test is graphed, and a rectangle is superimposed as defined by the four judgements of the standard setters. A diagonal is drawn through the box, and the standard is the point where it intersects the examinee performance curve. Figure 36.3 provides an illustration.

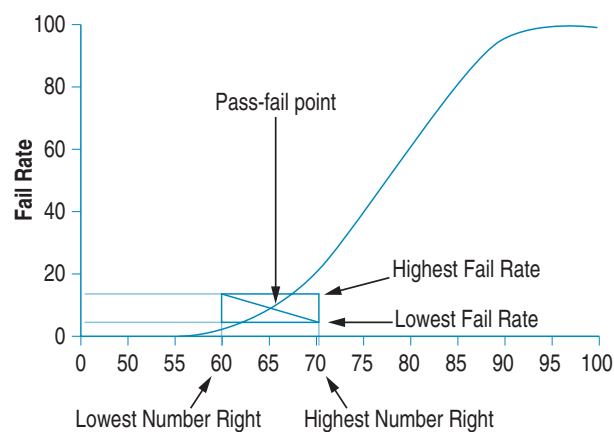



Fig. 36.3 Application of the Hofstee method to a 100-item test; the scores in this case are the numbers of items answered correctly. The standard setters are asked to answer four questions: What are the minimum and maximum acceptable fail rates, and what are the minimum and maximum acceptable cutoff scores? After a discussion where these estimates can be changed, means over all standard setters are calculated. These are graphed as a rectangle and a diagonal is drawn through it. The performance of the students is graphed, showing what the fail rate would be at each possible value for the cutoff.

This method is efficient, and the standard setters feel capable of making the judgements necessary for it. In some instances, however, the curve for examinee performance does not fall within the area circumscribed by the rectangle; in this case, the minimum or maximum acceptable pass rate is selected by default to provide a standard. Because it includes elements of a relative standard, this method is not ideal for regular use in a high-stakes setting. However, it is suitable occasionally, or for use in a low-stakes application.

 The contrasting groups method is especially useful when there is a need to directly manipulate the number of false-positive or false-negative decisions.

Putting it together

BEFORE THE MEETING

Prior to the meeting, the method for setting standards needs to be selected depending on the purpose of the test, the stakes and the resources available. Once this is done, the standard setters should be chosen to be broadly representative of the relevant perspectives. They should all review the assessment in detail so that they are familiar with the content and scoring. This is particularly important for the methods that do not require a review of the test as part of the standard setting process.

DURING THE MEETING

At the beginning of the meeting the methods should be explained to the standard setters. They should then engage in a discussion of the purpose and content of the test and the abilities of the examinees. These discussions are critical because they focus the standard setters on the task and guide them as they begin to make judgements. Once this discussion is completed, the standard setters should practise the method, where reasonable, with data that are not part of the test. Throughout the practice period and the remainder of the meeting, the standard setters should be given feedback about the consequences of their judgements (e.g. what percentage of examinees they would pass). It is important that all standard setters attend the entire meeting and that there are not interruptions. Absences, even for a short time, will generate missing data and could influence the standard more broadly by altering the discussion.

AFTER THE MEETING

Outliers

Common to all of the methods is the possibility that a standard setter with extreme views might

significantly influence the results. Livingston and Zieky (1982) review methods for dealing with this problem, such as removing outlying judgements from the calculations or using the median instead of the mean. The removal of data should be a last resort, however, since it undermines the credibility of the process and the selection of standard setters.

Reliability

It is important to determine whether the results would be the same if the method was repeated with more or different standard setters. This is reliability or reproducibility; there are a variety of ways of calculating it, but generalizability theory offers a good alternative (Brennan & Lockwood 1980). If the results of this analysis are unacceptable given the purpose of the assessment, standard setters can be added in a second application of the method. The data from this second application should be combined with those of the first unless there were significant problems associated with it.

Outcomes

A standard that produces unreasonable outcomes (far too many or too few examinees pass) will not be viewed as credible regardless of the care with which it was derived. Therefore, it is important to collect data that support the fact that the stakeholders believe the standard is correct and that it has reasonable relationships with other markers of competence. For example, it would be supportive of the standard if the faculty generally believed that an appropriate number of students passed the summative assessment at the end of medical school. Moreover, it would be useful to gather evidence that the students who passed also performed well in the next phase of their training. In an assessment programme that continues over time, it is important to ensure that the stakeholders view the results as reasonable and that these results are related to the other indicators of proficiency.

Summary

A standard is a single score on a test that serves as the boundary between qualitatively different performances. These standards are an expression of judgement in the context of a particular assessment, its purpose and the wider social/professional environment. Consequently, methods for selecting them are a means for gathering a variety of perspectives, blending them together and expressing them as a single score. This chapter described the two types of standards, the characteristics that lead to their credibility, the more popular methods for setting standards and some efficient variations for use by medical educators.

References

- Berk RA: A consumer's guide to setting performance standards on criterion-referenced tests, *Review of Educational Research* 56:137-172, 1986.
- Brennan RL, Lockwood RE: A comparison of the Nedelsky and Angoff cutting score procedures using generalizability theory, *Applied Psychological Measurement* 4:219-240, 1980.
- Cusimano MD: Standard setting in medical education, *Academic Medicine* 71:s112-s120, 1996.
- Dauphinee WD, Blackmore DE, Smee S, et al: Using the judgements of physician examiners in setting standards for a national multi-center high stakes OSCE, *Advances in Health Sciences Education* 2:201-211, 1997.
- Livingston SA, Zieky MJ: *Passing Scores: A Manual for Setting Standards of Performance on Educational and Occupational Tests*, Princeton, NJ, 1982, Educational Testing Service.
- Kramer A, Muijtjens A, Jansen K, et al: Comparison of a rational and an empirical standard setting procedure for an OSCE. Objective structured clinical examinations, *Medical Education* 37: 132-139, 2003.
- McKinley DW, Boulet JR, Hambleton RK: A work-centered approach for setting passing scores on performance-based assessments, *Evaluation and the Health Professions* 28:349-369, 2005.
- Norcini JJ: Setting standards on educational tests, *Medical Education* 37:464-469, 2003.
- Norcini JJ, Guille RA: Combining tests and setting standards. In Norman G, van der Vleuten C, Newble D, editors: *International Handbook of Research in Medical Education*, Dordrecht, The Netherlands, 2002, Kluwer Academic, pp 811-834.
- Norcini JJ, Shea JA: The credibility and comparability of standards, *Applied Measurement in Education* 10:39-59, 1997.
- Shea JA, Bellini LM, McOwen KS, Norcini JJ: Setting standards for teaching evaluation data: An application of the contrasting groups method, *Teaching and Learning in Medicine* 21:82-86, 2009.
- Stern DT, Ben-David MF, Norcini JJ, et al: Setting school-level outcome standards, *Medical Education* 40:166-172, 2006.
- Wood TJ, Humphrey-Murto SM, Norman GR: Standard setting in a small scale OSCE: a comparison of the modified borderline-group method and the borderline regression method, *Advances in Health Sciences Education Theory and Practice* 11:115-122, 2006.

15 กรกฎาคม 2563

Grading

GRADING

รศ.นพ. เชิดศักดิ์ ไอร่มณิรัตน์
ภาควิชาศัลยศาสตร์ คณะแพทยศาสตร์ศิริราชพยาบาล
มหาวิทยาลัยมหิดล

1

“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 ข้อดีของการตัดสินผลการเรียนแบบอิงเกณฑ์และอิงกลุ่มได้
 - เลือกใช้วิธีการตัดเกรดที่เหมาะสมกับบริบทของสถาบันในการตัดสินผลการเรียนของนักศึกษา
 - 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 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.

Grading

5

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

Grading

6

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

Grading

7

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%

Grading

8

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.

Grading

9

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

Grading

10

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

www.menti.com

Grading

11

Standardization of Scores

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

Z = standard score

X = raw score

M = mean

SD = standard deviation

Grading

12

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.

13

Activity

- แบ่งกลุ่ม 4 กลุ่ม
- แต่ละกลุ่ม เป็นกรรมการรายวิชาหนึ่ง
- ให้เสนอแนวทางการจัดสัดส่วนคะแนน และนโยบายการตัดเกรดของวิชาที่รับผิดชอบ
- ตัวอย่าง
 - รายวิชา SISx4xx ศัลยศาสตร์ 4 หน่วยกิต
 - คะแนน mcq 30 + meq 20 + report 20 + osce 20 + class activity 10
 - Grading: Raw score, absolute grading : A > 85, B+ > 80, B > 75, C+ > 70, C > 65

(เวลา 8 นาที)

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.

Grading

15

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

Assessment

17

การตัดเกรด

อาจารย์ ดอกเตอร์ นายแพทย์เชิดศักดิ์ ไธรมณีนันท์ พ.บ.

ภาควิชาศัลยศาสตร์, คณะแพทยศาสตร์ศิริราชพยาบาล, มหาวิทยาลัยมหิดล, กรุงเทพมหานคร ๑๐๗๐๐.

เมื่อเดือนพฤษภาคมที่ผ่านมาทางฝ่ายการศึกษา ก่อนปริญญาได้จัดให้มีการสัมมนาระหว่างภาควิชาและโรงเรียนต่าง ๆ ในสังกัดของคณะแพทยศาสตร์ศิริราชพยาบาลเรื่องการตัดเกรด เพื่อเป็นการแลกเปลี่ยนแนวคิดและประสบการณ์ในการตัดเกรดของนักศึกษาที่อยู่ในความดูแลของคณะแพทยศาสตร์ศิริราชพยาบาล การสัมมนาดังกล่าวได้รับความร่วมมือเป็นอย่างดีจากโรงเรียนต่าง ๆ และภาควิชาทั้งระดับปริศลินิกและคลินิก มีการอภิปรายกันอย่างกว้างขวางถึงวิธีการที่ทางภาควิชาและโรงเรียนต่าง ๆ ใช้ในการตัดเกรด ปัญหาที่พบ ทัศนคติ และข้อบ่งชี้ต่าง ๆ ในการตัดเกรดที่อาจารย์ควรนำมาพิจารณา บทความนี้เป็นสรุปสาระสำคัญของการสัมมนาที่ผู้บันทึกเห็นว่าน่าจะเป็นประโยชน์ต่อคณาจารย์ในคณะแพทยศาสตร์ศิริราชพยาบาลเพื่อนำไปใช้ปรับปรุงวิธีการตัดเกรด

สภาพการณ์ในปัจจุบัน

จากการสำรวจแนวทางการตัดเกรดของนักศึกษาโดยภาควิชาและโรงเรียนต่าง ๆ ในคณะแพทยศาสตร์ศิริราชพยาบาลพบว่าวิธีการที่ใช้กันอยู่ในปัจจุบันมีความแตกต่างกันไป หลายภาควิชาส่งคะแนนให้งานบริการการศึกษาตัดเกรดให้ ภาควิชาและโรงเรียนส่วนใหญ่ใช้วิธีการที่เคยถือปฏิบัติกันมานาน โดยไม่ได้ทบทวนถึงหลักการ เหตุผล และความเหมาะสมของวิธีการตัดเกรดที่ใช้มาเป็นเวลานาน อาจารย์ผู้รับผิดชอบในการตัดเกรดจำนวนมากมีปัญหาหรือข้อ

สงสัยในวิธีการตัดเกรดแต่ไม่ทราบว่าจะไปหาคำตอบได้จากที่ใด โดยภาพรวมแล้วภาควิชาส่วนใหญ่ใช้วิธีการตัดเกรดนักศึกษาแบบอิงกลุ่มโดยกำหนดสัดส่วนนักศึกษาที่ควรจะได้เกรดต่าง ๆ ไว้ให้ค่อนข้างคงที่ในแต่ละกลุ่ม ภาควิชาทางปริศลินิกให้น้ำหนักของคะแนนสอบภาคทฤษฎีค่อนข้างมากในการคำนวณคะแนนรวมเพื่อนำมาตัดเกรด ในขณะที่ภาควิชาทางคลินิกมีคะแนนปฏิบัติค่อนข้างมาก โดยให้น้ำหนักอยู่ในช่วง ๓๐-๗๕% ของคะแนนรวม หลายภาควิชาใช้วิธีการตัดเกรดแบบผสมระหว่างการตัดเกรดแบบอิงเกณฑ์และอิงกลุ่ม โดยมีการกำหนดเกณฑ์ผ่านแบบอิงเกณฑ์ (ส่วนใหญ่ตั้งเกณฑ์ไว้ว่านักศึกษาต้องได้คะแนนไม่ต่ำกว่าร้อยละ ๖๐ จึงจะผ่าน) แต่ใช้การตัดสินแบบอิงกลุ่มในการให้เกรดว่าจะมีนักศึกษาคนใดได้ A, B, หรือ C ในการตัดเกรดแบบอิงกลุ่มในระดับคลินิกนั้น บางภาควิชาพิจารณาตัดสินเกรดปีละครั้งโดยนำคะแนนของนักเรียนทั้งชั้นปีมาตัดเกรดรวมกัน แต่ก็มีหลายภาควิชาที่พิจารณาตัดสินเกรดแยกตามกลุ่มย่อยของนักเรียนที่ชั้นปฏิบัติงานพร้อมกัน

หลักการพื้นฐานของการตัดเกรด

๑. คำจำกัดความ

การตัดเกรดเป็นการใช้วิจารณญาณของอาจารย์ผู้สอนในรายวิชาหนึ่ง ๆ ในการตัดสินว่านักเรียนหรือนักศึกษาประสบผลสำเร็จในการศึกษาวิชานั้น ๆ มากน้อยเพียงใดโดยอาศัยข้อมูลจากการประเมินผลการเรียน

๑๐๗

พฤษภาคม-สิงหาคม ๒๕๕๑, ปีที่ ๑, ฉบับที่ ๒

เวบบันทึกศิริราช

บทความทั่วไป

หรือการปฏิบัติงานของนักเรียนที่มีความถูกต้องเที่ยงตรง ในระยะเวลาที่กำหนด โดยผลที่ได้คือเกรดจะเป็นดัชนีที่ชี้วัดถึงความสำเร็จในการศึกษาของนักเรียน หรือนักศึกษาคนนั้น^๑

จากนิยามข้างต้นมีประเด็นที่น่าสนใจหลายประการด้วยกัน คือ

(๑) การตัดเกรดเป็นกระบวนการที่ต้องใช้วิจารณญาณของอาจารย์ในการตัดสิน อาจารย์ซึ่งเป็นผู้ที่เห็นความก้าวหน้าของนักศึกษาตลอดระยะเวลาที่อยู่ในความดูแลของอาจารย์ต้องทำการประมวลข้อมูลที่มีอยู่อย่างเป็นธรรมเพื่อตัดสินผล ไม่มีสูตรคำนวณ หรือวิธีการทางสถิติใดที่จะนำมาตัดสินเกรดให้นักศึกษาได้ แทนการใช้วิจารณญาณของอาจารย์ ข้อกำหนดหรือการคิดคำนวณทางสถิตินั้นเป็นเพียงเครื่องมือที่ช่วยให้อาจารย์สามารถตัดสินใจได้ง่ายขึ้น แต่อาจารย์ไม่จำเป็นต้องยึดติดกับการตัดคะแนนตามข้อกำหนดทางสถิติเสมอไป หากผลการตัดสินเกรดที่ได้รับจากวิธีการทางสถิติขัดแย้งกับสิ่งที่อาจารย์เห็นสมควร (เช่นเมื่อใช้การตัดเกรดแบบอิงกลุ่มแล้วพบว่านักศึกษาที่ได้คะแนน $\leq 60\%$ ถูกตัดสินให้สอบตก) อาจารย์สามารถทบทวนขั้นตอนในการรวมคะแนน และการตัดเกรดได้ เพื่อให้เกิดความเป็นธรรม ภายใต้ระเบียบและข้อกำหนดของภาควิชา โรงเรียน และ คณะฯ ซึ่งได้แจ้งให้นักศึกษาทราบ

(๒) ถึงแม้การตัดเกรดจะเป็นกระบวนการที่อยู่ในดุลยพินิจของอาจารย์ผู้สอนในรายวิชานั้น ๆ แต่กระบวนการดังกล่าวมิได้เป็นสิ่งที่ทำอย่างเลื่อนลอยปราศจากหลักการ การตัดเกรดนั้นต้องวางอยู่บนพื้นฐานของข้อมูลการประเมินผลที่มีความถูกต้องและเที่ยงตรง หากคะแนนดิบที่นำมาใช้ในการตัดเกรดเป็นคะแนนที่ได้มาจากการประเมินผลที่ไม่เหมาะสม หรือมีความคลาดเคลื่อนสูง ก็เป็นการยากที่จะทำให้การตัดเกรดมีความเที่ยงตรงและเป็นธรรมกับนักศึกษา

(๓) ผลลัพธ์ของการตัดเกรดคือเกรด ซึ่งเป็นดัชนีสรุปผลสำเร็จทางการศึกษาโดยรวมของนักศึกษาคณะหนึ่ง ๆ แต่เนื่องจากว่าเกรดมีที่มาจากหลายแหล่งในหลายวิชา ซึ่งแต่ละแหล่งของข้อมูลคะแนนดิบก็มี

วัตถุประสงค์ และวิธีการที่แตกต่างกัน เมื่อนำคะแนนมารวมกันก็จะทำให้ความหมายของคะแนนดิบนั้นสูญเสียไป ตัวอย่างเช่น หากในรายวิชาหนึ่งอาจารย์ตัดเกรดโดยรวมคะแนนจากการสอบ multiple-choice questions, Objective Structured Clinical Examination (OSCE), และคะแนนรายงานผู้ป่วย หากมีนักศึกษาทำคะแนนสอบ multiple-choice questions และ OSCE ได้ดี แต่ไม่ส่งรายงานผู้ป่วยเลย เมื่อเอาคะแนนมารวมกันแล้วคะแนนสอบที่สูงก็จะบดบังปัญหาในการเขียนรายงานของนักศึกษาคนนั้นไป ดังนั้นอาจารย์จะใช้เกรดที่เป็นตัวอักษรตัวเดียวมาสรุปว่านักศึกษาคนใดคนหนึ่งนั้นดีหรือไม่ดีในทุกด้านตามเกรดที่ได้รับนั้นอาจไม่ถูกต้อง อาจารย์ต้องย้อนกลับไปดูรายละเอียดของคะแนนดิบด้วยจึงจะได้ข้อมูลที่สมบูรณ์เกี่ยวกับความรู้ ความสามารถของนักศึกษา

๒. ประโยชน์ของการตัดเกรด

ข้อมูลที่ได้จากการตัดเกรดของนักศึกษาสามารถนำไปใช้ประโยชน์ได้หลายอย่างได้แก่

(๑) อาจารย์สามารถนำเกรดที่ได้ไปใช้ในการบริหารการศึกษา เพื่อตัดสินว่านักศึกษาคนใดควรได้เลื่อนชั้น นักศึกษาคนใดควรได้รับรางวัลเรียนดี หรือนักศึกษาคนใดควรจะได้รับปริญญาเกียรตินิยม การใช้ประโยชน์ของเกรดในลักษณะนี้เป็นสิ่งที่เห็นได้ชัดเจนที่สุดและสถานศึกษาในทุกระดับใช้อยู่

(๒) เกรดที่นักศึกษาได้รับนี้ หากมีการแจ้งให้ผู้ปกครองของนักศึกษาทราบด้วยก็จะเป็นการรายงานความก้าวหน้าในการศึกษาของนักศึกษาให้ผู้ปกครองทราบ เป็นการสื่อสารระหว่างอาจารย์กับผู้ปกครอง หากมีปัญหาในการเรียน ผู้ปกครองก็จะเห็นว่าเกรดไม่ดี อาจมีการร่วมมือกันกับอาจารย์ในการให้ความช่วยเหลือเพื่อแก้ปัญหาในการเรียนของนักศึกษาผู้นั้น ในปัจจุบันจัดว่า ทางคณะแพทยศาสตร์ศิริราชพยาบาลใช้ประโยชน์ในด้านนี้น้อยมาก เนื่องจากเราตัดเกรดเป็นเพียงตัวอักษร A, B, C, D, F โดยไม่มีข้อมูลที่เป็นประโยชน์ในการปรับปรุงตัวของนักศึกษาแต่อย่างใด ผู้ปกครองเพียงแต่ทราบว่านักศึกษาเรียนวิชานี้ได้ดี แต่เรียนอีกวิชาหนึ่งได้ไม่ดี แต่ไม่ทราบว่าไม่ดีในด้านใด ผู้ปกครอง

เวบบิ้นทีกีรียา

บทความทั่วไป

จะให้ความช่วยเหลืออย่างไร ที่คะแนนของบุตรหลานเขาไม่ดีขึ้นเป็นเพราะเนื้อหายาก หรือข้อสอบยาก หรือนักศึกษาไม่ใส่ใจเรียน ขาดเรียน ไม่ส่งรายงาน

(๓) หากการรายงานเกรดของนักศึกษาทำอย่างมีประสิทธิภาพ ข้อมูลที่นักศึกษาได้รับสามารถนำไปใช้เป็นประโยชน์ในการปรับปรุงการเรียนของตนได้ด้วย คณะแพทยศาสตร์ศิริราชพยาบาลยังไม่ได้ใช้ประโยชน์ด้านนี้ในปัจจุบัน การรายงานเกรดที่จะใช้ประโยชน์ในลักษณะนี้ได้ควรต้องมีการชี้แจงให้นักศึกษาทราบว่าเขามีข้อดี หรือข้อด้อยอย่างไรบ้าง เขาทำคะแนนด้านใดได้ดี สิ่งใดที่ทำได้ดีอยู่แล้วนักศึกษาจะได้มีกำลังใจทำได้ดีขึ้นต่อไป ทักษะหรือความสามารถด้านใดที่เขายังบกพร่องอยู่ ก็ควรมีการบอกให้นักศึกษาทราบเพื่อที่เขาจะได้รู้ว่าควรต้องปรับปรุงตนอย่างไร

๓. วิธีการที่ใช้ในการตัดเกรด

สามารถแบ่งวิธีการตัดเกรดออกได้เป็น ๒ วิธีใหญ่ ๆ คือ

(๑) การตัดเกรดแบบอิงเกณฑ์ (Absolute grading) เป็นการกำหนดว่าหากนักศึกษาทำคะแนนได้ถึงเท่าไรจะได้เกรดเท่าไร ตัวอย่างเช่น กำหนดว่าหากคะแนนตั้งแต่ ๙๐ คะแนนขึ้นไปได้ A หากคะแนนอยู่ในช่วง ๘๐ - ๘๙ คะแนนได้ B หากคะแนนอยู่ในช่วง ๗๐ - ๗๙ คะแนนได้ C หากคะแนนอยู่ในช่วง ๖๐ - ๖๙ คะแนนได้ D และหากคะแนนต่ำกว่า ๖๐ ได้ F เป็นต้น^{๒-๓}

(๒) การตัดเกรดแบบอิงกลุ่ม (Relative grading) เป็นการกำหนดสัดส่วนของนักศึกษาที่จะได้เกรดต่าง ๆ ให้คงที่ การที่นักศึกษาคนใดจะได้เกรดอะไรนั้นให้เทียบกับเพื่อนในกลุ่มที่ตัดเกรดด้วยกัน หากทำได้ดีกว่าเพื่อนจะได้เกรดดี หากทำได้ดีไม่ดีกว่าเพื่อนจะได้เกรดไม่ดี ตัวอย่างเช่นกำหนดว่านักเรียนที่ได้คะแนนสูงสุด ๑๕% ของกลุ่มจะได้ A ผู้ที่คะแนนรองลงไปอีก ๒๕% จะได้ B ผู้ที่คะแนนรองลงไปอีก ๔๕% จะได้ C และผู้ที่คะแนนต่ำลงไปอีก ๑๐% จะได้ D และนักศึกษาที่ได้คะแนนต่ำสุด ๕% ของกลุ่มจะได้ F เป็นต้น^{๒-๓}

ในปัจจุบันการตัดเกรดทั้ง ๒ วิธีเป็นที่ยอมรับกันทั่วไป มีทั้งสถาบันที่ใช้การตัดเกรดแบบอิงเกณฑ์ และสถาบันที่ใช้การตัดเกรดแบบอิงกลุ่ม ในบทบาท

หน้าที่ของอาจารย์ซึ่งต้องพิจารณาให้เกρνนักศึกษาที่ควรต้องเข้าใจว่าวิธีการตัดเกรดที่ใช้อยู่มีข้อดี และข้อเสียอย่างไร เหมาะสมกับระบบการเรียนการสอนที่อาจารย์จัดให้นักศึกษาหรือไม่อย่างไร ในที่นี้ก็จะขอสรุปข้อดี และข้อเสียของการตัดเกรดทั้ง ๒ วิธี

การตัดเกรดแบบอิงเกณฑ์นั้นข้อดีคือเกρνที่นักเรียนได้สามารถระบุได้ชัดเจนว่านักศึกษามีความรู้ความสามารถดีหรือไม่ มากน้อยเพียงใด โดยไม่ขึ้นกับว่านักศึกษาอยู่ในกลุ่มเพื่อนที่เรียนเก่งหรือไม่^๒ นักศึกษาที่ได้เกรด A ก็แสดงว่ามีความรู้ ความสามารถผ่านเกณฑ์ขั้นสูงที่อาจารย์กำหนดไว้ ซึ่งหากเกณฑ์ตัดสินคงที่ตลอดระยะเวลาหลายปี เกρνที่นักเรียนได้ในแต่ละปีก็สามารถเทียบกันได้ การตัดเกรดวิธีนี้ไม่มีการจำกัดจำนวนของนักเรียนที่จะได้แต่ละเกรด ดังนั้นหากนักเรียนทุกคนทำคะแนนได้ดีเยี่ยม ทุกคนในชั้นก็มีสิทธิ์ที่จะได้ A โดยไม่ต้องแข่งขันกับเพื่อนในกลุ่ม ดังนั้นการตัดเกรดวิธีนี้จึงเป็นการส่งเสริมให้นักศึกษาช่วยกันเรียน แต่วิธีการตัดเกรดแบบอิงเกณฑ์นี้ก็อาจถูกวิจารณ์ได้ว่าเกณฑ์ที่ตั้งไว้นั้นไม่มีที่มาที่ชัดเจน^๒ อาจารย์ใช้ความรู้สึก หรือความเห็นส่วนตัวในการกำหนดเกณฑ์ ซึ่งอาจไม่เหมาะสม อย่างไรก็ตาม หากเกณฑ์ที่ตั้งขึ้นนั้นวางอยู่บนพื้นฐานข้อมูลคะแนนสอบที่ผ่านมาของนักศึกษาหลายปี ก็น่าจะเป็นหลักฐานสนับสนุนความน่าเชื่อถือของเกณฑ์ที่เพียงพอ เนื่องจากอาจารย์ได้ศึกษาแล้วว่าในระยะหลายปีที่นักศึกษาที่เยี่ยมยอดนั้นควรมีคะแนนอยู่ในช่วงใด นอกจากนี้วิธีการตัดเกรดแบบอิงเกณฑ์ยังอาจประสบปัญหาในการแปลผลเกรดเทียบระหว่างกลุ่มนักศึกษา หรือระหว่างปีการศึกษาในบางกรณี เนื่องจากคะแนนที่ได้นั้นไม่ได้ขึ้นกับความรู้ความสามารถของนักศึกษาอย่างเดียว แต่มีปัจจัยภายนอกมารบกวนคะแนนของนักศึกษาได้ เช่น ความยากง่ายของข้อสอบ หรือประสิทธิภาพในการสอนของอาจารย์^๒ ดังนั้นหากนักศึกษาได้เกรดสูง อาจเป็นได้ว่าข้อสอบที่ใช้สอบนั้นง่ายกว่ากลุ่มอื่น หากต้องการให้เกณฑ์การให้เกρνมีมาตรฐานที่คงที่ และสามารถเปรียบเทียบผลการศึกษานักศึกษาที่เรียนไม่พร้อมกันได้ อาจารย์ควรมีกระบวนการในการควบคุมระดับความยากง่ายของ

เวบบิ้นทีกีรียา

บทความทั่วไป

ข้อสอบให้คนที่ หรือมีการปรับคะแนนนักศึกษาตามระดับความยากง่ายของข้อสอบ ดังตัวอย่างที่เห็นได้จากการสอบ TOEFL (Test of English as a Foreign Language) ซึ่งจะพบว่าผู้เข้าสอบทำข้อสอบคนละชุดกัน แต่สุดท้ายทาง ETS (Educational Testing Service) ก็มีการปรับคะแนนให้อยู่บนมาตรฐานเดียวกัน สามารถเทียบผลสอบกันได้^๕

ข้อดีของการตัดสินแบบอิงกลุ่มก็คือสามารถควบคุมจำนวนของนักศึกษาที่ได้เกรดต่าง ๆ ได้ค่อนข้างคงที่ ทำให้การบริหารการศึกษาทำได้ง่าย ไม่เกิดเหตุการณ์ที่ทำให้ต้องมีการปรับเปลี่ยนกิจกรรมการเรียนการสอนอย่างคาดไม่ถึง เช่นต้องจัดสอบซ่อมให้นักศึกษาทั้งชั้นปี หรือต้องจัดตารางให้นักศึกษาครั้งชั้นปีขึ้นปฏิบัติงานเพิ่มในภาควิชาใดภาควิชาหนึ่ง แต่การตัดสินแบบอิงกลุ่มนี้มีข้อเสียหลายประการด้วยกันคือการที่อาจารย์ไม่สามารถเทียบระดับความรู้ความสามารถของนักศึกษาที่อยู่ต่างรุ่นหรือกลุ่มได้ กล่าวคือนักศึกษาที่ได้ A ในรุ่นปัจจุบัน หากนำไปตัดเกรดกับนักศึกษาในกลุ่มอื่นอาจได้เกรด B ก็ได้ นั่นคือเกรดที่นักศึกษาจะได้รับนอกจากจะขึ้นกับว่านักศึกษาทำได้ดีมากน้อยเพียงใดแล้ว ยังขึ้นกับคะแนนของเพื่อนในกลุ่มที่ทำเกรดรวมกันด้วย ซึ่งจะนำไปสู่การสร้างบรรยากาศการเรียนที่มีการแข่งขันกัน นักศึกษาไม่ช่วยกันเรียนเท่าที่ควรเนื่องจากเกรงว่าหากช่วยเพื่อนแล้วจะทำให้คะแนนของเพื่อนสูงขึ้นซึ่งอาจส่งผลให้ตัวนักศึกษาเองได้เกรดต่ำ^{๕-๖}

โดยทั่วไปแล้วหากไม่ได้มีการนำเกรดไปใช้ในการคัดเลือกนักเรียนเข้าในโควตาพิเศษซึ่งมีที่นั่งจำกัด (เช่นให้นักเรียนที่ได้เกรด A ได้ไปเข้าประชุมวิชาการที่ต่างประเทศ ให้นักเรียนที่ได้เกรด D ทุกคนเข้าในโปรแกรมติวพิเศษ) แนะนำให้ใช้การตัดสินแบบอิงเกณฑ์ เนื่องจากเกรดที่ได้เป็นตัวบอกถึงระดับความรู้ความสามารถของนักศึกษาได้โดยไม่เกี่ยวข้องว่านักศึกษายู่ในกลุ่มที่เก่งหรือไม่เก่ง หากนักศึกษาทุกคนในชั้นเรียนมีความรู้ความสามารถไม่เพียงพอที่จะไปดูแลผู้ป่วยในสาขาวิชานั้น ๆ นักศึกษาทุกคนก็ควรจะถูกตัดสินให้ไม่ผ่านและมีการเรียนเสริมเพื่อให้ความรู้ถึงเกณฑ์

มาตรฐานประกอบวิชาชีพเวชกรรม หากนักศึกษาครั้งชั้นมีความรู้ความสามารถดีมาก ก็สามารถตัดสินให้นักศึกษาทั้งครั้งชั้นได้เกรด A ได้ ไม่จำเป็นต้องตัดสินให้นักศึกษาที่ได้คะแนน ๗๐ - ๘๐% ต้องขึ้นปฏิบัติงานเพิ่มเติมเนื่องจากนักศึกษาคนดังกล่าวอยู่ในกลุ่มเพื่อนที่เก่งมาก ทุกคนได้คะแนนอยู่ในช่วง ๘๕ - ๙๕% หากจะใช้การตัดสินแบบอิงกลุ่ม แนะนำให้ตัดสินได้หรือตกด้วยวิธีการอิงเกณฑ์ก่อน แล้วจึงใช้การตัดสินอิงกลุ่มเพื่อแยกนักเรียนออกเป็นกลุ่มที่ได้เกรด A, B, หรือ C^{๖-๗}

๔. การรวมคะแนน

ข้อมูลสำคัญที่ใช้ในการตัดสินคือคะแนนของนักเรียนซึ่งได้มาจากการประเมินผลการเรียนหลายวิธีเข้าด้วยกัน เช่น การสอบ multiple-choice questions, OSCE, คะแนนรายงานผู้ป่วย เป็นต้น การนำคะแนนจากหลายวิธีการประเมินมารวมกันนี้ต้องทำอย่างเหมาะสมเพื่อให้คะแนนรวมที่ได้มีความถูกต้องและเป็นธรรม หลักการพื้นฐานคือคะแนนรวมที่ได้นั้นจะมีน้ำหนักคะแนนของการสอบแต่ละส่วนเท่าไรนั้นขึ้นกับสัดส่วนของคะแนนที่อาจารย์กำหนด และการกระจายตัวของคะแนนในการสอบครั้งนั้น โดยทั่วไปอาจารย์มักคำนึงถึงปัจจัยแรกเท่านั้น และทำการรวมคะแนนโดยทำคะแนนเต็มของการสอบแต่ละครั้งให้เท่ากับสัดส่วนของคะแนนที่ต้องการในคะแนนรวม แล้วทำการบวกคะแนนทั้งหมดเข้าด้วยกันให้ได้คะแนนเต็ม ๑๐๐ แล้วนำคะแนนรวมที่ได้ไปใช้ในการตัดสิน ปัญหาที่จะพบได้ในการรวมคะแนนวิธีนี้คือ คะแนนสอบทางทฤษฎีนั้นมักมีการกระจายตัวของคะแนนมาก (มีค่า standard deviation (SD) สูง) ในขณะที่คะแนนปฏิบัตินั้นมักไม่ค่อยมีความแตกต่างของคะแนน นักศึกษามักมีคะแนนปฏิบัติที่ใกล้เคียงกันมาก หากอาจารย์กำหนดให้คะแนนภาคทฤษฎีและปฏิบัติมีน้ำหนักเท่ากัน โดยทำคะแนนสอบทฤษฎีให้เต็ม ๕๐ คะแนน และทำคะแนนปฏิบัติให้เต็ม ๕๐ คะแนน แล้วรวมคะแนนเข้าด้วยกัน คะแนนภาคทฤษฎีจะเป็นตัวกำหนดเกรดของนักเรียน โดยที่คะแนนปฏิบัติส่งผลน้อยมาก

การรวมคะแนนที่ถูกต้องนั้นต้องมีการปรับให้คะแนนการสอบย่อยแต่ละครั้งมีคะแนนเต็มเท่ากัน และ

เวบบินทีกีรียา

บทความทั่วไป

มีการกระจายตัวของคะแนนเหมือนกันเสียก่อน โดยการแปลงคะแนนดิบเป็นคะแนนมาตรฐาน (Standardized score) แล้วจึงคูณคะแนนมาตรฐานดังกล่าวด้วยน้ำหนักที่ต้องการ แล้วจึงทำการรวมคะแนน^๓ วิธีการคิดคะแนนมาตรฐานนั้นสามารถทำได้โดยง่าย โดยใช้สูตรต่อไปนี้

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

เมื่อ Z คือคะแนนมาตรฐาน, x คือคะแนนดิบ, M คือคะแนนเฉลี่ยของการสอบนั้น, และ SD คือค่าเบี่ยงเบนมาตรฐานของคะแนนสอบนั้น^๓ คะแนนมาตรฐานที่ได้ออกมาจากการคำนวณตามสูตรนี้จะมีค่าเฉลี่ยเท่ากับ ๐ และมีค่าเบี่ยงเบนมาตรฐานเท่ากับ ๑ หากอาจารย์ต้องการปรับคะแนนให้ไม่มีคะแนนติดลบ และไม่มีคะแนนเป็นจุดทศนิยมสามารถแปลงเป็นคะแนน T score โดยใช้สูตร

$$T = 10Z + 50$$

คะแนน T score นี้จะมีค่าเฉลี่ยเท่ากับ ๕๐ และมีค่าเบี่ยงเบนมาตรฐานเท่ากับ ๑๐

ยกตัวอย่างเช่นอาจารย์ต้องการตัดเกรดโดยรวมคะแนนจากการสอบ ๓ ครั้ง แต่ละครั้งให้มีสัดส่วนเป็น ๓๐% ของคะแนนรวม และคะแนนรายงานผู้ป่วยอีก ๑๐% อาจารย์ควรคิดคะแนนดังนี้

(๑) คำนวณ T score จากคะแนนดิบทั้ง ๔ ครั้งได้เป็น T_{exam1} , T_{exam2} , T_{exam3} และ T_{report}

(๒) คูณ T score แต่ละส่วนด้วยน้ำหนักคะแนนตามความเหมาะสมแล้วรวมคะแนนเข้าด้วยกัน

$$\text{Total score} = 3 T_{exam1} + 3 T_{exam2} + 3 T_{exam3} + 1 T_{report}$$

(๓) เทียบบัญญัติไตรยางศ์ให้คะแนนเต็มเป็น ๑๐๐ คะแนน แล้วนำคะแนนดังกล่าวไปใช้ในการตัดเกรดตามเกณฑ์ที่ทางภาควิชาตั้งไว้

๕. ความผิดพลาดของการตัดเกรด

การวัดผลการศึกษานั้นสามารถเกิดความผิดพลาดขึ้นได้เช่นเดียวกันกับการวัดอื่น ๆ เช่นการวัด

ระดับน้ำตาลในเลือด หรือการวัดความดันโลหิต ดังนั้นคะแนนที่นักศึกษาได้จากการสอบแต่ละครั้งก็เกิดความผิดพลาดคลาดเคลื่อนได้จากปัจจัยต่าง ๆ เช่นความไม่เที่ยงตรงของเครื่องมือวัดผล (ข้อสอบ) เป็นต้น ดังนั้นการสรุปผลการศึกษานักศึกษาเป็นเกรดนั้นอาจารย์ก็ต้องคำนึงด้วยว่ามีโอกาสเกิดความผิดพลาดขึ้นได้ โดยความผิดพลาดของการตัดสินผลนั้นสามารถเกิดขึ้นได้ ๒ ลักษณะด้วยกัน คือ

(๑) False positive หมายถึง การตัดสินให้นักศึกษาที่สมควรสอบตกให้สอบผ่าน

(๒) False negative หมายถึง การตัดสินให้นักศึกษาที่สมควรสอบผ่านให้สอบตก

หากการสอบที่กำลังพิจารณาตัดสินผลนั้นมีความสำคัญต่อความปลอดภัยของสังคม หากตัดสินให้ผู้ที่ไม่มีความรู้ความสามารถเพียงพอผ่านไปได้อาจเกิดผลเสียต่อสังคม เช่นการตัดสินให้ใบอนุญาตประกอบวิชาชีพเวชกรรม อาจารย์จะต้องระมัดระวังให้เกิด false positive น้อยที่สุด แต่หากการสอบนั้นเป็นการสอบย่อยซึ่งจะมีการสอบอื่น ๆ มาตรวจสอบนักศึกษาอีกหลายครั้งดังเช่นการสอบทั่วไปที่ใช้ในคณะฯ อาจารย์อาจยอมรับ false positive ได้พอสมควร เพราะหากนักเรียนไม่มีความรู้เพียงพอจริงเขาก็คงจะไม่ผ่านการสอบอื่น ๆ ที่จะตามมา แต่การเกิด false negative จะทำให้นักเรียนเสียกำลังใจ และเสียประวัติการศึกษาไป

การจะพิจารณาว่าการตัดสินผลการสอบนั้นจะเกิด false positive หรือ false negative มากน้อยเพียงใดทำได้โดยการคำนวณ standard error of measurement (SEM) กล่าวคือในการสอบแต่ละครั้งคะแนนที่นักศึกษาได้รับนั้นประกอบไปด้วยคะแนนที่แท้จริง (true score) ของนักศึกษาคนนั้น กับความคลาดเคลื่อน (error) ที่เกิดจากความไม่เที่ยงตรงของเครื่องมือที่ใช้วัดผล หากเราสามารถทำการวัดผลซ้ำหลาย ๆ ครั้งในรูปแบบเดิมกับนักศึกษาคนนั้น โดยที่ปัจจัยทุกอย่างถูกควบคุมให้คงที่ (นักศึกษามีความรู้เท่าเดิม ข้อสอบมีความยากเท่าเดิม) คะแนนเฉลี่ยที่ได้จะเท่ากับคะแนนที่แท้จริงของนักศึกษาคนนั้น และคะแนนจะมีการกระจายตัวแบบ normal distribution โดยมีค่าเบี่ยงเบนมาตรฐาน

เวบบิ้นทีกศิริราช

บทความทั่วไป

(standard deviation) เท่ากับ SEM เราสามารถคำนวณค่า SEM ได้จากสูตร

$$SMD = SD\sqrt{(1-r)}$$

เมื่อ r คือค่า internal consistency reliability ของคะแนนสอบครั้งนั้น ๆ^๒

อาศัยความรู้พื้นฐานทางสถิติเราก็จะได้ว่า โอกาสที่คะแนนที่แท้จริงจะอยู่ในช่วง score ± 1 SEM เท่ากับ ๖๘% และโอกาสที่คะแนนที่แท้จริงจะอยู่ในช่วง score ± 2 SEM เท่ากับ ๙๕% ดังนั้นหากอาจารย์ต้องการตัดสินใจให้นักศึกษาสอบผ่านรายวิชาหนึ่งโดยให้ความมั่นใจว่าจะมี false negative เกิดขึ้นไม่เกิน ๕% ก็ต้องปรับลดเกณฑ์สอบผ่านลงจากค่าที่ตั้งไว้อีก ๒ SEM

ข้อกำหนดเกี่ยวกับการตัดเกรดของนักศึกษาใน คณะแพทยศาสตร์ศิริราชพยาบาล

นอกจากหลักการทางทฤษฎีแล้ว อาจารย์ยังต้องทราบถึงข้อบังคับของคณะแพทยศาสตร์ศิริราชพยาบาล และของมหาวิทยาลัยมหิดลด้วยเพื่อให้การตัดเกรดของนักศึกษาเป็นไปอย่างถูกต้อง ในที่นี้ผู้นิพนธ์จะขอสรุปประเด็นสำคัญที่เกี่ยวกับการตัดเกรดจากข้อบังคับมหาวิทยาลัยมหิดล^๓ และประกาศของคณะแพทยศาสตร์ศิริราชพยาบาล^๔

รายวิชาส่วนใหญ่ที่จัดสอนในคณะแพทยศาสตร์ศิริราชพยาบาลทำการตัดเกรดโดยใช้สัญลักษณ์ตัวอักษร A, B+, B, C+, C, D+, D, และ F ซึ่งมีค่าประจำเป็น ๔.๐, ๓.๕, ๓.๐, ๒.๕, ๒.๐, ๑.๕, ๑.๐, และ ๐ ตามลำดับ มีเพียงไม่กี่รายวิชาที่ทางคณะฯ พิจารณาแล้วว่าไม่สมควรจำแนกผลการศึกษออกเป็นระดับ จะตัดสินผลเพียงว่าผ่าน หรือไม่ผ่าน โดยให้แสดงผลเป็น S (Satisfactory) หรือ U (Unsatisfactory) ตามลำดับ

นอกจากสัญลักษณ์ที่ใช้เป็นประจำข้างต้นแล้ว ยังมีสัญลักษณ์ที่อาจารย์สามารถใช้ในการรายงานผลการศึกษาอื่น ๆ อีกได้แก่ I (Incomplete) สำหรับรายวิชาที่ยังไม่สามารถตัดสินผลได้เนื่องจากนักศึกษาไม่ส่งงาน หรือไม่สอบเพราะเจ็บป่วย หรือด้วยเหตุสุดวิสัย,

P (In progress) สำหรับรายวิชาที่ยังไม่สิ้นสุดการเรียนการสอนเนื่องจากมีการเรียนต่อเนื่องมากกว่า ๑ ภาคการศึกษา, W (Withdrawal) สำหรับรายวิชาที่นักศึกษาขอถอนตัวจากการศึกษาหรือถูกสั่งพักการศึกษา, AU (Audit) สำหรับรายวิชาที่นักศึกษาเข้าเรียนโดยไม่นับหน่วยกิต, และ X (No report) สำหรับรายวิชาที่คณะฯ ไม่ได้รับรายงานผลการประเมิน

นักศึกษาจะได้รับการประเมินว่าผ่านในรายวิชาใดได้จะต้องได้รับเกรดที่มีค่าประจำตั้งแต่ ๒.๐ ขึ้นไป หรือได้เกรด S (Satisfactory) หากนักศึกษาได้เกรด D, D+, หรือ U ถือว่าสอบไม่ผ่านในรายวิชานั้น นักศึกษาแพทย์ที่สอบไม่ผ่านจะสามารถสอบแก้ตัวได้เมื่อมีเกรดเฉลี่ยสะสมในปีการศึกษานั้นไม่ต่ำกว่า ๒.๐ หากมีเกรดเฉลี่ยประจำปีต่ำกว่า ๒.๐ จะต้องลงทะเบียนเรียนใหม่ สำหรับนักศึกษาในหลักสูตรอื่นจะสามารถสอบแก้ตัวได้หรือไม่ขึ้นกับดุลยพินิจของกรรมการประจำหลักสูตร

หากนักศึกษาที่ได้เกรด D หรือ D+ สอบแก้ตัวผ่านจะได้รับเกรด C ในรายวิชานั้น แต่หากสอบแก้ตัวไม่ผ่านจะได้รับเกรด F นักศึกษาที่ได้รับการตัดสินให้ได้เกรด F จะต้องลงทะเบียนเรียนรายวิชานั้นซ้ำ

การพัฒนาระบบการตัดเกรด

จากการสัมมนาระหว่างภาควิชาที่จัดขึ้นนี้พบว่าภาควิชา และโรงเรียนต่าง ๆ ในสังกัดคณะแพทยศาสตร์ศิริราชพยาบาลยังมีโอกาสที่จะพัฒนาระบบการตัดเกรดให้ดีขึ้นได้ในประเด็นต่าง ๆ ต่อไปนี้

(๑) พิจารณาหาแนวทางพัฒนาการรายงานผลการศึกษาให้มีข้อมูลที่มากขึ้นเพื่อเป็นประโยชน์ต่อการพัฒนาหรือปรับปรุงตัวของนักศึกษา กล่าวคือ นอกจากจะตัดสินผลการศึกษาเป็นเกรดให้นักศึกษาแล้ว หากทางภาควิชาหรือโรงเรียนสามารถให้ข้อมูลแก่นักศึกษาเพิ่มเติมว่าเหตุใดเขาจึงได้เกรดดังกล่าว เขาทำคะแนนส่วนใดได้ดี เขามีปัญหาในคะแนนส่วนใด ก็จะเป็นประโยชน์ต่อนักศึกษามากในการที่เขาจะได้ไปพัฒนาปรับปรุงตนให้ดีขึ้นในส่วนที่เขาทำคะแนนได้ไม่ดีนัก

(๒) พิจารณาทบทวนวิธีการตัดเกรดที่ใช้อยู่ว่า

๑๑๒

พฤศจิกายน-สิงหาคม ๒๕๕๑, ปีที่ ๑, ฉบับที่ ๒

เวบบิ้นทีกีธีรธา

บทความทั่วไป

เป็นแบบอิงเกณฑ์ หรืออิงกลุ่ม วิธีที่ใช้อยู่ในมีความเหมาะสมกับรูปแบบการเรียนการสอน หรือการวัดผลที่ทางภาควิชาหรือโรงเรียนใช้อยู่หรือไม่ หากเป็นวิธีการที่มีความเหมาะสมอยู่แล้วก็ควรจะคงไว้ต่อไป แต่หากไม่เหมาะสมก็พิจารณาปรับเปลี่ยน

(๓) ทบทวนวิธีประเมินผลนักศึกษาที่ใช้อยู่ (เช่น multiple-choice questions, multiple essay questions, OSCE, และ clinical performance ratings เป็นต้น) ว่ามีความถูกต้อง เทียบตรงมากน้อยเพียงใด หากวิธีประเมินผลที่ใช้นั้นให้คะแนนที่มีความเที่ยงตรง เชื่อถือได้ ก็ควรคงวิธีการประเมินนั้นไว้ หากวิธีการประเมินผลบางวิธีมีปัญหา ให้คะแนนที่ไม่เที่ยงตรง มี reliability ต่ำ (มี SEM สูง) ก็จะต้องหาวิธีปรับปรุงวิธีการประเมินผลดังกล่าวให้ดีขึ้น

(๔) ทบทวนวิธีการรวมคะแนนดิบจากแหล่งต่าง ๆ เป็นคะแนนรวม หากใช้วิธีการรวมคะแนนที่ถูกต้องอยู่แล้วก็ดำเนินการต่อไป แต่หากวิธีการรวมคะแนนที่ใช้ในปัจจุบันเป็นวิธีการที่ไม่เหมาะสมก็ควรพิจารณาปรับเปลี่ยน โดยทำการแปลงคะแนนดิบเป็น T score ก่อนรวมคะแนน

(๕) พิจารณาถึงโอกาสเกิดความผิดพลาดในการตัดสินผลการศึกษาให้ผ่าน หรือไม่ผ่านแก่นักศึกษาว่าทางภาควิชาหรือโรงเรียนยอมรับอัตราการเกิด false positive หรือ false negative ในการตัดสินผลมากน้อยเพียงใด แล้วทำการปรับเกณฑ์ผ่านโดยใช้ค่า SEM ตามความเหมาะสม

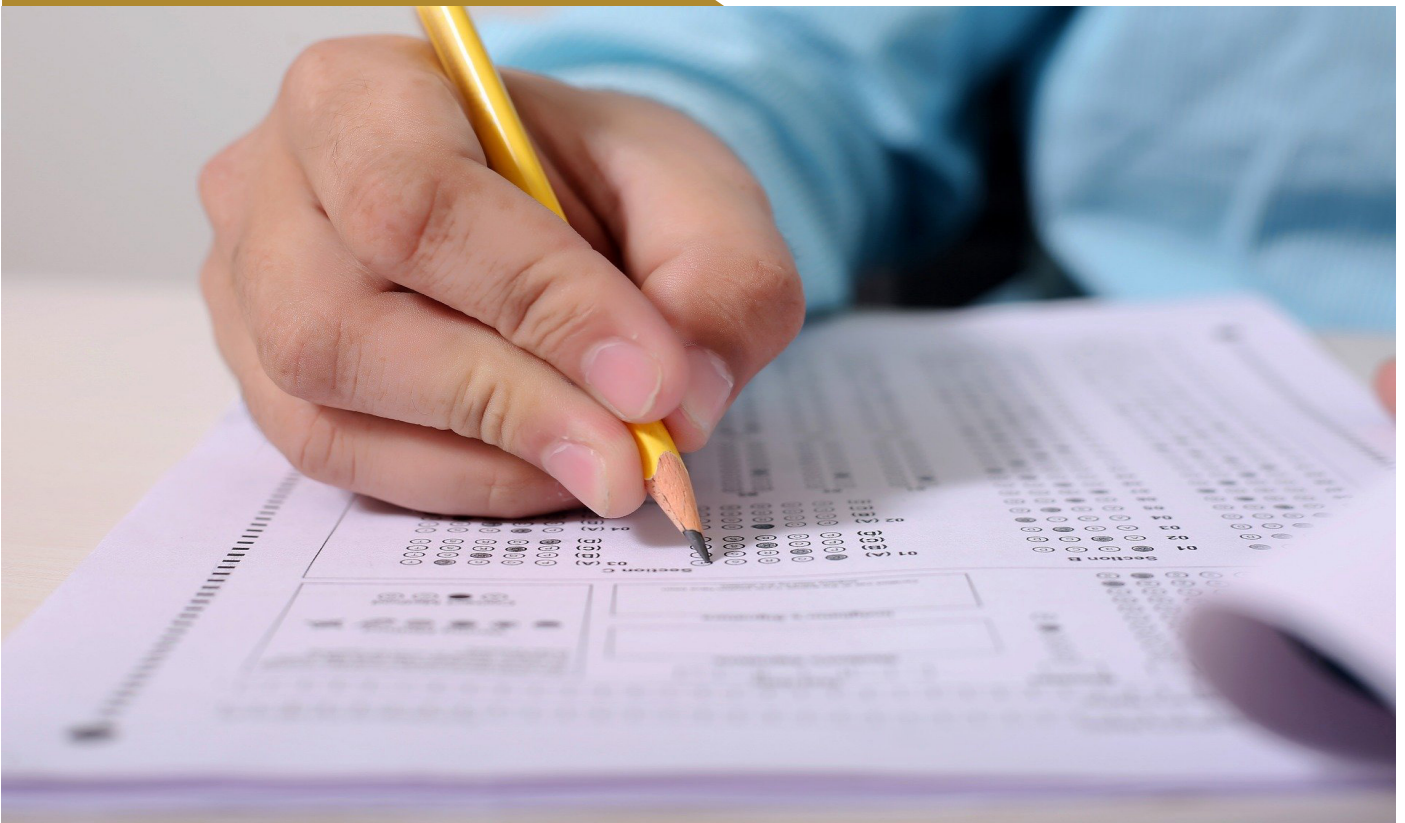
เอกสารอ้างอิง

๑. Guskey TR, Bailey JM. Developing grading and reporting systems for student learning. Thousand Oaks, CA: Corwin, 2001.
๒. Gronlund NE. Assessment of student achievement, 7th ed. Boston: Allyn & Bacon, 2003.
๓. Linn RL, Miller MD. Measurement and assessment in teaching, 9th ed. Upper Saddle River, NJ: Prentice Hall, 2004.
๔. Educational Testing Service. TOEFL iBT Tips: How to prepare for the TOEFL iBT. Princeton, NJ: ETS, 2007.
๕. Gray K. Why we will lose: Taylorism in America's high schools. Phi Delta Kappan 1993;74:370-4.
๖. Haladyna TM. A complete guide to student grading. Boston: Allyn & Bacon, 1999.
๗. มหาวิทยาลัยมหิดล. ข้อบังคับมหาวิทยาลัยมหิดลว่าด้วยการศึกษาระดับอนุปริญญาและปริญญาตรี พ.ศ. ๒๕๓๘.
๘. คณะแพทยศาสตร์ศิริราชพยาบาล. ประกาศคณะแพทยศาสตร์ศิริราชพยาบาล เรื่องแนวทางปฏิบัติสำหรับข้อบังคับมหาวิทยาลัยมหิดลว่าด้วยการศึกษาระดับอนุปริญญาและปริญญาตรี พ.ศ. ๒๕๓๘.

กระดาษบันทึก

กระดาษบันทึก

เอกสารประกอบการอบรม



16 กรกฎาคม 2563

16 กรกฎาคม 2563

Multiple-choice questions item development

MCQ Item Development

รศ.นพ. เข็ดศักดิ์ ไอร่มณีรัตน์
ภาควิชาศัลยศาสตร์ คณะแพทยศาสตร์ศิริราชพยาบาล
มหาวิทยาลัยมหิดล

1

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

Multiple-Choice Questions

- Advantages
 - Objective scoring
 - High internal consistency reliability
 - Strong research evidence to support its validity
 - Efficiency in testing and scoring

3

Multiple-Choice Questions

- Limitations
 - Cueing of correct answer
 - Random guessing
 - Testing of trivial knowledge
 - Difficulty of development of good MCQ items
 - Focus only in cognitive abilities, not good for assessing psychomotor skills or attitudes

4

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

5

Activity

- Open a web browser
- Go to <http://socrative.com>
- Select [Student login]
- In Room name, type in: IRAMANEERAT
- Click [Join]
- Type in your own name

A Good MCQ Item

1. Content
2. Structure

Guidelines for MCQ items

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

8 8

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

9 9

Format Guidelines

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

10 10

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

11 11

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

12 12

Guidelines for MCQ items

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

13

ข้อผิดพลาดในการสร้างข้อสอบปรนัย

14

Activity

- Open a web browser
- Go to <http://socrative.com>
- Select [Student login]
- In Room name, type in: IRAMANEERAT
- Click [Join]
- Type in your own name

Common Pitfalls

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

Questions & Comments

Cherdsak.ira@mahidol.ac.th

17

การสร้างข้อสอบปรนัย

ผู้ช่วยศาสตราจารย์ นายแพทย์เชิดศักดิ์ โอรมนรัตน์

ภาควิชาสรีรศาสตร์, คณะแพทยศาสตร์ศิริราชพยาบาล, มหาวิทยาลัยมหิดล, กรุงเทพมหานคร ๑๐๓๐๐.

ข้อสอบปรนัย (multiple-choice question) เป็นรูปแบบการประเมินผลที่นิยมใช้กันอย่างแพร่หลายในวงการแพทยศาสตรศึกษาเนื่องด้วยคุณสมบัติที่ดีหลายประการด้วยกัน ได้แก่ ประสิทธิภาพในการประเมินความรู้ปริมาณมากในเวลาอันสั้น ผลการประเมินที่ไม่มีผลกระทบจากความรู้สึกส่วนตัวของผู้ตรวจให้คะแนน คะแนนที่มีความเที่ยงสูง รวมถึงผลการวิจัยจำนวนมากที่สนับสนุนความถูกต้องของผลการประเมินด้วยข้อสอบปรนัย^{๑-๖} ข้อสอบปรนัยที่พัฒนาขึ้นอย่างดีนั้นสามารถวัดความรู้ได้ทั้งระดับการจดจำ การทำความเข้าใจ และการประยุกต์ความรู้ไปใช้ในการดูแลคนไข้^{๑-๔} อย่างไรก็ตาม การศึกษาวินิจฉัยเกี่ยวกับคุณภาพของข้อสอบปรนัยที่พัฒนาขึ้นใช้ในโรงเรียนแพทย์หลายแห่งพบว่าข้อสอบจำนวนไม่น้อยมีลักษณะที่ไม่เหมาะสม^{๕-๖} ข้อสอบปรนัยที่ถูกพัฒนาขึ้นอย่างไม่ถูกหลักการนั้นส่งผลเสียหลายอย่าง เช่น ทำให้ข้อสอบยากขึ้นโดยไม่จำเป็น ทำให้ผู้สอบเกิดความสับสน ทำให้ผู้สอบบางกลุ่มเสียเปรียบผู้สอบคนอื่น ทำให้การตัดสินใจผิดพลาด เป็นต้น^{๖-๘} ดังนั้นการออกข้อสอบปรนัยที่ดี วางอยู่บนหลักการที่ต้องจึงมีความสำคัญมากในการควบคุมคุณภาพการศึกษาในโรงเรียนแพทย์ บทความนี้จึงเขียนขึ้นเพื่อเป็นการรวบรวมหลักการพื้นฐานในการออกข้อสอบปรนัยที่ได้รับการยอมรับกันทั่วไปในวงการวัดและประเมินผล ผู้มีพันธหวังว่าข้อแนะนำต่าง ๆ ที่ได้นำเสนอในบทความนี้จะเป็นแนวทางที่เป็นประโยชน์ในการพัฒนาข้อสอบปรนัยที่มีคุณภาพให้ผู้อ่านไม่มากนักน้อย

รูปแบบพื้นฐานของข้อสอบปรนัย

ข้อสอบปรนัยคือข้อสอบชนิดที่มีคำถามแล้วมีตัวเลือกให้ผู้สอบเลือกตัวเลือกที่เหมาะสมเพื่อตอบคำถามดังกล่าว ข้อสอบปรนัยสามารถแบ่งออกได้เป็น ๒ รูปแบบ^๙ ได้แก่

๑. ข้อสอบถูกผิด (True/false item)

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

ข้อสอบชนิดถูกผิดนี้เคยเป็นที่นิยมมากในวงการแพทยศาสตรศึกษาอยู่ระยะหนึ่งเนื่องจากสามารถทดสอบความรู้ได้ปริมาณมาก แต่ข้อสอบชนิดนี้มีข้อจำกัดที่สำคัญคือสามารถใช้ได้เฉพาะกับเนื้อหาที่มีความถูกต้องชัดเจนเท่านั้น ซึ่งการตัดสินใจทางการแพทย์ส่วนมากไม่เป็นเช่นนั้น การตัดสินใจในการวินิจฉัย การตรวจค้นเพิ่มเติม หรือการรักษาผู้ป่วยส่วนใหญ่นั้นแพทย์ตัดสินใจเลือกระหว่างทางเลือกที่แตกต่างกันสามสี่อย่างซึ่งทุกทางเลือกมีความเป็นไปได้ มีส่วนถูก หรือมีความเหมาะสมในบางด้าน

เวบบทีกีรยา

บทความทั่วไป

แต่ก็มีความไม่เหมาะสมในด้านอื่นด้วย เช่นการเลือกใช้ยาในผู้ป่วยที่มีการติดเชื้อ นักศึกษาแพทย์มักรู้ว่าควรวินิจฉัยปฏิบัติวิธีอะไร ซึ่งยาปฏิชีวนะหลายชนิดก็รักษาการติดเชื้อชนิดนั้นๆ ได้ แต่นักศึกษาต้องเลือกระหว่างยาที่ล้นใช้ได้ในการรักษานั้นว่ายาใดที่มีประสิทธิภาพสูงสุด 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 หรือภาวะ หรือผู้ป่วยตามด้วยคำถาม หรือเว้นช่องว่างสำหรับเติมคำ หรือข้อความที่เหมาะสมลงไป

๒. ตัวเลือก (options) คือคำ หรือข้อความที่

ผู้ออกข้อสอบนำเสนอตามหลังจากโจทย์เพื่อให้ผู้สอบเลือกไปใช้ตอบคำถาม หรือเติมลงในช่องว่างในโจทย์

๒.๑ ตัวเลือกที่ถูกต้อง (correct option) เป็น

คำตอบที่ถูกต้องมีเพียงตัวเลือกเดียวต่อข้อสอบข้อหนึ่ง

๒.๒ ตัวลวง (distractors) เป็นคำตอบที่ผิด หรือ

ไม่เหมาะสม มีไว้ลวงให้ผู้สอบที่ไม่มีความรู้ หรือมีความเข้าใจไม่ถูกต้องในเนื้อหาที่นำมาออกข้อสอบเลือกตอบ ตัวลวงไม่จำเป็นต้องเป็นคำตอบที่ผิดชัดเจนเสมอไป ตัวลวงที่ดีมักมีส่วนถูกบ้าง แต่มีระดับของความถูกต้องเหมาะสมน้อยกว่าคำตอบที่ถูก

ข้อแนะนำพื้นฐานของการเขียนข้อสอบปรนัย

มีผู้เชี่ยวชาญทางการประเมินผลให้ข้อแนะนำจำนวนมากในการเขียนข้อสอบปรนัย เคยมีผู้รวบรวมไว้ถึง ๔๓ ข้อ^{๒,๓} ในที่นี้ผู้นิพนธ์ขอนำเสนอเฉพาะข้อแนะนำที่ได้รับยอมรับอย่างกว้างขวางและสามารถประยุกต์ใช้ได้ชัดเจนในการพัฒนาข้อสอบทางการแพทย์ โดยจะทำการจัดหมวดหมู่ของข้อแนะนำเหล่านี้ออกเป็น ๔ กลุ่มด้วยกัน ได้แก่ (๑) เนื้อหาข้อสอบ, (๒) การจัดรูปแบบข้อสอบ, (๓) การเขียนโจทย์, และ (๔) การเขียนตัวเลือก

๑. เนื้อหาข้อสอบ

๑.๑ ข้อสอบหนึ่งข้อควรมุ่งเน้นประเมินความรู้เพียงเรื่องเดียว

ก่อนเริ่มเขียนข้อสอบอาจารย์ผู้ออกข้อสอบควรตั้งวัตถุประสงค์ให้ชัดเจนว่าต้องการประเมินความรู้ของผู้สอบในเรื่องใด และเขียนโจทย์เพื่อตอบสนองวัตถุประสงค์ดังกล่าวเท่านั้น เนื่องจากเนื้อหาวิชาทางการแพทย์มีมาก อาจารย์แต่ละท่านเมื่อทำการสอนไปแล้วจึงอยากจะทำทดสอบความรู้ในหลายเรื่องที่ตนได้สอนไป แต่กลับมีโควตาจำกัดในการออกข้อสอบ ทำให้อาจารย์จำนวนไม่น้อยเขียนข้อสอบหนึ่งข้อถามทั้งเรื่องการวินิจฉัยโรค การตรวจค้นเพิ่มเติม การรักษาโรค และ ภาวะแทรกซ้อนของโรคไปพร้อมกัน ลักษณะข้อสอบเช่นนี้ไม่ควรใช้ เพราะมักซับซ้อนเกินไป เมื่อผู้สอบตอบข้อสอบผิด ก็ไม่สามารถวินิจฉัยได้ว่าผู้สอบขาดความรู้ ความเข้าใจในเรื่องใด

๑.๒ หลีกเลี่ยงการถามความรู้ในรายละเอียดปลีกย่อยที่ไม่มีที่ใช้ทางคลินิก (trivial content)

๓๐

มกราคม-มิถุนายน ๒๕๕๕, ปีที่ ๕, ฉบับที่ ๑

เวบบิ้นทีกสิริราช

บทความทั่วไป

องค์ความรู้ทางการแพทย์นั้นปริมาณมาก ไม่มีผู้ใดที่จดจำเนื้อหาที่มีในตำรา หรือวารสารทางการแพทย์ได้ทั้งหมด แม้ว่าจะองค์ความรู้หลายเรื่องมีความน่าสนใจ แต่มีประโยชน์ในการประยุกต์ใช้ทางคลินิกค่อนข้างน้อย องค์ความรู้ดังกล่าวจัดเป็นรายละเอียดปลีกย่อย (trivial content) ซึ่งไม่แนะนำให้ทำการทดสอบ สิ่งที่ดีควรทำการประเมินคือความสามารถในการประยุกต์ใช้ความรู้ในทางคลินิก (application of knowledge) ไม่แนะนำการทดสอบวัดความสามารถในการจดจำเป็นหลัก อย่างไรก็ตามก็ตามการที่แนะนำให้ออกข้อสอบที่เน้นการประยุกต์ใช้ความรู้ ไม่ได้หมายความว่าความจำเป็นที่ผู้ป่วยนั้นไม่ต้องใช้ความจำเลย ตรงกันข้ามการจดจำเนื้อหาเป็นพื้นฐานที่สำคัญในการแก้ปัญหาทางคลินิก ผู้สอบย่อมต้องจำเนื้อหาได้บ้าง จึงจะประยุกต์องค์ความรู้ดังกล่าวไปแก้โจทย์ปัญหาที่นำเสนอได้

๑.๓ หลีกเลียงการถามความรู้ในเรื่องที่ยังมีความขัดแย้งกันในแนวทางปฏิบัติ (controversy)

ความรู้ทางการแพทย์ในหลายหัวข้อยังเป็นเรื่องที่ยังมีผู้เชี่ยวชาญยังมีความเห็นแตกต่างกัน ผู้ป่วยรายเดียวกันไปพบแพทย์สองคนอาจได้รับการรักษาที่แตกต่างกันซึ่งวิธีการรักษาทั้งสองวิธีก็มีการวิจัยสนับสนุนด้วยกันทั้งคู่ อย่างไรก็ตามยังคงมีความขัดแย้ง (controversy) ในเรื่องดังกล่าวอยู่ เนื้อหาในลักษณะนี้ไม่ควรนำมาออกสอบด้วยข้อสอบปรนัย เนื่องจากในขณะที่ทำข้อสอบอยู่นั้น ผู้สอบไม่มีทางรู้ได้เลยว่าอาจารย์ผู้ออกข้อสอบอ้างอิงจากตำราหรือบทความวิชาการใด เนื้อหาที่ยังมีความขัดแย้งที่ผู้เชี่ยวชาญจากต่างสถาบันมีแนวทางในการปฏิบัติที่ต่างกันนี้แนะนำให้ใช้ข้อสอบในรูปแบบอื่นในการทดสอบเช่นข้อสอบอัตนัย เป็นต้น

๑.๔ หลีกเลียงการลอกประโยคหรือข้อความจากตำราโดยตรง

ดังได้กล่าวแล้วว่าข้อสอบที่ดีควรมุ่งเน้นการประเมินความเข้าใจ หรือ การประยุกต์ใช้ความรู้ ไม่ควรออกข้อสอบที่ประเมินความสามารถในการจำรายละเอียดปลีกย่อย การออกข้อสอบโดยวิธีการเปิดตำราแล้วคัดลอกประโยคจากตำราโดยตรงมักจะลงเอยด้วยข้อสอบที่ทดสอบความจำว่าผู้สอบท่องเนื้อหาในตำราตรงส่วนนั้นได้หรือไม่

ข้อสอบที่ดีควรได้จากการดูผู้ป่วย โจทย์ที่ดีควรเป็นปัญหาของผู้ป่วยที่พบในการทำงานนั่นเอง ตัวเลือกก็ได้จากข้อผิดพลาดที่นักศึกษาหรือแพทย์ประจำบ้านมักปฏิบัติกับผู้ป่วยแล้วทำให้ผลการรักษาไม่ดีขึ้นเอง

๑.๕ หลีกเลียงการนำเสนอข้อสอบที่ประเมินความรู้ในเรื่องเดียวกันสองข้อในข้อสอบชุดเดียวกัน

เนื่องจากเนื้อหาวิชาที่ต้องทำการประเมินในการสอบแต่ละครั้งนั้นมีมาก ดังนั้นองค์ความรู้ในแต่ละเรื่องแต่ละโรคจึงมักมีสัดส่วนของข้อสอบที่จะออกได้เพียงหนึ่งหรือสองข้อเท่านั้น การที่อาจารย์ออกข้อสอบในเรื่องหรือโรคเดียวกันซ้ำสองข้อในชุดข้อสอบเดียวกันจึงมักเป็นการลดโอกาสในการประเมินความรู้เรื่องอื่นซึ่งก็มีความสำคัญเช่นกัน การออกข้อสอบที่ดีนั้นควรต้องครอบคลุมวัตถุประสงค์การเรียนรู้ตามที่กำหนดในหลักสูตร หรือในเกณฑ์มาตรฐานผู้ประกอบวิชาชีพเวชกรรมอย่างสมดุล การที่จะบรรจุเป้าหมายดังกล่าวได้นั้นต้องเริ่มต้นจากการกำหนดสัดส่วนข้อสอบสร้างเป็นตารางกำหนดจำนวนข้อสอบ (table of specification) เมื่ออาจารย์ได้รับมอบหมายให้ออกข้อสอบควรต้องตรวจสอบให้ชัดเจนว่าเนื้อหาที่ต้องออกข้อสอบนั้นอยู่ในส่วนใดของตารางดังกล่าว การออกข้อสอบซ้ำซ้อนในเนื้อหาเรื่องเดียวกันเป็นสัญญาณบ่งบอกว่าอาจไม่ได้สร้างข้อสอบตามข้อกำหนดในตาราง นอกจากนี้การมีโจทย์สองข้อประเมินความรู้เรื่องเดียวกันมีความเป็นไปได้สูงที่เนื้อหาในข้อสอบข้อหนึ่งอาจบอกคำตอบในข้อสอบอีกข้อหนึ่งได้

๒. การจัดรูปแบบข้อสอบ

๒.๑ เลือกใช้คำศัพท์หรือรูปประโยคที่ง่ายต่อการทำความเข้าใจ

อาจารย์ผู้ออกข้อสอบต้องระลึกไว้เสมอว่าข้อสอบที่อาจารย์ออกเพื่อใช้ในการประเมินผลนักศึกษาแพทย์หรือแพทย์ประจำบ้านนั้นมีวัตถุประสงค์เพื่อทดสอบความรู้ทางการแพทย์เป็นสำคัญ มิใช่การประเมินความรู้ทางภาษาศาสตร์ ดังนั้นการเขียนข้อสอบของอาจารย์ควรเลือกใช้รูปแบบประโยคที่ง่ายต่อการทำความเข้าใจ อย่าเขียนประโยคซับซ้อนที่มีความยาวประโยคหลายบรรทัด มุ่งเน้นให้ภาษาเป็นสื่อในการนำเสนอความคิดของอาจารย์ผู้ออกข้อสอบไปยังผู้สอบ อย่าให้

เวบบันเทิงศิริราช

บทความทั่วไป

ภาษาเป็นอุปสรรคในการสื่อสาร การจะเลือกใช้ภาษาใดในการเขียนข้อสอบนั้นให้พิจารณาตามข้อกำหนดขององค์กรหรือหน่วยงานที่ควบคุมการสอบที่อาจารย์ส่งข้อสอบไปให้ใช้ ข้อสอบที่ใช้ในระดับการศึกษาหลักสูตรแพทยศาสตรบัณฑิตทั้งในระดับคณะ หรือข้อสอบที่ใช้ในการสอบระดับประเทศในปัจจุบันยังนิยมใช้ข้อสอบที่เขียนด้วยภาษาไทยโดยมีการใช้ศัพท์เทคนิคเป็นภาษาอังกฤษเหมือนดังภาษาที่แพทย์ใช้สื่อสารกันในการทำงานปกติ ส่วนข้อสอบในระดับหลังปริญญาบัตรหลายการสอบที่ภาควิชา หรือราชวิทยาลัยที่เกี่ยวข้องกำหนดให้ใช้ภาษาอังกฤษทั้งหมด ก่อนที่อาจารย์จะสร้างข้อสอบต้องมีการศึกษาข้อกำหนดของแต่ละการสอบให้ดี

๒.๒ หลักเลี่ยงการนำเสนอข้อมูลที่ไม่เกี่ยวข้องกับการแก้ปัญหาของโจทย์ข้อนั้น

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

๒.๓ จัดให้มีการตรวจสอบเนื้อหา คำศัพท์ และรูปแบบประโยคที่ใช้ในข้อสอบแต่ละข้อก่อนนำไปใช้

ถึงแม้ว่าอาจารย์ผู้เขียนข้อสอบจะได้มีการอ่านทวนสิ่งที่ตนเองเขียนแล้วเข้าใจเนื้อหาได้ดีและคิดว่าข้อสอบอยู่ในรูปแบบที่สามารถนำไปใช้ได้แล้ว ก็ไม่ควร

นำข้อสอบข้อนั้นไปใช้สอบเลย ควรให้มีคณะกรรมการข้อสอบซึ่งประกอบไปด้วยอาจารย์หลายท่านช่วยกันตรวจสอบและพิจารณาปรับแก้ข้อสอบทุกข้อก่อนนำไปใช้จริงเสมอ เนื่องจากผู้เขียนข้อสอบย่อมเข้าใจสิ่งที่ตนเขียนเสมอ แต่เมื่อผู้อื่นอ่านแล้วอาจพบว่ามีเนื้อหาที่กำกวมหรือเข้าใจโจทย์ต่างออกไปได้ การปรับแก้เนื้อหาที่มีความกำกวมหรือเฉลยซึ่งอาจารย์บางท่านอาจไม่เห็นด้วยให้ได้ข้อสอบที่มีความชัดเจน และอาจารย์ทุกท่านย่อมรับในค่าเฉลี่ยได้ก่อนจะนำข้อสอบไปทำการสอบจริงย่อมเป็นสิ่งที่ดีกว่าการตรวจพบปัญหาหลังจากสอบเสร็จแล้วซึ่งต้องมาตัดสินใจกันอีกว่าจะทำอย่างไรกับการคิดคะแนนของข้อสอบข้อดังกล่าว

๓. การเขียนโจทย์

๓.๑ เขียนโจทย์ให้มีความชัดเจน ผู้สอบทุกคนอ่านแล้วมีความเข้าใจตรงกัน

ข้อแนะนำนี้อาจดูเหมือนตรงไปตรงมา แต่กลับเป็นปัญหาที่พบบ่อยมากในการพัฒนาข้อสอบปรนัยประเด็นสำคัญคือโจทย์ที่ตั้นต้องมีความสมบูรณ์ในตัวเองโดยไม่ต้องอาศัยตัวเลือก โจทย์ข้อสอบที่ตั้นเมื่ออ่านโจทย์เสร็จแล้ว หากผู้สอบมีความรู้ในเรื่องที่ทำการประเมินนั้น เขาจะบอกคำตอบได้โดยไม่ต้องอ่านตัวเลือกเลย ดังนั้นเมื่ออาจารย์เขียนข้อสอบเสร็จแล้วแนะนำให้ลองปิดตัวเลือกแล้วอ่านเฉพาะโจทย์ดู หากอาจารย์อ่านแล้วบอกได้ว่าโจทย์ถามอะไรและบอกได้ว่าควรตอบอะไรโดยไม่ต้องอ่านตัวเลือกจัดว่าข้อสอบข้อดังกล่าวมีโจทย์ที่มีความชัดเจน

๓.๒ เรียบเรียงเนื้อหาให้ใจความสำคัญของข้อสอบอยู่ในโจทย์

เนื่องจากข้อสอบปรนัยมีตัวเลือกที่อาจารย์ต้องสร้างขึ้นหลายตัวเลือก บางครั้งอาจารย์ผู้พัฒนาข้อสอบอาจเผลอเรอเอาใจความสำคัญไปใส่ไว้ในตัวเลือกซึ่งทำให้เนื้อหาในโจทย์ขาดสาระสำคัญ อ่านโจทย์แล้วไม่เข้าใจว่าผู้ออกข้อสอบต้องการถามความรู้เรื่องอะไร ตัวอย่างข้อสอบที่ไม่เป็นไปตามข้อแนะนำนี้คือข้อสอบที่ถามว่า ข้อใดต่อไปนี้เป็นไปตั้นถูก หรือข้อใดต่อไปนี้เป็นไปตั้นผิดแล้วเขียนรายละเอียดเกี่ยวกับโรค หรือการรักษาบางอย่างในตัวเลือกแต่ละข้อ ข้อสอบในลักษณะนี้มักทำให้

เวบบิ้นทีกีรราช

บทความทั่วไป

ผู้สอบต้องอ่านข้อสอบย้อนไปมาหลายรอบกว่าจะเข้าใจจุดประสงค์ของข้อสอบ แล้วจึงตัดสินใจเลือกคำตอบโดยทั่วไปแนะนำให้อาจารย์นำเสนอรายละเอียดต่าง ๆ ไว้ในตัวโจทย์ให้มากที่สุด ส่วนตัวเลือกเขียนเป็นคำหรือข้อความสั้น ๆ

๓.๓ หลักเขียนการเขียนโจทย์ที่มีรูปประโยคเป็นเชิงปฏิเสธ

โจทย์ที่ดีไม่ควรอยู่ในประโยคเชิงปฏิเสธ เช่นถามถึงสิ่งที่เป็นข้อยกเว้น สิ่งที่ไม่ควรปฏิบัติ สิ่งทีพบน้อยที่สุด หรือสิ่งที่ไม่น่านึกถึง เป็นต้น งานวิจัยส่วนใหญ่พบว่าข้อสอบที่มีโจทย์ในรูปแบบปฏิเสธเหล่านี้มีระดับความยากง่ายไม่ต่างจากข้อสอบอื่น ๆ แต่งานวิจัยบางชิ้นพบว่าข้อสอบที่มีโจทย์ในรูปแบบปฏิเสธมีความยากมากกว่าข้อสอบอื่นชัดเจนโดยเฉพาะในข้อสอบวัดความรู้ระดับสูง^{๑๑-๑๒} แต่ผู้เชี่ยวชาญในการประเมินผลส่วนใหญ่มีความเห็นพ้องกันว่าข้อสอบประเภทนี้สามารถสร้างความสับสนให้กับผู้สอบได้ จึงไม่แนะนำให้ใช้ แต่หากอาจารย์ผู้ออกข้อสอบมีความจำเป็นต้องใช้ข้อสอบที่มีการใช้คำปฏิเสธในโจทย์ แนะนำให้พิมพ์คำปฏิเสธให้เด่นชัด โดยใช้ตัวหนาและขีดเส้นใต้เพื่อให้ผู้สอบเห็นชัด^{๑๐}

๔. การเขียนตัวเลือก

๔.๑ เขียนตัวเลือกที่มีประสิทธิภาพให้มีจำนวนมากที่สุดเท่าที่เหมาะสมกับบริบท

เรื่องจำนวนตัวเลือกที่เหมาะสมนี้เป็นเรื่องที่มีผู้เชี่ยวชาญด้านการประเมินผลจำนวนมากสนใจ มีงานวิจัยเกี่ยวกับเรื่องจำนวนตัวเลือกที่เหมาะสมในข้อสอบปรนัยอยู่มากมาย^{๑๓} อาจารย์ผู้ออกข้อสอบส่วนมากจะคุ้นเคยกับข้อสอบปรนัยชนิดที่มีห้าตัวเลือก ปกติครั้งที่อาจารย์ออกข้อสอบแล้วนึกตัวเลือกได้เพียงสามหรือสี่ตัว จึงเกิดคำถามว่าจำเป็นต้องมีตัวเลือกครบห้าตัวเลือกหรือไม่ งานวิจัยบางชิ้นพบว่าการลดจำนวนตัวเลือกลงทำให้ข้อสอบง่ายขึ้น^{๑๓-๑๔} แต่งานวิจัยบางชิ้นพบว่าการลดจำนวนตัวเลือกลงทำให้ได้ข้อสอบยากขึ้น^{๑๕-๑๖} ผู้เชี่ยวชาญในการประเมินผลเสนอว่าข้อสอบปรนัยที่มีตัวเลือกเพียงสามตัวเลือกก็สามารถทดสอบความรู้ได้อย่างมีประสิทธิภาพ^{๑๗-๑๙} แต่มีอาจารย์จำนวนไม่น้อยที่ไม่สบายใจที่มีตัวเลือกในข้อสอบแต่ละข้อน้อยกว่าห้าตัว

เลือกด้วยกังวลว่าจะทำให้มีโอกาสสูงที่ผู้สอบที่ไม่มีความรู้จะเดาสุ่มได้คำตอบที่ถูกต้อง แต่จากข้อมูลที่ปรากฏในปัจจุบันพบว่าผู้สอบในการสอบในระดับสูงนั้นพฤติกรรมเดาสุ่มโดยที่ผู้สอบปราศจากความรู้ไม่น่าจะมีบทบาทน้อยมาก ผู้สอบส่วนใหญ่มักมีความรู้บ้างและสามารถตัดตัวเลือกที่ไม่สมเหตุสมผลอย่างชัดเจนได้^{๑๐} ในการศึกษาข้อสอบปรนัยส่วนใหญ่พบตัวเลือกที่ไม่ทำงานเป็นจำนวนไม่น้อย^{๑๔} ข้อมูลที่ได้จากการวิเคราะห์ข้อสอบปรนัยที่ใช้ในทางแพทยศาสตรศึกษาในประเทศไทยหลายครั้งก็สอดคล้องกับงานวิจัยในต่างประเทศที่พบว่าข้อสอบส่วนใหญ่มักมีตัวเลือกที่ทำงานจริงราวสามหรือสี่ตัวเลือก มีข้อสอบน้อยข้อมากที่ตัวเลือกทั้งห้าตัวเลือกทำงานอย่างมีประสิทธิภาพ

ด้วยข้อมูลจากการศึกษาต่าง ๆ ขอนำเสนอในการออกข้อสอบปรนัยในปัจจุบันคือให้อาจารย์เขียนจำนวนตัวเลือกมากที่สุดที่มีความเหมาะสมกับเนื้อหาโจทย์ ไม่จำเป็นต้องเขียนตัวเลือก ๕ ตัวเลือกเสมอไป เนื่องจากตัวเลือกที่ห้าที่เขียนขึ้นเพื่อเติมเต็มโดยไม่สมเหตุสมผลนั้นมักไม่ค่อยมีคนเลือก หากเนื้อหาที่อาจารย์นำมาสอบมีตัวเลือกที่เหมาะสมเพียงสามหรือสี่ตัวเลือกก็เขียนจำนวนตัวเลือกเพียงสามหรือสี่ตัวเลือก^{๑๐} แต่อย่างไรก็ตามให้อาจารย์ศึกษาข้อกำหนดของแต่ละการสอบที่อาจารย์เกี่ยวข้องด้วย เนื่องจากนโยบายของแต่ละการสอบแตกต่างกันไป องค์กรที่จัดสอบทางแพทยศาสตรศึกษาจำนวนไม่น้อยยังคงตั้งข้อกำหนดให้ใช้ข้อสอบ ๕ ตัวเลือกเสมอ ซึ่งหากอาจารย์ไม่ทำตามข้อกำหนดดังกล่าวข้อสอบที่ออกไปอาจไม่ได้รับการพิจารณาได้

๔.๒ จัดให้ตัวเลือกที่ถูกต้องมีการกระจาย

ตำแหน่งไปให้มีจำนวนพอ ๆ กันในทุกตัวเลือก

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

๔.๓ เขียนตัวเลือกแต่ละข้อให้เป็นอิสระ ไม่ขึ้นต่อกัน

๓๓

มกราคม-มิถุนายน ๒๕๕๕, ปีที่ ๕, ฉบับที่ ๑

เวบบิ้นทีกีริราช

บทความทั่วไป

ในการเขียนตัวเลือกของข้อสอบแต่ละข้อ อาจารย์ต้องระมัดระวังให้ตัวเลือกแต่ละตัวเลือกไม่มีความซ้ำซ้อนกัน เช่นตัวเลือก ก เป็นยากลุ่มย่อยของตัวเลือก ข ตัวเลือก ก เป็นช่วงอายุ ๒ - ๑๐ ปี ตัวเลือก ข เป็นช่วงอายุ ๕ - ๑๑ ปี เป็นต้น การเขียนตัวเลือกที่ซ้ำซ้อนกันนี้ หากเกี่ยวข้องกับตัวเลือกที่ถูกต้องอาจมีผู้สอบแย้งว่ามีตัวเลือกที่ถูกต้องมากกว่าหนึ่งตัวเลือก 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 ก็น่าจะเป็นยาปฏิชีวนะ เช่นกันไม่ใช่ยาเคมีบำบัด หรือยาต้านเชื้อรา) การที่มีตัวเลือกที่ไม่เข้าพวก ไม่มีความเป็นเนื้อเดียวกันกับตัวเลือกอื่นเป็นคำบอกใบ้ในการตัดตัวเลือกที่ผู้สอบนิยมใช้มาก ดังนั้นอาจารย์ผู้ออกข้อสอบควรหลีกเลี่ยง

ในบางบริบทของการดูแลรักษาผู้ป่วย สิ่งแพทย์ต้องตัดสินใจเลือกอาจมีทั้งการเลือกที่จะให้การรักษาเลยหรือจะส่งตรวจค้นเพิ่มเติมก่อน ในกรณีนี้ อาจารย์สามารถเขียนตัวเลือกที่มีการรักษาและการตรวจเพิ่มเติมปะปนกันได้ 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^{๑๑} จึงแนะนำให้หลีกเลี่ยงการใช้

ตัวเลือก “ไม่มีข้อใดถูก” เป็นประเด็นที่ผู้เชี่ยวชาญในการประเมินผลยังคงถกเถียงกันอยู่บ้าง ผู้เชี่ยวชาญบางส่วนเห็นว่าไม่ควรใช้ตัวเลือกประเภทนี้ แต่ผู้เชี่ยวชาญบางส่วนให้ความเห็นว่าสามารถใช้ได้ในบางกรณี^{๑๒} เหตุผลที่ตัวเลือกชนิดนี้เป็นปัญหาคือการใช้ตัวเลือกนี้มักสร้างความลำบากใจให้กับผู้สอบในการเลือกคำตอบที่ถูกในกรณีที่ตัวเลือกแต่ละตัวเลือกไม่ถูกหรือผิดชัดเจน เพราะผู้สอบจะต้องทำการเปรียบเทียบตัวเลือกที่นำเสนอในข้อสอบกับทางเลือกอื่น ๆ ที่เขานึกได้^{๑๓} หากโจทย์ถามว่ายาใดที่ควรให้แก่ผู้ป่วย แล้วมีซีอียาสี่ชนิด และมีตัวเลือก “ไม่มีข้อใดถูก” นอกจากที่ผู้สอบต้องนึกว่าในบรรดา ยาที่ปรากฏในตัวเลือกนั้นเหมาะสมหรือไม่แล้วเขายังนึกต่อไปอีกว่ามียาอื่นใดที่สามารถให้ในผู้ป่วยรายนี้ได้อีก หากเขานึกออกว่ามียาอื่นที่น่าจะเหมาะสมกับผู้ป่วยมากกว่ายาในตัวเลือก (ด้วยเหตุผลที่อาจแตกต่างไปจากที่อาจารย์ผู้ออกข้อสอบคิด) เขาก็จะเลือก “ไม่มีข้อใดถูก”

การใช้ตัวเลือก “ไม่มีข้อใดถูก” จะยังเป็นปัญหามากขึ้นในข้อสอบที่ถามถึงสิ่งที่ไม่ควรทำ เช่นยาใดไม่ควรใช้ในผู้ป่วย ซึ่งนอกจากยาที่นำเสนอในตัวเลือกแล้วย่อมมียาชนิดอื่นอีกมากมายในบัญชียาที่ไม่เหมาะสม ซึ่งไม่มี

เวบบันทึทศิรราช

บทความทั่วไป

ทางที่ใครจะรู้ได้ว่าการที่ผู้สอบเลือกตอบ “ไม่มีข้อใดถูก” นั้นเขาคิดถึงยาใด และยานั้นไม่เหมาะสมมากไปกว่ายาที่มีอยู่ในตัวเลือกหรือไม่ งานวิจัยทั้งหมดที่ศึกษาถึงตัวเลือกชนิดนี้ได้ข้อสรุปที่ตรงกันว่าข้อสอบที่ใช้ตัวเลือกประเภทนี้เพิ่มระดับความยากให้ข้อสอบ^{๖๖} โดยทั่วไปแล้วจึงไม่แนะนำให้ใช้ตัวเลือกประเภทนี้ในการสอบทางแพทยศาสตรศึกษาซึ่งทางเลือกสำหรับสถานการณ์ที่น่าเสนาหามีได้มากและการตัดสินใจเลือกคำตอบต้องอาศัยการเปรียบเทียบข้อดีข้อเสียของแต่ละตัวเลือก

สรุป

ในบทความนี้ผู้พิมพ์ได้กล่าวถึงข้อแนะนำขั้นพื้นฐานในการพัฒนาข้อสอบปรนัยชนิดเลือกคำตอบที่ถูกที่สุดโดยสรุปข้อแนะนำเหล่านี้ออกเป็นสี่กลุ่มด้วยกัน ได้แก่ (๑) เนื้อหาข้อสอบ, (๒) การจัดรูปแบบข้อสอบ, (๓) การเขียนโจทย์, และ (๔) การเขียนตัวเลือก ผู้พิมพ์หวังว่าข้อแนะนำเหล่านี้คงพอเป็นแนวทางสำหรับอาจารย์แพทย์ในการพัฒนาข้อสอบปรนัยที่มีคุณภาพเพื่อใช้ในการประเมินนักศึกษาแพทย์และแพทย์ประจำบ้านได้บ้าง อย่างไรก็ตามบทความนี้เป็นการกล่าวถึงข้อแนะนำเบื้องต้นเท่านั้น ยังมีข้อแนะนำอื่นๆ ที่ผู้พิมพ์ไม่ได้นำมารวบรวมไว้ในบทความนี้เพื่อต้องการทำให้เนื้อหากระชับโดยข้อแนะนำอื่นๆ ที่ผู้พิมพ์ไม่ได้กล่าวถึงนี้พบว่าเป็นปัญหาน้อยในการออกข้อสอบทางการแพทย์ หรือเป็นข้อแนะนำที่ไม่ได้รับการสนับสนุนอย่างกว้างขวางจากผู้เชี่ยวชาญทางการวัดและประเมินผล หากผู้อ่านสนใจรายละเอียดของข้อแนะนำอื่นๆ ที่มีผู้กล่าวไว้สามารถศึกษาเพิ่มเติมได้จากเอกสารอ้างอิงที่แสดงไว้ท้ายบทความ

มีข้อควรพิจารณาในการประยุกต์ใช้ข้อแนะนำเหล่านี้ในการพัฒนาข้อสอบที่ผู้พิมพ์ขอล่าถึงประการหนึ่งคือ แม้ว่าข้อแนะนำที่กล่าวถึงเหล่านี้หลายข้อมีการศึกษาวิจัยสนับสนุนที่ชัดเจน แต่สิ่งเหล่านี้ก็เป็นเพียงข้อแนะนำว่าผู้ออกข้อสอบควรปฏิบัติ ไม่ใช่กฎเกณฑ์ตายตัว การเขียนข้อสอบปรนัยนั้นเป็นงานที่ต้องอาศัยทั้งศาสตร์และศิลปะผสมผสานกันอย่างเหมาะสม

หาใช้สูตรคณิตศาสตร์ที่ไม่มีข้อยกเว้น ผู้พิมพ์ไม่คาดหวังให้อาจารย์ผู้พัฒนาข้อสอบยึดข้อแนะนำเหล่านี้เสมือนกฎเกณฑ์ตายตัวที่ต้องทำตามในทุกกรณี หากแต่ต้องการให้อาจารย์ใช้เป็นแนวทางในการสร้างข้อสอบ ในบางบริบทผู้ออกข้อสอบอาจเลือกที่จะไม่ปฏิบัติตามข้อแนะนำบางประการได้บ้าง แต่การที่จะไม่ปฏิบัติตามข้อแนะนำเหล่านี้จำเป็นต้องมีเหตุผลที่เหมาะสม และควรทำไม่บ่อยนัก ยกตัวอย่างเช่นข้อแนะนำว่า โจทย์ไม่ควรเขียนถามข้อยกเว้น จะพบได้ว่ามีบางบริบทที่การรู้ข้อยกเว้น หรือข้อห้ามปฏิบัติก็เป็นองค์ความรู้ที่สำคัญในการดูแลรักษาผู้ป่วย ดังนั้นในบริบทที่เหมาะสมผู้พิมพ์เองก็เห็นด้วยว่าอาจเขียนโจทย์ที่ถามข้อยกเว้นได้ แต่อย่างไรก็ตามการจะไม่ปฏิบัติตามข้อแนะนำนี้ต้องไม่ทำบ่อยจนเกินจำเป็น หากออกข้อสอบ ๑๐๐ ข้อ จะมีข้อสอบที่ถามข้อยกเว้น ประมาณบ้าง 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 medical education. Dordrecht: Kluwer Academic Publishers, 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 Associates, 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, Koepfen 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.

เวชบันทึกศิริราช

บทความทั่วไป

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.

๓๖

มารากน-นิพนธ์ ๒๕๕๕, ปีที่ ๕, ฉบับที่ ๑

ข้อผิดพลาดที่ควรระวังในการสร้าง ข้อสอบปรนัย

ผู้ช่วยศาสตราจารย์ นายแพทย์เชิดศักดิ์ โสมณิรัตน์

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 ทำให้ข้อสอบยากขึ้นสร้างความสับสนให้ผู้สอบ ทำให้ผู้สอบบางกลุ่มเสียเปรียบและทำให้การตัดสินผลสอบผิดพลาด เป็นต้น^{๓-๕} ดังนั้นการออกข้อสอบปรนัยที่มีคุณภาพดีจึงเป็นงานที่มีความสำคัญและท้าทายความสามารถ

การสร้างข้อสอบปรนัยที่มีคุณภาพดีนั้นควรเริ่มต้นจากการมีองค์ความรู้พื้นฐานในการสร้างข้อสอบแล้วเกิดการฝึกฝนทักษะ สังคมประสบการณ์ในการออกข้อสอบจนเกิดความชำนาญ ปัญหาที่พบบ่อยในโรงเรียนแพทย์หลายแห่งคือมีอาจารย์จำนวนไม่น้อยที่ได้รับมอบหมายให้ออกข้อสอบปรนัย โดยไม่ได้มีการพัฒนาองค์ความรู้พื้นฐานที่เหมาะสมก่อน ซึ่งเป็นเหตุให้มีข้อสอบปรนัยที่มีลักษณะไม่เหมาะสมตามหลักการออกข้อสอบปะปนมาในข้อสอบที่ให้นักศึกษาแพทย์และแพทย์ประจำบ้านทำอยู่บ้าง ผู้นิพนธ์จึงเห็นความสำคัญของการเผยแพร่องค์ความรู้พื้นฐานของการออกข้อสอบปรนัย องค์ความรู้พื้นฐานในการสร้างข้อสอบปรนัยนั้นมีสองส่วน ส่วนแรกเป็นหลักการของการสร้างข้อสอบทั่วไปซึ่งได้มีผู้รวบรวมเป็นข้อเสนอแนะดีพิมพ์ในตำราและวารสารทางวิชาการอยู่บ้าง^{๑, ๕-๗} ส่วนที่สองเป็นข้อผิดพลาดในการสร้างข้อสอบที่อาจารย์ผู้ออกข้อสอบพึงหลีกเลี่ยง ในบทความนี้ผู้นิพนธ์จะมุ่งเน้นในส่วนที่สองนี้โดยจะรวบรวมข้อผิดพลาดในการสร้างข้อสอบปรนัย ที่อาจเป็นตัวบอกใบ้ให้ผู้สอบที่ไม่มีความรู้ในเรื่องที่ทำการทดสอบสามารถเลือกคำตอบที่ถูกต้องได้ ดังนั้นการที่อาจารย์ผู้ออกข้อสอบทราบถึงสิ่งเหล่านี้และหลีกเลี่ยงเสียจะส่งผลให้ข้อสอบปรนัยที่สร้างขึ้นสามารถใช้วัดองค์ความรู้ทางการแพทย์ได้จริง โดยปราศจากปัจจัยรบกวนจากการสังเกตพบสิ่งบอกรับคำตอบ

๓/๗

กรกฎาคม-ธันวาคม ๒๕๕๕, ปีที่ ๕, ฉบับที่ ๒

เวบบิ้นทีกีรียา

บทความทั่วไป

ข้อสอบปรนัยที่กล่าวถึงในบทความนี้มุ่งประเด็นไปที่ข้อสอบปรนัยชนิดเลือกคำตอบที่ถูกต้องที่สุด (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) การไปคำตอบด้วยหลักตรรกะ, (๓) การใช้คำคุณศัพท์บอกระดับของความแน่ชัด, (๔) ความยาวของตัวเลือก, (๕) การใช้คำซ้ำในโจทย์และตัวเลือก, (๖) การเข้าพวกของคำ หรือข้อความที่ปรากฏในตัวเลือก, และ (๗) การบอกไปคำตอบโดยโจทย์ข้ออื่น

๑. ข้อผิดพลาดในไวยากรณ์

ตัวเลือกทุกตัวต้องสามารถตอบโจทย์ได้อย่างถูกต้องตามหลักไวยากรณ์ บ่อยครั้งอาจารย์ผู้ออกข้อสอบมุ่งความสนใจไปที่คำตอบที่ถูกต้อง และให้ความสนใจกับตัวลวงน้อยไปจนทำให้ตัวลวงผิดหลักไวยากรณ์ โดยมักพบบ่อยในข้อสอบที่เป็นภาษาอังกฤษ ข้อผิดพลาดที่พบได้บ่อยเช่น ความไม่เข้ากันของ article (A, An, The) กับคำนามที่ตามหลัง, คำนามกับกริยาที่ไม่เข้ากันในเชิงเอกพจน์หรือพหูพจน์, การเติมคำในประโยคที่เว้นว่างไว้สำหรับเติมคำนามแต่ตัวลวงเป็นกริยาหรือเป็นคำนามในลักษณะที่ไม่เข้ากับรูปประโยค 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 gonorrhoea*
- 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. การไปคำตอบด้วยหลักตรรกะ

ในการเขียนตัวเลือก อาจารย์ผู้ออกข้อสอบต้องระมัดระวังไม่ให้ผู้สอบสามารถตัดตัวเลือกได้ด้วยหลักตรรกศาสตร์ เนื่องจากผู้สอบที่มีทักษะการทำข้อสอบดีจะสามารถพิจารณาความเป็นไปได้ของตัวเลือกต่าง ๆ และตัดตัวลวงที่ไม่มีทางเป็นไปได้ตามหลักของเหตุและผลออกไปได้โดยไม่ต้องอาศัยความรู้เรื่องที่ว่าอาจารย์ตั้งเป้าหมายว่าจะทดสอบ

ตัวอย่างที่ ๓. ภาวะไส้เลื่อนบริเวณขาหนีบ (inguinal hernia)

- A. พบในผู้ชายบ่อยกว่าผู้หญิง
- B. พบในผู้หญิงบ่อยกว่าผู้ชาย
- C. พบเกิดขึ้นในผู้หญิงและผู้ชายในอัตราเท่ากัน
- D. พบบ่อยในผู้ที่มีกระดูกสันหลังคด
- E. พบในผู้ที่มีภูมิคุ้มกันต่ำในทวีปเอเชีย มากกว่าผู้ที่มีภูมิคุ้มกันต่ำในทวีปยุโรป

ในตัวอย่างที่ ๓ นี้อาจารย์ผู้ออกข้อสอบต้องการวัดความรู้เรื่องอุบัติการณ์ของไส้เลื่อนขาหนีบ แต่หาก

เวบบันทึทศิรราช

บทความทั่วไป

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

๓. การใช้คำคุณศัพท์บอกระดับของความแน่ชัด

อาจารย์ผู้ออกข้อสอบพึงระมัดระวังการใช้คำคุณศัพท์ที่บอกระดับความแน่ชัดของข้อความ ซึ่งจะมีหลายระดับ โดยทั่วไปแล้วคำคุณศัพท์ที่แสดงความแน่ชัดมาก แสดงความมั่นใจมาก (เช่น always, never) มักไม่ถูกต้อง เนื่องจากในทางการแพทย์นั้นมีความไม่แน่นอนเกิดขึ้นเป็นประจำ ข้อความที่บอกเล่าถึงสิ่งที่เป็นไปได้โดยไม่ชี้ชัดลงไปว่าต้องเกิดขึ้นแน่นอน (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 เป็นข้อความบอกว่ามีโอกาสเป็นไปได้โดยไม่ชี้ชัดว่าต้องเกิด

ผู้สอบที่มีทักษะการทำข้อสอบดีจะตัดตัวเลือก B, C, และ D ได้โดยไม่ต้องอาศัยความรู้ทางการแพทย์เลย

๔. ความยาวของตัวเลือก

มีการตั้งข้อสังเกตว่าอาจารย์แพทย์มักชอบสอนและอธิบายแม้กระทั่งในการสอบอาจารย์แพทย์หลายท่านก็ติดนิสัยรักการสอนนี้มาด้วย ทำให้อาจารย์มักเขียนตัวเลือกที่ถูกต้องที่มีคำอธิบายประกอบอย่างครบถ้วนทำให้ตัวเลือกที่ถูกมักมีความยาวมากกว่าตัวลวง^๔ นักศึกษาผู้เข้าสอบจำนวนไม่น้อยรู้จักความจริงข้อนี้และมักเลือกตัวเลือกที่มีความยาวมากที่สุด หากเขาไม่สามารถหาคำตอบได้ด้วยความรู้ทางการแพทย์ที่เขา

ตัวอย่างที่ ๕. ผู้หญิงอายุ ๒๘ ปี แต่งงานมานาน ๑ ปี ยังไม่มีบุตร คุณกำเนิดโดยการกินยาคุมเป็นประจำ สังเกตว่าตนเองน้ำหนักตัวเพิ่มขึ้นหลังจากกินยาคุมมาขอคำแนะนำเรื่องการคุมกำเนิด ท่านจะแนะนำอย่างไร

- A. ให้เปลี่ยนไปใช้การใส่ห่วงอนามัย
- B. ให้ใช้ถุงยางอนามัย
- C. ให้กินยาคุมกำเนิดต่อไปได้เนื่องจากมีการศึกษาแล้วว่ายาคุมกำเนิดชนิดกินไม่ส่งผลให้เกิดการเพิ่มขึ้นของน้ำหนักตัว

D. ให้รับประทานยาลดความอ้วน

ในตัวอย่างที่ ๕ นี้จะสังเกตเห็นว่าตัวเลือก C มีการอธิบายเหตุผลประกอบส่งผลให้มีความยาวมากกว่าตัวเลือกอื่นชัดเจน ลักษณะเช่นนี้จะเป็นการบอกใบ้ให้นักศึกษาเลือกตัวเลือกนี้

๕. การใช้คำซ้ำในโจทย์และตัวเลือก

การใช้คำเดียวกัน หรือคำที่มีความหมายเหมือนกันในโจทย์และตัวเลือก มักเป็นการบอกใบ้ว่าตัวเลือกดังกล่าวเป็นตัวเลือกที่ถูกต้อง^๕

ตัวอย่างที่ ๖. Which of the following statements is true regarding sacular 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.

เวบบิ้นทีกีรียา

บทความทั่วไป

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 การที่มีคำที่มีความหมายซ้ำกันเช่นนี้เป็นตัวบอกใบ้คำตอบที่อาจารย์ผู้ออกข้อสอบต้องตรวจตราให้ดีกว่าก่อนนำข้อสอบไปใช้

๖. การเข้าพวกของคำ หรือข้อความที่ปรากฏในตัวเลือก

ข้อสอบจำนวนไม่น้อยนำเสนอรายการของหลายอย่างในตัวเลือก (เช่น ชื่อการตรวจค้นเพิ่มเติม ชื่อโรค ชื่อยา ฯลฯ) มีผู้เชี่ยวชาญในการประเมินผลตั้งข้อสังเกตว่าในข้อสอบเหล่านี้ตัวเลือกที่ถูกต้องมักมีลักษณะเข้าพวกกับตัวเลือกอื่นมากที่สุด หากเป็นรายการของตัวเลือกที่ถูกก็คือข้อที่มีจำนวนรายการซ้ำกับตัวเลือกอื่นมากที่สุด ดังนั้นในการนำเสนอตัวเลือกอาจารย์ผู้ออกข้อสอบพึงระมัดระวังอย่าให้ตัวเลือกที่ถูกต้องมีลักษณะที่เข้าพวกได้อย่างชัดเจน พยายามทำตัวหลงอื่นให้มีลักษณะเข้าพวกให้ใกล้เคียงกับตัวเลือกที่ถูกต้อง

ตัวอย่างที่ ๗. โรคที่แพทย์วินิจฉัยผิดว่าเป็นได้ตั้งอันดับยี่สิบเรียงลำดับจากมากไปน้อยคือ

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) ส่วนโรคที่เหลือนกล่าวถึงโรคละครั้งเดียว ดังนั้นตัวเลือกที่มีพวกรวมมากที่สุดคือตัวเลือก A ซึ่งเป็นคำตอบที่ถูกต้อง

การเข้าพวกของตัวเลือกที่ถูกนั้น ไม่จำเป็นต้องเป็นลักษณะของการมีจำนวนหรือความถี่ของคำมากที่สุดเพียงเท่านั้น อาจหมายรวมถึงการมีรูปร่างลักษณะ หรือความหมายคล้ายคลึงกันได้ด้วย

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

A. Amoxicillin 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 ในบรรดา ยาฉีดจะเห็นว่าเม็ดต้านเชื้อราที่ไม่เข้าพวก (ตัวเลือก C) ดังนั้นจะเหลือตัวเลือกที่นักศึกษาต้องคิดเลือกจริง ๆ เพียงตัวเลือก B กับ D ซึ่งหากดูกลุ่มยาาก็จะพบว่ายากุ่ม Cephalosporin เข้าพวกมากที่สุด ทำให้ผู้สอบที่มีทักษะการทำข้อสอบดีสามารถเลือกคำตอบที่ถูกต้อง (ตัวเลือก B) ได้โดยไม่ต้องมีความรู้เรื่องการรักษาผู้ป่วย febrile neutropenia

๗. การบอกใบ้คำตอบโดยโจทย์ข้ออื่น

ข้อผิดพลาดนี้เป็นข้อผิดพลาดที่ตัวผู้เขียนข้อสอบไม่ค่อยรู้แต่ผู้ที่ตรวจพบข้อผิดพลาดนี้คืออาจารย์ผู้เลือกข้อสอบไปใช้ เนื่องจากในการสอบแต่ละครั้งใช้ข้อสอบจำนวนมาก หากเลือกข้อสอบโดยไม่ระมัดระวังอาจมีข้อสอบสองข้อที่ถามเกี่ยวกับโรคหรือกลุ่มอาการเดียวกัน ซึ่งข้อมูลจากโจทย์ในข้อหนึ่งอาจเป็นตัวบอกใบ้คำตอบของข้อสอบอีกข้อได้ ดังนั้นเมื่อทำการเลือกข้อสอบเสร็จแล้วจัดหน้ากระดาษเข้ารูปเล่มข้อสอบแล้วอาจารย์ควรอ่านข้อสอบฉบับสมบูรณ์นี้อีกหนึ่งหรือสองรอบก่อนส่ง

๘๐

กรกฎาคม-ธันวาคม ๒๕๕๕, ปีที่ ๕, ฉบับที่ ๒

เวบบิ้นทีกีธีรธา

บทความทั่วไป

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 ซึ่งผู้สอบที่มีทักษะการทำข้อสอบดีสามารถย้อนกลับไปอ่านโจทย์ในข้อก่อนหน้านั้นแล้วได้ข้อมูลว่าผู้ป่วยที่นำเสนอว่าเป็น febrile neutropenia มีไข้ขึ้น ๑๔ วันหลังได้ยาเคมีบำบัด ก็สามารถตอบข้อสอบข้อนี้ถูกต้องได้ง่าย

สรุป

ผู้นิพนธ์ได้รวบรวมข้อผิดพลาดในการสร้างข้อสอบปรนัยที่ผู้สอบอาจใช้เป็นแนวทางในการเลือกคำตอบที่ถูกต้องโดยไม่ต้องอาศัยความรู้ทางการแพทย์ที่อาจารย์ต้องการประเมินผล โดยเรียงเรียงเป็นเจ็ดกลุ่มข้อผิดพลาดด้วยกัน ผู้อ่านทุกท่านพึงตระหนักว่าสิ่งเหล่านี้ไม่ใช่หลักการทางวิทยาศาสตร์ที่ชัดเจนดังกฎทางคณิตศาสตร์หรือฟิสิกส์ หากแต่เป็นการรวบรวมข้อสังเกต

และคำแนะนำของผู้เชี่ยวชาญทางการวัดและประเมินผล จึงเป็นเพียงแนวทางเบื้องต้นในการพิจารณาตรวจสอบเนื้อหาของข้อสอบเท่านั้น การประยุกต์ใช้องค์ความรู้นี้คงต้องอาศัยศิลปะพอสมควรเพื่อที่จะได้ข้อสอบที่ดีสามารถวัดองค์ความรู้ทางการแพทย์ของนักศึกษาหรือแพทย์ประจำบ้านที่เข้าสอบได้ตามวัตถุประสงค์ของการสอบ

เอกสารอ้างอิง

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

16 กรกฎาคม 2563

Multiple-choice questions item analysis

MCQ Item Analysis

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

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

MCQ item analysis

Outline

- Item statistics
- Test statistics
- Applications
- Limitations

MCQ item analysis

Two Parts of 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)

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

MCO item analysis

Item Discrimination

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

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

MCO 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

MCO 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

MCO item analysis

11

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	

MCO item analysis

12

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	

MCO item analysis

13

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					

MCO item analysis

14

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					
P-VALUE	0	0	0.73	0	0.26					
PT BISERIAL	0	0	-0.05	-0.01	0.05					

MCO item analysis

15

Siriraj Hospital's IA report

No. : 1		p Value : 0.64				rpb1 : 0.23			
A	B	C	* D	E					
rpb1	%	rpb1	%	rpb1	%	rpb1	%	rpb1	%
0.02	6.98	-0.18	5.08	-0.17	8.57	0.23	63.81	-0.07	15.56

MCO item analysis

Item Analysis and Option Analysis

Faculty of Medicine Siriraj Hospital
Mahidol University

No. : 1		p Value : 0.64				rpb1 : 0.23				No. : 2		p Value : 0.34				rpb1 : 0.19			
A	B	C	* D	E					A	B	C	D	* E						
rpb1	%	rpb1	%	rpb1	%	rpb1	%	rpb1	%	rpb1	%	rpb1	%	rpb1	%	rpb1	%	rpb1	%
0.02	6.98	-0.18	5.08	-0.17	8.57	0.23	63.81	-0.07	15.56	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				rpb1 : 0.35				No. : 4		p Value : 0.50				rpb1 : 0.33			
A	B	* C	D	E					A	* B	C	D	E						
rpb1	%	rpb1	%	rpb1	%	rpb1	%	rpb1	%	rpb1	%	rpb1	%	rpb1	%	rpb1	%	rpb1	%
-0.03	8.89	-0.26	23.17	0.35	55.87	-0.05	3.17	-0.16	8.89	-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				rpb1 : 0.06				No. : 6		p Value : 0.53				rpb1 : 0.20			
A	B	C	* D	E					A	B	* C	D	E						
rpb1	%	rpb1	%	rpb1	%	rpb1	%	rpb1	%	rpb1	%	rpb1	%	rpb1	%	rpb1	%	rpb1	%
-0.06	3.49	-0.08	53.02	0.05	12.06	0.06	23.81	0.02	7.62	-0.16	23.17	-0.11	3.81	0.20	53.33	-0.02	5.40	-0.02	14.29

MCO item analysis

Test Statistics

Looking at the whole test

MCO 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

20

Internal Consistency Reliability

- Consistency of test scores: If we test the students again, will they get the same scores?
- In MCQ exam, one commonly reported index of reliability is Cronbach's Alpha

n = number of testlets
 = score variance of total scores
 = score variance of the i^{th} testlet

MCQ item analysis

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

22

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

23

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

MCO 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.

MCO item analysis

การวิเคราะห์ข้อสอบปรนัย

อาจารย์ นายแพทย์เชิดศักดิ์ ไธรมณีนรัตน์

ภาควิชาวิทยาศาสตร์, คณะแพทยศาสตร์ศิริราชพยาบาล, มหาวิทยาลัยมหิดล, กรุงเทพมหานคร ๑๐๗๐๐.

การวิเคราะห์ข้อสอบปรนัย (Item analysis) เป็นการใช้วิธีการทางสถิติเพื่อวิเคราะห์คำตอบที่ผู้สอบตอบข้อสอบปรนัยในการสอบครั้งหนึ่ง เพื่อประเมินว่าข้อสอบที่นำมาใช้ในการสอบครั้งนั้นมีคุณสมบัติอย่างไร ทำงานได้ตามที่ต้องการหรือไม่ มีระดับความยากง่ายของข้อสอบเหมาะสมหรือไม่ มีข้อบกพร่องหรือไม่ และควรได้รับการปรับปรุงแก้ไขอย่างไร การวิเคราะห์ข้อสอบเป็นศาสตร์ที่ได้รับการพัฒนาอย่างต่อเนื่องมาเป็นเวลานาน มีเทคนิคและวิธีการต่าง ๆ มากมายที่ผู้วิเคราะห์สามารถใช้เพื่อบอกคุณสมบัติของข้อสอบแต่ละข้อ ตั้งแต่วิธีการง่าย ๆ ไปจนถึงวิธีการที่มีความซับซ้อนมาก โดยแต่ละเทคนิคการวิเคราะห์ก็มีจุดประสงค์แตกต่างกันไป ตั้งแต่การบอกระดับความยากง่าย การบอกถึงความสามารถในการแยกผู้สอบที่เก่งออกจากผู้สอบที่ไม่เก่ง ไปจนถึงเทคนิคขั้นสูงที่สามารถบอกได้ว่าข้อสอบมีความลำเอียงต่อผู้สอบเพศใดเพศหนึ่ง หรือผู้สอบจากสถาบันใดสถาบันหนึ่งเป็นพิเศษหรือไม่ มีการเดาข้อสอบมากน้อยเพียงใด ผู้สอบรู้ข้อสอบมาก่อนเข้าสอบหรือไม่ หรือมีความน่าจะเป็นมากน้อยเพียงใดที่ผู้สอบลอกคำตอบ ในบทความนี้ผู้เขียนไม่ได้ตั้งเป้าประสงค์ที่จะรวบรวมและอภิปรายเทคนิคการวิเคราะห์ข้อสอบทุกวิธีที่มีใช้อยู่ในปัจจุบัน แต่ต้องการเพียงนำเสนอความรู้พื้นฐานที่เกี่ยวกับการวิเคราะห์ข้อสอบและอธิบายถึงวิธีการวิเคราะห์ข้อสอบที่นิยมใช้กันในทางแพทยศาสตรศึกษา โดยเฉพาะในประเทศไทย โดยประสงค์ให้อาจารย์ผู้อ่านสามารถนำเอาความรู้ที่ได้จากบทความนี้ไปใช้แปลผลการวิเคราะห์ข้อสอบที่ตน

เกี่ยวข้อง และดำเนินการปรับปรุงคุณภาพของข้อสอบได้อย่างเหมาะสม

ความรู้พื้นฐานเกี่ยวกับข้อสอบปรนัย

ก่อนที่จะกล่าวถึงรายละเอียดในการวิเคราะห์ข้อสอบ ผู้นิพนธ์ก็จะขอทบทวนความรู้พื้นฐานเกี่ยวกับข้อสอบปรนัยก่อน โดยทั่วไปข้อสอบปรนัยแต่ละข้อมีส่วนประกอบสำคัญ ๒ ส่วนด้วยกันคือ

๑. โจทย์ (stem) เป็นข้อมูลของโรค หรือภาวะหรือผู้ป่วยตามด้วยคำถาม หรือเว้นช่องว่างสำหรับเติมคำหรือข้อความที่เหมาะสมลงไป

๒. ตัวเลือก (options) คือคำ หรือข้อความที่ผู้สอบข้อสอบนำเสนอตามหลังจากโจทย์เพื่อให้ผู้สอบเลือกไปใช้ตอบคำถาม หรือเติมลงในช่องว่างในโจทย์

๒.๑ ตัวเลือกที่ถูกต้อง (correct option) เป็นคำตอบที่ถูกต้องมีเพียงตัวเลือกเดียวต่อข้อสอบข้อหนึ่ง

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

ทฤษฎีที่ใช้ในการวิเคราะห์ข้อสอบ

ทฤษฎีที่ใช้ในการวิเคราะห์ข้อสอบในปัจจุบันนี้มี ๒ ทฤษฎีด้วยกัน ได้แก่ทฤษฎีการสอบแบบดั้งเดิม

๓๑

มกราคม-เมษายน ๒๕๕๓, ปีที่ ๓, ฉบับที่ ๑

เวบบันทึทกสิริราช

บทความทั่วไป

(classical test theory) และทฤษฎีการตอบสนองต่อข้อสอบ (item response theory) ทฤษฎีการสอบแบบดั้งเดิมนั้นเป็นทฤษฎีที่ได้ถูกพัฒนาขึ้นตั้งแต่ตอนต้นของศตวรรษที่ ๒๐ โดยมีการรวบรวมเป็นตำราในครั้งแรกตั้งแต่ปี ค.ศ. ๑๙๒๑ โดย William Brown และ Godfrey H Thomson^๒ หลังจากนั้นทฤษฎีนี้ก็ได้รับการใช้อย่างแพร่หลายในการวิเคราะห์ข้อสอบและได้รับการพัฒนาอย่างต่อเนื่อง ทฤษฎีการสอบแบบดั้งเดิมนี้อาศัยฐานอยู่บนสมมติฐานว่าคะแนนสอบที่ได้มานั้นประกอบไปด้วยคะแนนที่แท้จริง (true score) กับความผิดพลาดจากการวัด (error) ซึ่งสมมติฐานดังกล่าวต่อมาพบว่ามีข้อจำกัดหลายประการด้วยกัน ในราว ค.ศ. ๑๙๗๐ จึงได้มีความพยายามพัฒนาทฤษฎีที่ใช้ในการวิเคราะห์ข้อสอบแบบใหม่ขึ้นซึ่งให้หลักการของความน่าจะเป็นมาวิเคราะห์ข้อสอบ 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 เพราะจะเป็นสิ่งที้อาจารย์แพทย์ทุกท่านจะได้พบและใช้งานเป็นประจำ

การวิเคราะห์ข้อสอบตามทฤษฎีการสอบแบบดั้งเดิม

การวิเคราะห์ข้อสอบตามทฤษฎีการสอบแบบดั้งเดิมนั้นประกอบไปด้วย ๒ ส่วนใหญ่ ๆ คือ (๑) การ

วิเคราะห์ข้อสอบรายข้อ (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 หากข้อสอบข้อนั้นเป็นข้อสอบที่ง่ายผู้สอบทุกคนตอบถูกค่า p ก็จะเป็น ๑ หากไม่มีผู้สอบคนใดตอบถูกเลยข้อสอบข้อนั้นก็จะมีค่า p เป็น ๐ หากมีคนตอบถูก ๗๐% ข้อสอบข้อนั้นก็จะมีค่า 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 $r^{๐,๔}$ ซึ่งสามารถคำนวณได้จากสูตรต่อไปนี้^๗

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

๓๓

นางสาวณิชา นิมิต ๒๕๕๓, ปีที่ ๓, ฉบับที่ ๑

เวบบันทึทศึรศษ

บทควมท่วไป

เมื่อ Mp = คะแนนรวมเฉลี่ยของผู้สอบที่ตอบข้อสอบถูก

Mq = คะแนนรวมเฉลี่ยของผู้สอบที่ตอบข้อสอบผิด

SD = ค่าเบี่ยงเบนมาตรฐาน (standard deviation) ของคะแนนสอบ

p = สัดส่วนของผู้สอบที่ตอบข้อสอบถูกต้องของผู้สอบทั้งหมด

q = สัดส่วนของผู้สอบที่ตอบข้อสอบผิดของผู้สอบทั้งหมด

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^{๑๖} (โดยทั่วไปแล้วข้อสอบที่มีค่า point-biserial ติดลบ ให้สงสัยว่าจะเฉลยผิด)

๑.๓ ประสิทธิภาพของตัวลวง (distractor functionality)

ตัวลวงที่มีประสิทธิภาพนั้นมีคุณสมบัติ ๒ ประการคือ^{๑๗}

(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 หรือลวงให้ผู้ที่มี

ความรู้ดีมาเลือกจัดเป็นตัวลวงที่ไม่ดี สมควรพิจารณาตัดทิ้งหรือปรับเปลี่ยน

๒. การวิเคราะห์ข้อสอบโดยรวม (test analysis)

การวิเคราะห์ข้อสอบโดยรวมเป็นการพิจารณาว่าเมื่อข้อสอบทั้งชุดทำงานร่วมกันแล้วผลสอบที่ได้ออกมาเป็นอย่างไร มีระดับความยากง่ายเป็นอย่างไร มีการกระจายตัวของคะแนนเป็นอย่างไร มีความน่าเชื่อถือของคะแนนสอบมากน้อยเพียงใด ดัชนีต่าง ๆ ที่ต้องพิจารณาได้แก่

๒.๑ ความเที่ยงตรงของคะแนนสอบ (internal consistency reliability)

การประเมินความเที่ยงตรงของคะแนนสอบเป็นการตรวจสอบว่าคะแนนที่ได้ออกมานั้นมีความน่าเชื่อถือเพียงใด เป็นการตอบคำถามว่าหากนำผู้สอบมาสอบใหม่ในสภาวะการเดิม ด้วยข้อสอบที่มีระดับความยากง่ายเท่าเดิม และผู้สอบมีความรู้เท่าเดิมไม่ได้ไปศึกษาหาความรู้เพิ่มเติม จะได้คะแนนสอบเท่าเดิมหรือไม่^{๑๘}

ดัชนีชี้วัดความเที่ยงตรงของคะแนนสอบที่นิยมใช้ในการรายงานผลสอบด้วยข้อสอบปรนัยคือค่าสัมประสิทธิ์ อัลฟา (Coefficient Alpha) ซึ่งสามารถคำนวณได้จากสูตร^{๑๙}

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

เมื่อ α = สัมประสิทธิ์ อัลฟา (Coefficient Alpha)

n = จำนวนชุดย่อยของข้อสอบที่ทำกรแบ่งออกเพื่อหาความเที่ยง

σ_x^2 = การกระจายตัว (variance) ของคะแนนรวม

$\sigma_{x_i}^2$ = การกระจายตัว (variance) ของคะแนนข้อสอบย่อยชุดที่ i

ค่าสัมประสิทธิ์อัลฟานี้มีค่าอยู่ในช่วง ๐ - ๑ ค่าต่ำแสดงว่าคะแนนที่ได้มีความเชื่อถือได้น้อย ไม่แตกต่างไปจากการเดาสุ่ม ค่าสูงแสดงว่าคะแนนที่ได้นั้นมีความน่าเชื่อถือมาก หากทำการทดสอบซ้ำคะแนนที่ได้ก็จะใกล้เคียงเดิม โดยทั่วไประดับของความเที่ยงตรง

๓๓

มกราคม-เมษายน ๒๕๕๒, ปีที่ ๒, ฉบับที่ ๑

เวบบินทีกีรียา

บทความทั่วไป

ของคะแนนสอบที่ยอมรับได้นั้นขึ้นกับว่าต้องการนำเอาคะแนนสอบไปใช้ทำอะไร หากการตัดสินผลสอบนั้นมีความสำคัญมาก (high-stakes examination) เช่น การตัดสินผลสอบขอรับใบประกอบวิชาชีพเวชกรรม หรือประกาศนียบัตรแพทย์ผู้เชี่ยวชาญเฉพาะสาขา มักต้องการคะแนนสอบที่มีค่าสัมประสิทธิ์ อัลฟา ไม่ต่ำกว่า ๐.๙ หากการตัดสินผลสอบนั้นมีความสำคัญปานกลาง (medium-stakes examination) เช่น การสอบลงกอง การสอบเลื่อนชั้นเรียน มักต้องการคะแนนสอบที่มีค่าสัมประสิทธิ์ อัลฟา อยู่ในช่วง ๐.๘ - ๐.๘๙ หากการตัดสินผลสอบนั้นมีความสำคัญน้อย (low-stakes examination) เช่น การสอบย่อยในชั้นเรียน การสอบแบบ formative assessment มักต้องการคะแนนสอบที่มีค่าสัมประสิทธิ์ อัลฟา อยู่ในช่วง ๐.๗ - ๐.๗๙^{๑๒}

ประเด็นสำคัญที่ต้องพิจารณาคือเมื่อได้คะแนนสอบที่มีค่าสัมประสิทธิ์ อัลฟาต่ำ จะต้องดำเนินการอย่างไรเพื่อพัฒนาให้การสอบครั้งต่อไปไม่ประสบปัญหาเรื่องความไม่น่าเชื่อถือของคะแนนสอบอีก ปัจจัยหลักที่จะช่วยเพิ่มความเที่ยงตรงของคะแนนสอบปรนัยมี ๓ ปัจจัยด้วยกัน^{๑๓} คือ

(๑) เพิ่มจำนวนข้อสอบให้มากขึ้น ยังมีข้อสอบมากข้อคะแนนที่ได้ก็จะมีค่าสัมประสิทธิ์เพิ่มขึ้น

(๒) ปรับให้ข้อสอบมีการคละกันของข้อสอบที่ยากและง่ายอย่างเหมาะสม เพื่อปรับให้คะแนนมีการกระจายตัวมากขึ้น หากข้อสอบทั้งชุดประกอบไปด้วยข้อสอบที่ง่ายหมด ผู้สอบเกือบทั้งหมดได้คะแนนสูงมาก จะทำให้มีความแตกต่างของคะแนนน้อย โอกาสที่จะแยกแยะผู้สอบที่มีความรู้ดีออกจากผู้ที่มีความรู้ปานกลาง หรือไม่รู้ดีได้อย่างมั่นใจก็เป็นไปได้น้อย ดังนั้นหากอาจารย์ปรับให้มีการคละกันของข้อสอบยากและง่ายอย่างเหมาะสม ก็จะทำให้ผู้สอบมีระดับคะแนนแตกต่างกันมาก ค่าสัมประสิทธิ์อัลฟาก็จะสูงขึ้นด้วย

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

๒.๒ การกระจายตัวของคะแนน และคะแนน

เฉลี่ย (standard deviation and mean score)

การตรวจดูลักษณะพื้นฐานของคะแนนสอบนี้จะช่วยบอกได้คร่าว ๆ ว่าการเรียนการสอนมีประสิทธิภาพเพียงใด หากอาจารย์สอนได้ดี นักเรียนทั้งชั้นเรียนเข้าใจเนื้อหาดี คะแนนสอบที่ได้ออกมาก็ควรจะกระจายตัวมากนัก (คะแนนเกาะกลุ่มกัน) และคะแนนเฉลี่ยก็ควรจะค่อนข้างสูงเมื่อเทียบกับนักเรียนรุ่นอื่น ๆ หากคะแนนสอบของนักเรียนมีการกระจายตัวมากผิดปกติ แสดงว่าอาจมีปัญหาบางประการในการเรียนการสอนทำให้นักเรียนบางคนมีความรู้ความเข้าใจดี แต่มีนักเรียนบางกลุ่มที่ไม่ค่อยรู้เรื่อง^{๑๔}

๒.๓ ค่าความยากง่ายเฉลี่ยของข้อสอบ (average difficulty)

จากการวิเคราะห์ข้อสอบรายข้อ เราได้ค่าความยากง่ายของข้อสอบแต่ละข้อ (p) เมื่อนำค่า p ของข้อสอบทุกข้อมาหาค่าเฉลี่ย เราก็จะได้ค่าความยากง่ายของข้อสอบทั้งหมด ค่าที่ได้มานี้ใช้เป็นตัวชี้วัดว่าข้อสอบทั้งชุดโดยรวมแล้วมีระดับความยากง่ายเป็นอย่างไร หากผู้สอบเป็นนักศึกษาในกลุ่มใหญ่พอที่เราจะตั้งสมมติฐานว่าระดับความสามารถมีการกระจายตัวอย่างเหมาะสมและไม่ต่างจากระดับความสามารถเฉลี่ยของกลุ่มผู้สอบปีก่อน ๆ เราก็สามารถนำค่าความยากง่ายของข้อสอบทั้งหมดนี้มาเทียบได้ว่าข้อสอบที่นำมาใช้ในปีนี้อาจง่ายกว่าข้อสอบปีก่อน 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^{๑๕}

๓๔

นางสาวเนติชน ๒๕๕๓, ปีที่ ๒๓, ฉบับที่ ๑

เวบบันทึทศิรราช

บทความทั่วไป

การนำผลการวิเคราะห์ข้อสอบไปใช้

ผลการวิเคราะห์ข้อสอบด้วยดัชนีชี้วัดต่าง ๆ ดังกล่าวข้างต้นสามารถนำไปใช้ประโยชน์ได้หลายประการ เช่น

๑. ใช้เป็นประโยชน์ในการปรับแก้คะแนนสอบ

จากผลการวิเคราะห์ข้อสอบจะช่วยชี้แนะให้เรทราบว่าการข้อสอบข้อใดน่าจะเฉลยผิด ข้อสอบข้อใดน่าจะมีคำตอบที่ถูกมากกว่า ๑ ตัวเลือก ข้อสอบข้อใดน่าจะมีปัญหา เช่น มีความคลุมเครือในคำถาม หรือตัวเลือกมีความซ้ำซ้อนกัน หรือเนื้อหาของข้อสอบอยู่นอกเหนือไปจากสิ่งที่สอนนักเรียน เป็นต้น ข้อสอบที่มีปัญหาเหล่านี้ต้องได้รับการประเมินโดยคณะกรรมการตรวจข้อสอบซึ่งประกอบไปด้วยอาจารย์ผู้มีความรู้ความชำนาญในเนื้อหาวิชาที่ทำการสอบว่าจะดำเนินการอย่างไรกับการคิดคะแนน หากปัญหาที่พบมีความรุนแรงไม่มากจนทำให้การตัดสินใจเลือกคำตอบที่ถูกต้องเปลี่ยนไป คณะกรรมการอาจพิจารณาคิดคะแนนของข้อสอบข้อนั้นตามปกติ หากข้อสอบเฉลยผิดคณะกรรมการสามารถพิจารณาแก้คำตอบแล้วทำการตรวจให้คะแนนข้อสอบข้อนั้นใหม่ หากข้อสอบข้อใดมีคำตอบที่เหมาะสม ๒ ข้อ คณะกรรมการอาจพิจารณาให้ผู้สอบที่ตอบข้อใดข้อหนึ่งใน ๒ ข้อดังกล่าวได้คะแนนในข้อนั้น หากข้อสอบนั้นมีความคลุมเครือมากจนไม่สามารถตัดสินใจเลือกคำตอบที่เหมาะสมได้ คณะกรรมการสามารถตัดข้อสอบข้อนั้นออกจากการคิดคะแนน และปรับคะแนนเกณฑ์ผ่านลดลงตามความเหมาะสม

๒. ใช้เป็นประโยชน์ในการปรับปรุงคุณภาพข้อสอบ

ภายหลังจากการรายงานคะแนนสอบเป็นที่เรียบร้อยแล้ว คณะกรรมการสอบสามารถนำผลการวิเคราะห์ข้อสอบแต่ละข้อมาพิจารณาโดยละเอียดเพื่อดูว่าข้อสอบข้อใดสมควรได้รับการปรับปรุงแก้ไข ข้อสอบที่พบว่ายากเกินไปอาจเกิดจากโจทย์คำถามมีความคลุมเครือ ต้องทำการปรับแก้ให้โจทย์ชัดเจนขึ้น หรือเพิ่มเติมข้อมูลบางประการเข้าไปเพื่อให้การวินิจฉัย

ชัดเจนขึ้น ข้อสอบที่พบว่าง่ายเกินไปอาจพิจารณาปรับให้ยากขึ้นโดยการแก้ไขโจทย์หรือตัวเลือก ข้อสอบที่มีค่า point-biserial ต่ำมักเกิดจากโจทย์ที่คลุมเครือ สร้างความสับสนให้ผู้สอบ สมควรได้รับการปรับโจทย์คำถามใหม่

นอกจากนี้อาจารย์ยังต้องพิจารณาถึงการทำงานของตัวเลือกด้วย ปัญหาที่พบบ่อยมากในการวิเคราะห์ข้อสอบปรนัยคือมีตัวลวงจำนวนมากที่ไม่ทำงาน (มีผู้สอบเลือกน้อยมาก หรือลวงเฉพาะผู้ที่มีความรู้ดีให้มาเลือก) จากการศึกษาวิจัยข้อสอบปรนัยจำนวนมากพบว่าข้อสอบส่วนใหญ่มีตัวลวงที่ทำงานจริงเพียง ๓ ตัวเลือกเท่านั้น^๕ ตัวลวงที่เหลือเป็นตัวลวงที่ไม่มีประโยชน์ พิมพ์ลงมาให้ข้อสอบก็เป็นการเปลืองเนื้อที่หน้ากระดาษ และเสียเวลาอ่านโดยใช่เหตุ อาจารย์ควรพิจารณาตัดตัวลวงที่ไม่ทำงานออกเสียหรือเปลี่ยนเป็นตัวลวงอื่นที่น่าจะมีประสิทธิภาพมากขึ้น

๓. ใช้เป็นประโยชน์ในการบริหารคลังข้อสอบ

ข้อสอบแต่ละข้อนั้นได้มาด้วยความยากลำบาก อาจารย์แต่ละท่านต้องใช้เวลาและความคิดอย่างมากเพื่อพัฒนาข้อสอบที่ดีขึ้นมาใช้ ดังนั้นเมื่อนำข้อสอบมาใช้แล้วผลการวิเคราะห์ข้อสอบแสดงว่าข้อสอบข้อใดเป็นข้อสอบที่ดี มีระดับความยากง่ายเหมาะสม มีความสามารถในการจำแนกผู้สอบที่ดีก็ควรพิจารณาเลือกเก็บข้อสอบดังกล่าวไว้ในคลังข้อสอบเพื่อที่จะได้นำกลับมาใช้ใหม่ในอนาคต ในการเก็บข้อสอบเข้าในคลังข้อสอบก็ต้องมีการแนบข้อมูลเกี่ยวกับประวัติการใช้งานและผลการวิเคราะห์ข้อสอบในแต่ละครั้งไว้คู่กันด้วย เพื่อที่จะได้เป็นประโยชน์ในการเลือกข้อสอบมาใช้ใช้งาน หากอาจารย์ต้องการข้อสอบที่มีระดับความยากง่าย หรือความสามารถในการจำแนกผู้สอบมากนักน้อยเพียงใดจะได้ดึงเอาข้อสอบที่มีคุณลักษณะตามต้องการออกมาใช้ได้ตามต้องการ

๔. ใช้เป็นประโยชน์ในการพัฒนาคุณภาพการสอน

การพิจารณาผลการวิเคราะห์ข้อสอบโดยละเอียดในหัวข้อที่อาจารย์ท่านใดท่านหนึ่งรับผิดชอบ

๓๕

มกราคม-เมษายน ๒๕๕๓, ปีที่ ๒, ฉบับที่ ๑

เวบบินทักสิริราช

บทความทั่วไป

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^{๓๓,๓๔} หากได้ข้อมูลมาจากกลุ่มตัวอย่างที่มีขนาดใหญ่พอและมีการกระจายตัวของระดับความสามารถของผู้สอบที่เหมาะสม ค่าต่าง ๆ ที่ได้ (p , r , coefficient alpha) จะค่อนข้างเที่ยงตรง ปัญหาที่สำคัญในการวิเคราะห์ข้อสอบในโรงเรียนแพทย์คือการสอบจำนวนมากจัดในนักศึกษาในกลุ่มเล็ก และ

นักศึกษาแต่ละกลุ่มก็มีการกระจายตัวของระดับความสามารถแตกต่างกัน นักศึกษาบางกลุ่มมีความสามารถสูงกว่านักศึกษากลุ่มอื่น ดังนั้นผลการวิเคราะห์ข้อสอบไม่ว่าจะเป็นค่า p , r , coefficient alpha, mean, หรือ standard deviation อาจเปลี่ยนแปลงไปในแต่ละกลุ่มของนักศึกษา ดังนั้นการนำผลการวิเคราะห์ข้อสอบไปใช้ในทางปฏิบัติจึงมีข้อควรระวังดังต่อไปนี้

การพิจารณาว่าข้อสอบยากหรือง่ายโดยใช้ค่า p นั้นเป็นค่าที่ไม่คงที่ ขึ้นอยู่กับกลุ่มผู้สอบ หากนำข้อสอบข้อหนึ่งไปใช้กับนักเรียนกลุ่มที่มีความรู้ดี นักเรียนส่วนใหญ่จะทำข้อสอบได้ถูกต้องทำให้ค่า p สูง แต่เมื่อนำข้อสอบข้อเดิมไปใช้กับนักเรียนกลุ่มที่ความรู้ไม่ดีนัก สัดส่วนของนักเรียนที่ทำข้อสอบข้อเดียวกันได้ถูกต้องจะลดลงทำให้ค่า p ลดลง นอกจากนี้ในข้อสอบที่เน้นการท่องจำที่เคยใช้แล้ว เมื่อนำกลับมาใช้ใหม่ในนักเรียนกลุ่มใหม่ อาจมีนักเรียนจำนวนหนึ่งที่สามารถตอบข้อสอบถูกต้องเนื่องจากรู้ข้อสอบมาก่อนก็จะทำให้ค่า p สูงขึ้นกว่าเดิมได้

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

ค่าสัมประสิทธิ์อัลฟา เป็นค่าที่มีความเฉพาะเจาะจงกับการสอบของนักเรียนกลุ่มใดกลุ่มหนึ่งเท่านั้น หากใช้เป็นคุณสมบัติติดตัวข้อสอบแต่ละข้อไม่ หากข้อสอบชุดหนึ่งทำการสอบกับนักเรียนกลุ่มหนึ่งแล้วพบว่าคะแนนสอบที่ได้มานั้นมีค่าสัมประสิทธิ์อัลฟาสูงในระดับที่ต้องการก็ไม่ได้เป็นตัวรับประกันว่าหากนำข้อสอบชุดเดิมนั้นไปทำการสอบกับนักเรียนกลุ่มอื่นจะได้ค่าสัมประสิทธิ์อัลฟาที่สูงเช่นเดียวกัน นอกจากนี้ค่าสัมประสิทธิ์อัลฟาที่สูงไม่ได้เป็นตัวบอกถึงคุณภาพของข้อสอบรายข้อแต่อย่างใด

ค่าสัมประสิทธิ์อัลฟาที่สูงช่วยบอกแค่เพียงว่า

๓๖

กรกฎาคม-สิงหาคม ๒๕๕๓, ปีที่ ๒, ฉบับที่ ๑

เวบบิ้นทีกีธีรราช

บทความทั่วไป

คะแนนสอบในข้อสอบข้อหนึ่งมีความผันแปรไปในทิศทางเดียวกันกับคะแนนสอบในข้อสอบข้ออื่นในการสอบชุดเดียวกัน นั่นคือในข้อสอบชุดที่มีค่าสัมประสิทธิ์อัลฟ่าสูงก็อาจประกอบไปด้วยข้อสอบที่ดี และข้อสอบที่ไม่ดีรวมกันอยู่ ต้องไปตรวจสอบดัชนีชี้วัดคุณภาพของข้อสอบตัวอื่น ๆ ในแต่ละข้ออีกครั้ง

ข้อควรจำในการวิเคราะห์ข้อสอบที่ผู้นิพนธ์ข้อย้าในตอนท้ายของบทความนี้ก็คือค่าดัชนีชี้วัดคุณภาพต่าง ๆ ของข้อสอบที่กล่าวมาทั้งหมดนี้เป็นเพียงตัวช่วยให้อาจารย์เข้าใจข้อสอบดีขึ้นและช่วยแนะแนวทางในการพัฒนาปรับปรุงข้อสอบให้ดีขึ้น ดัชนีเหล่านี้ไม่ใช่ค่าตัดสินหรือตัวชี้ชะตาของข้อสอบ ไม่มีดัชนีใดที่ได้จากการวิเคราะห์ข้อสอบจะมาทดแทนดุลยพินิจของอาจารย์ไปได้ ดัชนีคุณภาพของข้อสอบไม่ว่าจะคำนวณมาด้วยวิธีการที่ถูกต้องแล้วก็ตามก็เป็นเพียงตัวเลขที่สามารถเกิดความผิดพลาดในการแปลผลได้ดังเช่นการแปลผลการวิเคราะห์ทางสถิติต่าง ๆ บทบาทของอาจารย์ในการวิเคราะห์ข้อสอบคงไม่ใช่การยึดถือตัวเลขดัชนีต่าง ๆ เป็นกฎตายตัว หากแต่ใช้ดัชนีเหล่านี้ช่วยเป็นแนวทางในการพิจารณาข้อสอบ หากดัชนีตัวใดระบุว่าข้อสอบอาจมีปัญหา อาจารย์ก็นำข้อสอบนั้นมาพิจารณากันโดยคณะกรรมการข้อสอบ หากหลังจากการพิจารณาโดยถี่ถ้วนแล้วอาจารย์คิดว่าข้อสอบข้อนั้นเหมาะสมแล้ว ไม่ควรทำการปรับแก้เนื้อหา อาจารย์ก็ยืนยันไปว่าไม่แก้ไข อาจารย์คงไม่ตัดสินการรักษาผู้ป่วยโดยใช้ผลเลือดตัวใดตัวหนึ่งเป็นเกณฑ์โดยไม่พิจารณาอาการและอาการแสดงของผู้ป่วยร่วมด้วย ฉันทัดก็ฉันทัน อาจารย์

ไม่ควรตัดสินชะตากรรมของข้อสอบโดยใช้เพียงค่า p หรือ r โดยไม่พิจารณาความเหมาะสมของเนื้อหาโจทย์และตัวเลือกต่าง ๆ ในข้อสอบข้อนั้น

เอกสารอ้างอิง

- Livingston SA. Item analysis. In: Downing SM, Haladyna TM, eds. Handbook of test development. Mahwah, NJ: Lawrence Erlbaum Associates; 2006:421-41.
- Brown W, Thomson GH. The essentials of mental measurement, 2nd ed. Cambridge, England: University Press; 1921.
- Yen WM, Fitzpatrick AR. Item response theory. In: Brennan RL, ed. Educational measurement, 4th ed. Westport, CT: Praeger Publishers; 2006:111-53.
- Haladyna TM. Writing test items to evaluate higher order thinking. Boston, MA: Allyn and Bacon; 1997.
- Haladyna TM. Writing multiple choice items. Chicago, IL: CAT Inc.; 2003.
- Haladyna TM. Developing and validating multiple-choice test items, 3rd ed. Mahwah, NJ: Lawrence Erlbaum Associates; 2004.
- 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.
- Haladyna TM, Downing SM. How many options is enough for a multiple-choice test item? Educ Psychol Meas 1993;53:999-1010.
- Gronlund NE. Assessment of student achievement, 7th ed. Boston: Allyn & Bacon, 2003.
- Linn RL, Miller MD. Measurement and assessment in teaching, 9th ed. Upper Saddle River, NJ: Prentice Hall, 2004.
- Haertel EH. Reliability. In: Brennan RL, editor. Educational measurement, 4th ed. Westport, CT: Praeger Publishers; 2006:65-110.
- Downing SM. Reliability: On the reproducibility of assessment data. Med Educ 2004;38:1006-12.
- Embretson SE, Reise SP. Item response theory for psychologists. Mahwah, NJ: Lawrence Erlbaum Associates, 2000.
- 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)

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

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 $\sqrt{1-rtt}$)

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

16 กรกฎาคม 2563

Constructed response item development

Constructed response item development

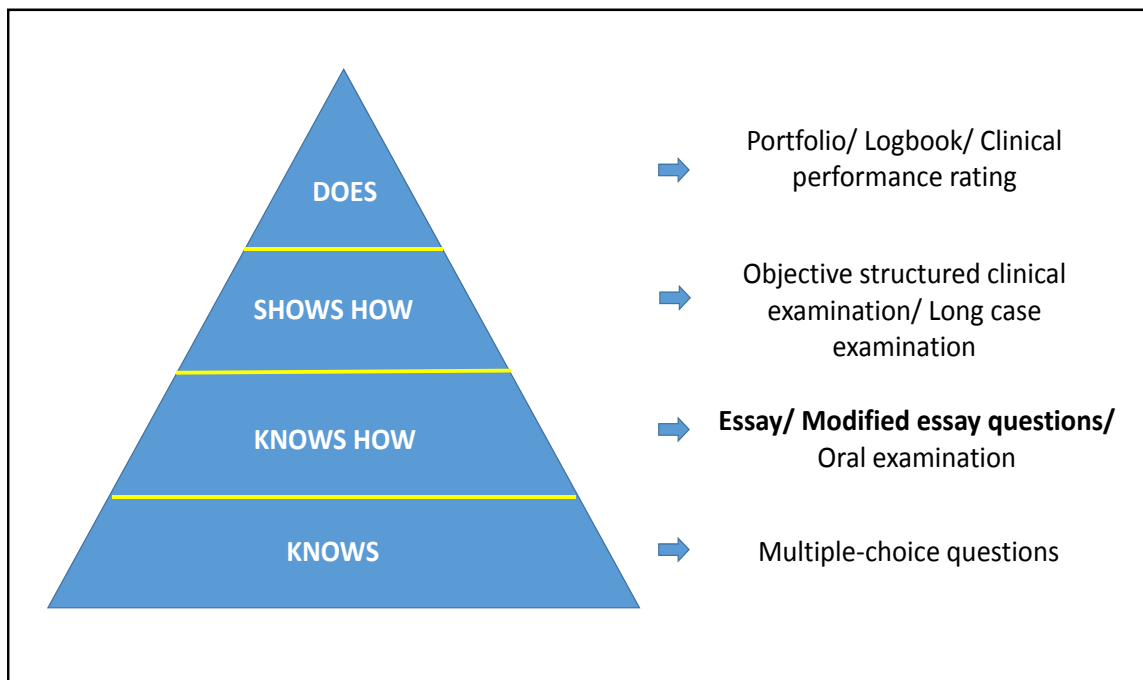
Assistant Professor Thos Harnroongroj
Department of Orthopaedics and Rehabilitation, Faculty of Medicine,
Siriraj Hospital, Mahidol University

Objectives

- Could describe the types of constructed response item question.
- Know the process of developing constructed response item question.

Written testing formats

- Selected response item
- Constructed response item



Selected response item VS constructed response item

	Selected response item	Constructed response item
Measured construct	Concrete knowledge, basic interpretation, some applications	Complex cognitive ability
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.

Types of constructed response item

Constructed response item

Constructed response item



Traditional essay questions:

- Long essay
- Short essay



Modified essay questions (MEQs):

- Standard modified essay questions
- Patient management problem
- Key features problem
- Short answer questions

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

Long Essay Questions (LEQs)

- Open-ended, unstructured questions
- Assess
 - Students' understanding
 - Writing ability

Journal of Educational Research & Medical Teacher 2015;3(1):8-12

Short Essay Questions (SEQs)

- Open-ended, structured question
- Expect specific answer
- Test the knowledge of
 - Analyzing
 - Reasoning
 - Application
 - Integration

Journal of Educational Research & Medical Teacher 2015;3(1):8-12

LEQs VS SEQs

	Long essay	Short essay
Content coverage	Narrow	Broad
Item development	Easy	Difficult
Scoring guideline development	Very difficult	Easier
Students' answers	Infinite possibilities	More focus
Reliability	Very low	Low
Time used	More	Less
Good use for	Complex cognitive abilities: Analysis, synthesis, evaluation and idea presentation	Assessment of simplified, structured problems with limited answers

Examples

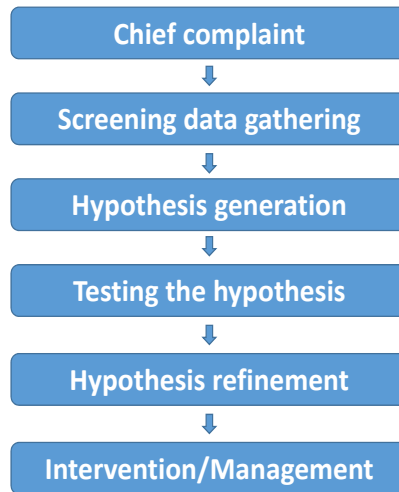
- Please provide the differential diagnosis of right lower quadrant abdominal pain
- Please explain about open fracture
- Please compare the difference between skin traction and skeletal traction

Clinical problem solving methods

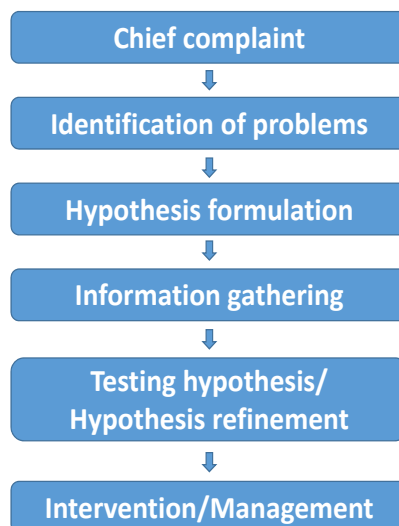
- Pattern recognition
- Algorithm
- Hypothesis testing
 - Forward reasoning (data driven process)
 - Backward reasoning (hypothesis driven process)

วิชญ์ ธรรมลิขิตกุล การประเมินความรู้ในการแก้ปัญหาผู้ป่วยทางคลินิก. สารศิริราช 2534, 43(2): 123-134.

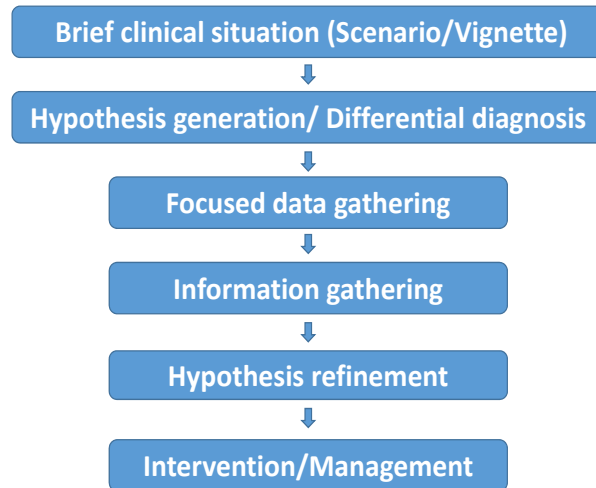
Forward reasoning method



Backward reasoning method



MEQ process



MEQ assessment

- Problem solving
- Decision making

Characteristics of MEQ

- Real life scenario
- Serial question and answer
- Serial additional information
- Irreversible

Standard MEQ construct

Chief complaint

Question on differential diagnosis
Question for collecting additional information

Additional information (1)

Question on provisional diagnosis
Question on further investigation

Additional information (2)

Question on investigation interpretation
Question on definite diagnosis
Question on management
Explore reasoning and idea.

History

- 30 year-old Thai man sustained motorcycle accident 30 minutes PTA.
- He was transferred to the emergency room by EMS.

Please provide the provisional diagnosis and initial management.



Provisional diagnosis

Open fracture right tibia, Gustilo-Anderson type 3 (at least)

Initial evaluation

- Primary survey: no immediate life threatening condition
- Secondary survey:
 - BP 130/80 mmHg PR 110/min
 - No associated injury
 - Right leg:
 - lacerated wound 10 cm at mid leg
 - Exposed proximal fragment of tibia
 - Distal neurovascular status was intact

Please send the proper radiographic investigation.



Please interpret the radiographs.

- Film right tibia AP, Lateral
 - Transverse fracture of mid-shaft of right tibia and fibular

Diagnosis

- Open fracture of mid-shaft of right tibia
- Gustilo-Anderson type IIIB

Please describe the treatment in this patient.

Key features problem (KFP)

- Clinical decision making skills
- Identification of critical steps
- Reliability of 0.8 in 4-hour testing

Developing MEQs

Steps

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

วิชญ์ ธรรมลิขิตกุล การประเมินความรู้ในการแก้ปัญหาผู้ป่วยทางคลินิก. สารศิริราช 2534, 43(2): 123-134.

Assembling problem-writing groups

- Item writers: Well experience/ Multidisciplinary
- Written problems: Well grounded/ Real life experience
- Group review

Selecting problems

- Base on table of specification.
- Appropriate problems:
 - Common problems/symptomatology
 - Pitfall tasks
 - Multi-system
- Emphasize on problem solving or decision making.

Defining key features

- Brainstorming in the group
 - Critical points
 - Medical ethics issues
- Commonly
 - Further history
 - Further examination
 - Further investigation
 - Describe the treatment.

Writing the questions

- Number of questions
 - Mostly 2-4 questions
 - 1 question per 1 key feature
- Number of answers
 - Vary 1-10
 - Typically 3-5

Selecting question formats

- Clear and specific
- Open-ended
- Examples:
 - Please provide the only one most likely diagnosis in this patient
 - Please give 3 most helpful further investigation
 - Please explain definitive treatment for this condition

Preparing scoring keys

- List the correct and incorrect responses
- Assign score in each response
 - Multiple answers: Weight the proportion of the score
 - One acceptable answer
- Penalty
 - Harmful treatment/ decision making
 - Unnecessary investigation/ treatment: depends on the committee.
 - Not cross the item

Timing

- Try doing the examination yourself.
- Add the time by 30%-50%.

Validation and references

- Validation:
 - Pilot the test within group
 - Discussion and revision
- References

Conclusions

- Types of constructed response item question
- Steps of developing constructed response item question

กระดาษบันทึก

เอกสารประกอบการอบรม



17 กรกฎาคม 2563

17 กรกฎาคม 2563

OSCE item development

OSCE Item Development

เชิดศักดิ์ ไอรรมณีรัตน์
ภาควิชาศัลยศาสตร์ คณะแพทยศาสตร์ศิริราชพยาบาล
มหาวิทยาลัย มหิดล

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 (คำแนะนำผู้ปฎิบัติมาตรฐาน)
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 มีอาการปวดท้องได้ชายโครงด้านซ้าย

6 ชั่วโมง ปวดตลอดเวลา

คำสั่ง : จงแสดงวิธีการตรวจหน้าท้องผู้ป่วย บรรยายสิ่งที่ตรวจพบและให้การวินิจฉัยโรคที่คิดถึงมากที่สุด 1 โรค

เวลา : 5 นาที (ตรวจร่างกาย 4 นาทีครึ่ง บอกสิ่งที่พบและวินิจฉัยครึ่งนาที)

Scenario 1 (cont.)

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

Instruction for Examinees

- ผู้ป่วยหญิงไทย 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

History

- Programmed patients (Barrows & Abrahamson, 1964)
- Simulated patients (Barrows, 1971)
- Patient instructors (Stillman, 1976)
- Simulated patients-based exam (Harden et al, 1975)
- Standardized patients (Barrows, 1993)

Perkowski LC. Standardized patients. In: Distlehorst LH, Dunnington GL, Foise JR. Teaching and learning in medical and surgical education: Lessons learned for the 21st century. Routledge, 2000.

Instruction for SPs

- 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.)

Instruction for SPs

- โจทย์:** นักศึกษาจะทำการซักประวัติท่านเพื่อให้การวินิจฉัยโรคให้ท่านให้ข้อมูลต่อไปนี้
- ข้อมูลจากโจทย์:** ท่านเป็นผู้ป่วยชายไทย อายุ 40 ปี มีอาการปวดขาหน้าข้างขวา 1 วัน การแต่งกาย: แต่งกายชุดลำลอง เป็นเสื้อ กางเกงที่สามารถเปิดหน้าท้องได้สะดวก การตกแคงบาดแผล: ไม่มี
- ข้อมูลที่นักศึกษาจะซักถามจากท่าน**
1. ตำแหน่งที่ปวดท้อง : ปวดบริเวณขาหน้าด้านขวา
 2. ลักษณะของอาการปวด : ช่วงแรกปวดทรมานๆ ตลอดเวลา
 3. มีอาการปวดร้าวไปที่อื่นหรือไม่ : ไม่มี
 4. ลักษณะของอาการปวดตอนเริ่มแรก เป็นอย่างไร เป็นทันทีทันใดหรือค่อยๆปวดเพิ่มขึ้นช้าๆ เป็นที่ตำแหน่งเดียวกันนี้หรือมีการย้ายที่ปวด : ค่อยๆปวดเพิ่มขึ้นช้าๆ ไม่มีการย้ายที่ปวด
 5. มีปัจจัยใดที่ทำให้ปวดเปลี่ยนแปลงหรือไม่ : ปวดเพิ่มมากขึ้นในขณะยืนหรือโอ

Instruction for SPs (cont.)

- 6.อาการร่วมอื่นๆ
 - 6.1 หัวใจ: มีไข้ต่ำๆ
 - 6.2 ระบบทางเดินอาหาร: มีอาการปวดท้องเป็นๆเป็นพักๆ คลื่นไส้และอาเจียน
- 7. ประวัติอดีต
 - 7.1 ประวัติการมีก้อนที่ขาหนีบ
สังเกตมีก้อนที่ขาหนีบข้างขวา มา 2 ปี
 - 7.2 ประวัติการเปลี่ยนแปลงของก้อนที่ขาหนีบ
ขนาดก้อนเท่าเดิม จะโตมากเวลารืนหรือเง็ง เวลานอนแล้ว ก้อนจะยุบได้เอง
 - ...
- 9. ประวัติส่วนตัว : อาชีพ การสูบบุหรี่ การดื่มสุรา
ทำงานเป็นเสมียน สูบบุหรี่วันละ 2 ซองมา 10 ปี ไม่ดื่มสุรา

Scoring Rubric General Format

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

Scoring Rubric

- กระชับ ได้ใจความ สื่อความหมายตรงกัน
- กำหนดประเด็นที่สำคัญ หรือเป็นจุดที่มักทำผิดพลาด
- บรรยายพฤติกรรมที่ผู้ประเมินสังเกตได้
- กำหนดน้ำหนักคะแนนตามความสำคัญ

17 กรกฎาคม 2563

Long case examination

Long Case Examination

Cherdsak Iramaneerat
Department of Surgery
Faculty of Medicine Siriraj Hospital

A Long Case Exam

- While OSCE focuses on individual components of clinical competence, it is widely agreed that there is still a need for assessing students on patient care as a whole.

Long Case Examination

- The examinees spend a long period of time (usually about an hour) to explore and work up a single patient case in detail.
- An examiner assesses history taking, physical examination, communication skills, diagnostic skills, plan of investigations, management, and professionalism of the examinees

Outline

- Objectives
- Advantages and limitations
- Objective Structured Long Case Examination Record (OSLER)
- Long case exam in Thailand

Assessment Objectives

- Knowledge
 - Lower order: Recall, Comprehension, Application
 - Higher order: Analysis, Synthesis, Evaluation
- Psychomotor skills
- Attitudes

Long Case Examination

- Advantages
 - Comprehensive competency evaluation
 - In-depth exploration of knowledge, skills

Long Case Examination

- Disadvantages
 - Subjective ratings
 - Unstructured settings
 - Adequacy of observation
 - Case specificity: construct underrepresentation
 - Fairness among students: A luck of draw
 - Time commitment from medical teachers
 - Low reliability
 - Divergence of objectives: oral examination

OSLER

- Objective Structured Long Case Examination Record (OSLER)
 - Ten items structured record
 - History taking
 - Physical exam
 - Investigation, management, clinical acumen
 - Objectivity: prior agreement on what to be examined
 - Assess both processes and products
 - Identification of case difficulty by an examiner

Gleeson F. Assessment of clinical competence using the Objective Structured Long Examination Record (OSLER), Medical Teacher 1997, 19: 7 – 14.

OSLER's components

- History taking
 - Clarity of presentation, communication process, systematic approach, establishment of case facts.
- Physical examination
 - Systematic approach, examination technique, establishment of correct physical findings.
- Investigations, Management, Clinical acumen
 - Ability to identify and solve problems

The Case Difficulty

- **Standard case**
 - Single problem
- **Difficult case**
 - Up to three problem
- **Very difficult case.**
 - More than three problem

Awarding marks in the OSLER

- P+: Very good/excellent. (60-80%)
- P: Pass/ bare pass. (50-55%)
- P-: Below pass
 - Each items has to be graded followed by overall grade of the complete performance

OBJECTIVE STRUCTURED LONG EXAMINATION RECORD (OSLER)				DATE: _____																																																																																																
CANDIDATE: _____ NAME		EXAMINATION NO. _____																																																																																																		
Examiner to sign in LEAD, each of the ten items below and assign an overall GRADE and MARK according to the scales. EXER is discussed with the candidate as follows:																																																																																																				
	GRADE	MARKS	COMMENTS																																																																																																	
P+	VERY GOOD/EXCELLENT	(80-100)	See next page for guidelines																																																																																																	
P	FAIR/GOOD/BEFORE PASS	(50-75)	See next page for guidelines																																																																																																	
P-	BELOW PASS	(0-40)	See next page for guidelines																																																																																																	
<table border="1" style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th>ITEMS</th> <th>GRADE</th> <th>MARKS</th> <th>ACCOMPLISH</th> </tr> </thead> <tbody> <tr> <td>IDENTIFICATION OF HISTORY</td> <td></td> <td></td> <td></td> </tr> <tr> <td>FACTUALITY</td> <td></td> <td></td> <td></td> </tr> <tr> <td>COMMUNICATION PROCESSES</td> <td></td> <td></td> <td></td> </tr> <tr> <td>SYSTEMATIC PRESENTATION</td> <td></td> <td></td> <td></td> </tr> <tr> <td>SYSTEMATIC INVESTIGATION</td> <td></td> <td></td> <td></td> </tr> <tr> <td>CORRECT FACTS ESTABLISHED</td> <td></td> <td></td> <td></td> </tr> <tr> <td>PHYSICAL EXAMINATION</td> <td></td> <td></td> <td></td> </tr> <tr> <td>SYSTEMATIC</td> <td></td> <td></td> <td></td> </tr> <tr> <td>TECHNIQUE</td> <td></td> <td></td> <td></td> </tr> <tr> <td>CORRECT FINDINGS ESTABLISHED</td> <td></td> <td></td> <td></td> </tr> <tr> <td>APPROPRIATE INVESTIGATIONS BY A LOGICAL SEQUENCE</td> <td></td> <td></td> <td></td> </tr> <tr> <td>APPROPRIATE MANAGEMENT</td> <td></td> <td></td> <td></td> </tr> <tr> <td>CLINICAL ACUMEN</td> <td></td> <td></td> <td></td> </tr> <tr> <td colspan="5">ADDITIONAL COMMENTS: _____</td> </tr> <tr> <td colspan="5"> <table border="1" style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th colspan="2">MARKS FOR THE CASE DIFFICULTY</th> <th colspan="2">INDIVIDUAL EXAMINER</th> <th colspan="2">PAIR OF EXAMINERS</th> </tr> <tr> <th>Standard</th> <th>Difficult</th> <th>Grade</th> <th>Mark</th> <th>Grade</th> <th>Mark</th> </tr> </thead> <tbody> <tr> <td>_____</td> <td>_____</td> <td>_____</td> <td>_____</td> <td>_____</td> <td>_____</td> </tr> <tr> <td>Overall Grade</td> <td>_____</td> <td>Overall Mark</td> <td>_____</td> <td>Overall Grade</td> <td>_____</td> </tr> <tr> <td>Very Difficult</td> <td>_____</td> <td>Overall Mark</td> <td>_____</td> <td>Overall Grade</td> <td>_____</td> </tr> </tbody> </table> </td> </tr> </tbody> </table>					ITEMS	GRADE	MARKS	ACCOMPLISH	IDENTIFICATION OF HISTORY				FACTUALITY				COMMUNICATION PROCESSES				SYSTEMATIC PRESENTATION				SYSTEMATIC INVESTIGATION				CORRECT FACTS ESTABLISHED				PHYSICAL EXAMINATION				SYSTEMATIC				TECHNIQUE				CORRECT FINDINGS ESTABLISHED				APPROPRIATE INVESTIGATIONS BY A LOGICAL SEQUENCE				APPROPRIATE MANAGEMENT				CLINICAL ACUMEN				ADDITIONAL COMMENTS: _____					<table border="1" style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th colspan="2">MARKS FOR THE CASE DIFFICULTY</th> <th colspan="2">INDIVIDUAL EXAMINER</th> <th colspan="2">PAIR OF EXAMINERS</th> </tr> <tr> <th>Standard</th> <th>Difficult</th> <th>Grade</th> <th>Mark</th> <th>Grade</th> <th>Mark</th> </tr> </thead> <tbody> <tr> <td>_____</td> <td>_____</td> <td>_____</td> <td>_____</td> <td>_____</td> <td>_____</td> </tr> <tr> <td>Overall Grade</td> <td>_____</td> <td>Overall Mark</td> <td>_____</td> <td>Overall Grade</td> <td>_____</td> </tr> <tr> <td>Very Difficult</td> <td>_____</td> <td>Overall Mark</td> <td>_____</td> <td>Overall Grade</td> <td>_____</td> </tr> </tbody> </table>					MARKS FOR THE CASE DIFFICULTY		INDIVIDUAL EXAMINER		PAIR OF EXAMINERS		Standard	Difficult	Grade	Mark	Grade	Mark	_____	_____	_____	_____	_____	_____	Overall Grade	_____	Overall Mark	_____	Overall Grade	_____	Very Difficult	_____	Overall Mark	_____	Overall Grade	_____
ITEMS	GRADE	MARKS	ACCOMPLISH																																																																																																	
IDENTIFICATION OF HISTORY																																																																																																				
FACTUALITY																																																																																																				
COMMUNICATION PROCESSES																																																																																																				
SYSTEMATIC PRESENTATION																																																																																																				
SYSTEMATIC INVESTIGATION																																																																																																				
CORRECT FACTS ESTABLISHED																																																																																																				
PHYSICAL EXAMINATION																																																																																																				
SYSTEMATIC																																																																																																				
TECHNIQUE																																																																																																				
CORRECT FINDINGS ESTABLISHED																																																																																																				
APPROPRIATE INVESTIGATIONS BY A LOGICAL SEQUENCE																																																																																																				
APPROPRIATE MANAGEMENT																																																																																																				
CLINICAL ACUMEN																																																																																																				
ADDITIONAL COMMENTS: _____																																																																																																				
<table border="1" style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th colspan="2">MARKS FOR THE CASE DIFFICULTY</th> <th colspan="2">INDIVIDUAL EXAMINER</th> <th colspan="2">PAIR OF EXAMINERS</th> </tr> <tr> <th>Standard</th> <th>Difficult</th> <th>Grade</th> <th>Mark</th> <th>Grade</th> <th>Mark</th> </tr> </thead> <tbody> <tr> <td>_____</td> <td>_____</td> <td>_____</td> <td>_____</td> <td>_____</td> <td>_____</td> </tr> <tr> <td>Overall Grade</td> <td>_____</td> <td>Overall Mark</td> <td>_____</td> <td>Overall Grade</td> <td>_____</td> </tr> <tr> <td>Very Difficult</td> <td>_____</td> <td>Overall Mark</td> <td>_____</td> <td>Overall Grade</td> <td>_____</td> </tr> </tbody> </table>					MARKS FOR THE CASE DIFFICULTY		INDIVIDUAL EXAMINER		PAIR OF EXAMINERS		Standard	Difficult	Grade	Mark	Grade	Mark	_____	_____	_____	_____	_____	_____	Overall Grade	_____	Overall Mark	_____	Overall Grade	_____	Very Difficult	_____	Overall Mark	_____	Overall Grade	_____																																																																		
MARKS FOR THE CASE DIFFICULTY		INDIVIDUAL EXAMINER		PAIR OF EXAMINERS																																																																																																
Standard	Difficult	Grade	Mark	Grade	Mark																																																																																															
_____	_____	_____	_____	_____	_____																																																																																															
Overall Grade	_____	Overall Mark	_____	Overall Grade	_____																																																																																															
Very Difficult	_____	Overall Mark	_____	Overall Grade	_____																																																																																															

Examination Time

- Examiner – candidate time must be sufficient to allow for a valid assessment.
- Identical time should be allowed for all candidates in the interest of examination reliability.
- A minimum of 20 minutes should be allowed.
- For high-stakes exam: 30 minutes is recommended.

National Medical Licensing Examination

- 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

Long Case Examination

- ข้อกำหนดของ ศร. ในการสอบ long case examination
 1. จำนวนผู้ป่วยอย่างน้อย 2 ราย
 2. โรค หรือ ปัญหาสอดคล้องกับเกณฑ์มาตรฐานผู้ประกอบการวิชาชีพเวชกรรมของแพทยสภา
 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

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)

Level of Competencies

- Very good
 - ความถูกต้องครบถ้วนมากกว่าร้อยละ 80
- Good
 - ความถูกต้องครบถ้วนร้อยละ 60 – 80
- Require improvement
 - ความถูกต้องครบถ้วนน้อยกว่าร้อยละ 60 (ไม่ผ่าน)

Summary

- Long case exam
 - Objectives
 - Advantages and limitations
 - Objective Structured Long Case Examination Record (OSLER)
 - Long case exam in Thailand

17 กรกฎาคม 2563

Portfolio

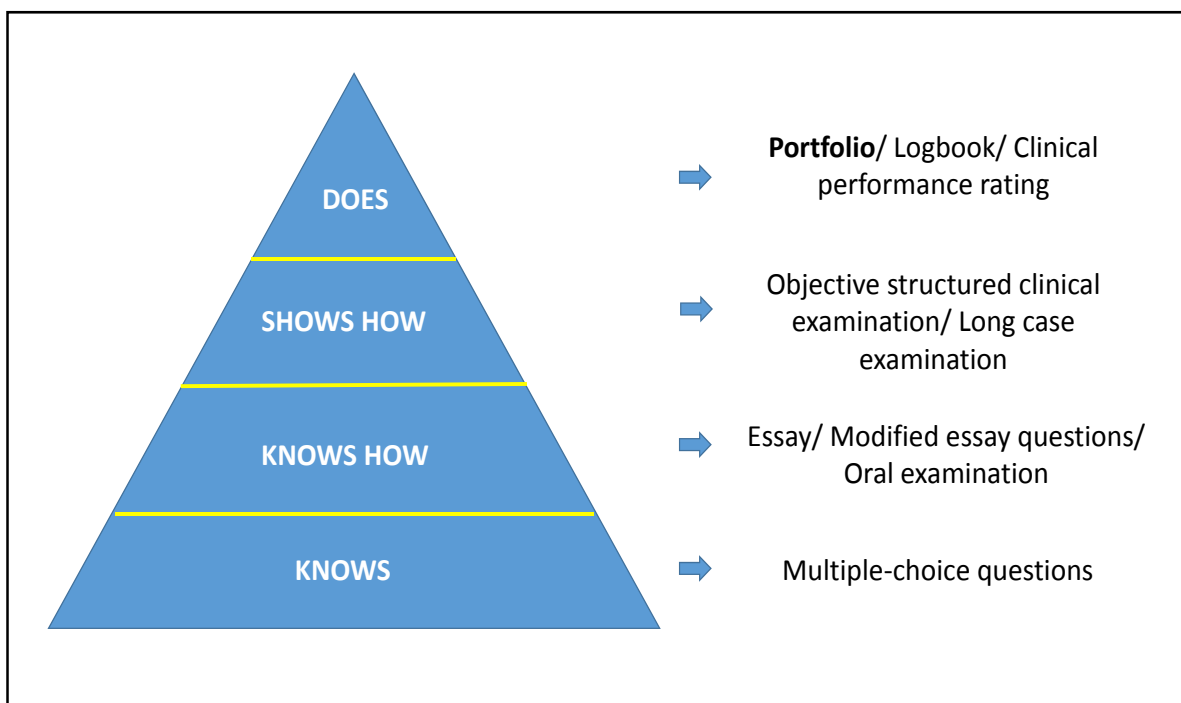
Portfolio

Assistant Professor Thos Harnroongroj
Department of Orthopaedics and Rehabilitation, Faculty of Medicine,
Siriraj Hospital, Mahidol University

Objectives

- Knows the characteristics of portfolio
- Knows how to develop the portfolio

What is portfolio?



What is portfolio ?

- One type of assessment
- “Does” level
- Wide ranges of assessment
- Linkage between assessment and learning
 - Reflection
 - Feedback

Haldane T. Gastroenterol Hepatol Bed Bench. 2014

Benefits and disadvantages of portfolio

Benefits	Disadvantages
<ul style="list-style-type: none">• Dynamic assessment (Longitudinal)• “Does” level assessment• Includes knowledge, skill and attitude• Reflective observation• Feedback and improvement	<ul style="list-style-type: none">• Validity• Reliability• Co-operation

Heeneman S. GMS Journal for Medical Education. 2017
Haldane T. Gastroenterol Hepatol Bed Bench. 2014

Developing the portfolio

Steps of developing portfolio

- Defining learning objective(s)
- Defining the proofed evidence(s)
- Observation-reflection part
- Evaluation and feedback part

Defining learning objective(s)

Learning outcomes	Why important ?

Learning objective(s)

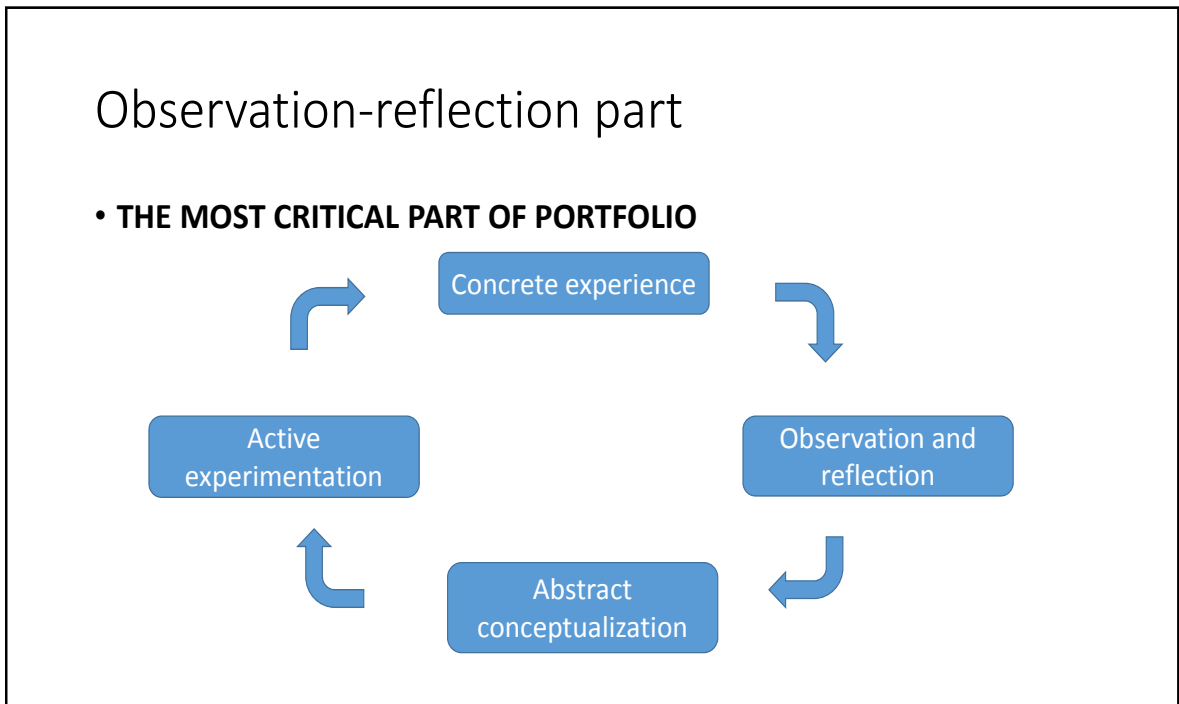
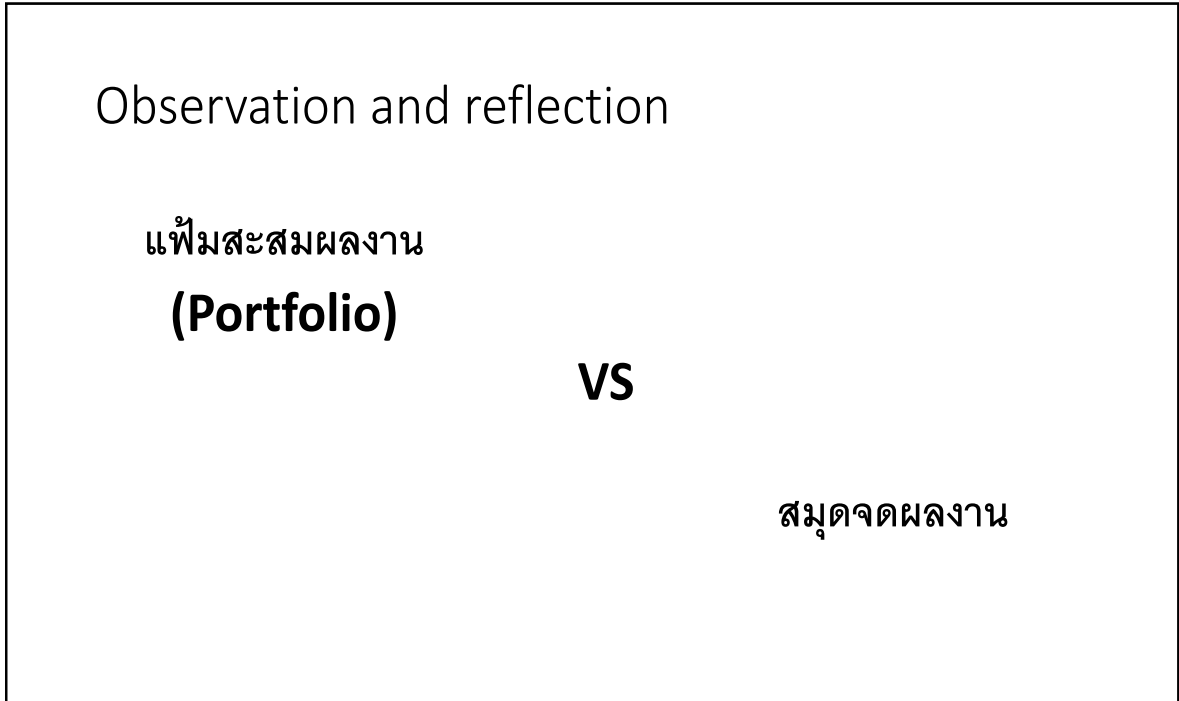
- Knowledge
- Skill
- Attitude

Defining the proof evidence(s)

Learning outcomes	Evidences	How

Characteristics of proof evidence(s)

Characteristics	Descriptions
Whole experiences	+ See whole pictures - Assess more difficult
Shopping cart	The learner choose the needed evidence.
Platinum level	The learner choose the best proof evidence. + Take less time for assessment - Validity



Levels of reflection

- Descriptive reflection (Superficial reflection)
- Practical reflection (Middle level reflection)
- **Critical reflection (Deep reflection)**

Effective observation and reflection

- Proper and adequate time
- Safe and supportive environment

Evaluation and feedback

- Summative
- Formative

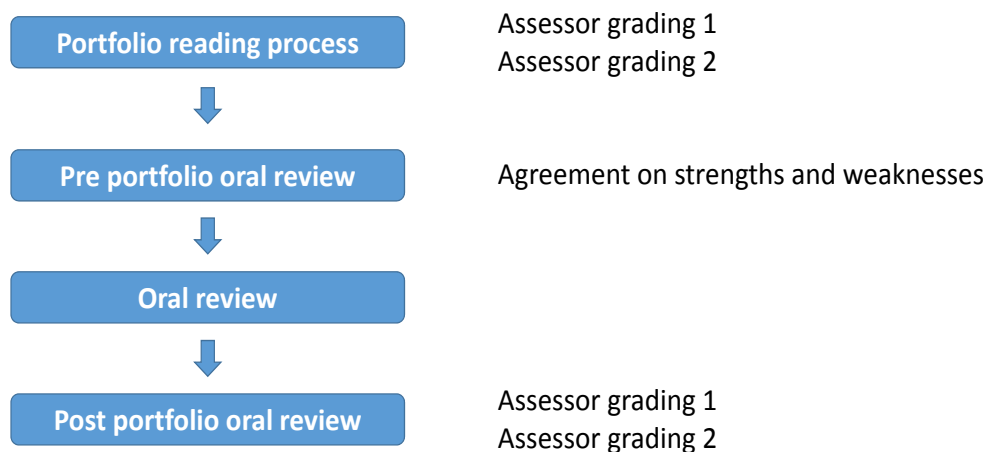
Formative evaluation

Advantages	Disadvantages
<ul style="list-style-type: none">• Less stress• Motivational support• Feedback• Enhances experiential learning	<ul style="list-style-type: none">• Less cooperation

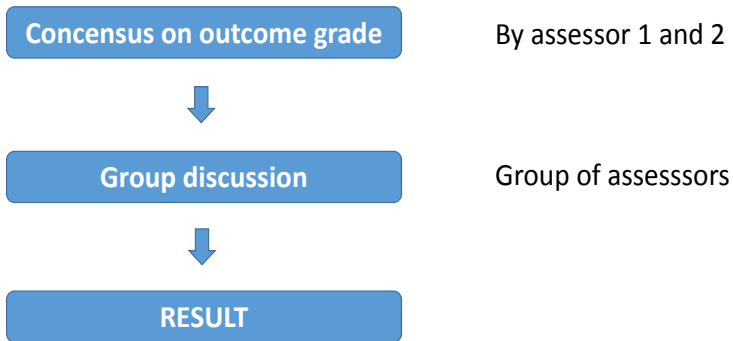
Summative evaluation

Advantages	Disadvantages
<ul style="list-style-type: none"> • Validity • Acceptable reliability • Practical 	<ul style="list-style-type: none"> • More stress • Less stimulation in experiential learning process

Summative evaluation process



Summative evaluation process



Successful portfolio

- Organization support
- Medical students
- **MEDICAL TEACHERS**

Conclusions

- Characteristics of portfolio
- How to develop a successful portfolio

17 กรกฎาคม 2563

Rating scale development

Rating Scale Development

เชิดศักดิ์ ไอรรมณีรัตน์
ภาควิชาศัลยศาสตร์ คณะแพทยศาสตร์ศิริราชพยาบาล
มหาวิทยาลัย มหิดล

Rating Scale

- Subjective rating of clinical skills and attitudes usually require rating scale or checklist
 - Rating scale: > 2 levels of score
 - Checklist: Yes/No

2

Rater Errors

- Construct-irrelevance variance in performance ratings that is associated with raters' behavior, not with the actual performance of ratees
- Valid use of 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.

3

Rater Errors

- Leniency/Severity**
 - difference in the levels of severity between raters
- Rater inconsistency**
 - instability of the level of severity within each rater
- Halo**
 - rater's tendency to let the rating of one trait influence his/her ratings on other traits
- Restriction of range**
 - clustering of ratings around a particular point on the rating scale

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.

4

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

Activity

- Open a web browser
- Go to <http://socrative.com>
- Select [Student login]
- In Room name, type in: IRAMANEERAT
- Click [Join]
- Type in your own name

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

แบบวัดผลสัมฤทธิ์ทางการเรียนของนักศึกษาพยาบาลปี 6
คณะพยาบาลศาสตร์ศิริราชพยาบาล
รศ. ชัยพร วัฒนศิริกุล
รศ.ดร.สุวิมล วัฒนศิริกุล

ข้อ	ข้อ (1)	ข้อ (2)	ข้อ (3)	ข้อ (4)
1. ความรู้	ความรู้เกี่ยวกับพยาธิวิทยา	ความรู้เกี่ยวกับพยาธิวิทยา	ความรู้เกี่ยวกับพยาธิวิทยา	ความรู้เกี่ยวกับพยาธิวิทยา
2. ทักษะ	ทักษะการวิเคราะห์ปัญหา	ทักษะการวิเคราะห์ปัญหา	ทักษะการวิเคราะห์ปัญหา	ทักษะการวิเคราะห์ปัญหา
3. เจตคติ	เจตคติในการเรียน	เจตคติในการเรียน	เจตคติในการเรียน	เจตคติในการเรียน
4. การสื่อสาร	การสื่อสารในชั้นเรียน	การสื่อสารในชั้นเรียน	การสื่อสารในชั้นเรียน	การสื่อสารในชั้นเรียน
5. การทำงานเป็นทีม	การทำงานเป็นทีม	การทำงานเป็นทีม	การทำงานเป็นทีม	การทำงานเป็นทีม
6. การแก้ปัญหา	การแก้ปัญหา	การแก้ปัญหา	การแก้ปัญหา	การแก้ปัญหา
7. การปรับตัว	การปรับตัว	การปรับตัว	การปรับตัว	การปรับตัว
8. การประเมินผล	การประเมินผล	การประเมินผล	การประเมินผล	การประเมินผล
9. การเรียนรู้	การเรียนรู้	การเรียนรู้	การเรียนรู้	การเรียนรู้
10. การสื่อสาร	การสื่อสาร	การสื่อสาร	การสื่อสาร	การสื่อสาร
11. การทำงานเป็นทีม	การทำงานเป็นทีม	การทำงานเป็นทีม	การทำงานเป็นทีม	การทำงานเป็นทีม
12. การแก้ปัญหา	การแก้ปัญหา	การแก้ปัญหา	การแก้ปัญหา	การแก้ปัญหา
13. การปรับตัว	การปรับตัว	การปรับตัว	การปรับตัว	การปรับตัว
14. การประเมินผล	การประเมินผล	การประเมินผล	การประเมินผล	การประเมินผล
15. การเรียนรู้	การเรียนรู้	การเรียนรู้	การเรียนรู้	การเรียนรู้

Questions & Comments

Cherdsakramaneerat@gmail.com

Group Work

- ให้อาจารย์ออกแบบใบประเมิน **performance** ในบริบทใดก็ได้ที่อาจารย์มีส่วนเกี่ยวข้อง
- ส่งตัวแทนนำเสนอ
 - บริบท: นักศึกษา, ชี้นี้, วิชา, สิ่งที่ต้องการประเมิน
 - ใบประเมิน(เวลาออกแบบ 10 นาที)
(เวลานำเสนอ กลุ่มละ 3 นาที)

13

17 กรกฎาคม 2563

Workplace-based assessment

Workplace-based
Assessment

เชิดศักดิ์ ไชยรัตน์
ภาควิชาศัลยศาสตร์ คณะแพทยศาสตร์ศิริราชพยาบาล
มหาวิทยาลัย มหิดล

1

Workplace-based Assessment

- A number of assessment methods, suitable for providing feedback based on observation of trainee performance in the workplace.
 - Mini-clinical Evaluation Exercise (mini-CEX)
 - Clinical Encounter Card (CEC)
 - Blinded Patient Encounter (BPE)
 - Direct Observation of Procedural Skills (DOPS)
 - Procedure based assessment (PBA)
 - Case-based Discussion (CbD)
 - Multisource Feedback (MSF)

WPBA: Characteristics

- เป็นการประเมินที่ให้ผู้เรียนเป็นผู้เริ่มต้น
- ผู้เรียนสามารถขอให้อาจารย์ประเมินได้ตลอดเวลาปฏิบัติงาน
- เป็นการประเมินในสถานที่ปฏิบัติงานจริง
- ประเมินซ้ำได้ หากคิดคะแนนจะใช้คะแนนครั้งที่ดีที่สุด
- จุดมุ่งหมายสำคัญคือการเปิดโอกาสให้อาจารย์ได้ **feedback**

WPBA: Strengths

- **Validity:** assessment of “does” level
- Identify students in needs of support early
- Provide feedback
- Create a nurturing culture
- Samples widely in many workplaces
- Utilize a number of assessors

General Medical Council. Workplace based assessment: A guide for implementation, April 2010.

WPBA: Limitations

- Low reliability
- Can be opportunistic
- Trainees may delay or avoid assessment
- Learner dependent and vulnerable
- Require time and training
- Bias due to the interaction between trainers and trainees

General Medical Council. Workplace based assessment: A guide for implementation, April 2010.

Mini-Clinical Evaluation Exercise

- นักศึกษาแสดงการ **approach** ผู้ป่วยจริงในคลินิกหรือหอผู้ป่วย ขณะที่ได้รับการสังเกตการณ์โดยอาจารย์
 - Focused history taking
 - Focused physical examination
 - Making clinical diagnosis
 - Develop a management plan
- ใช้เวลาในการ **approach** ผู้ป่วย 15 นาทีต่อราย ตามด้วยการให้ **feedback** จากอาจารย์อีก 5 นาที
- อาจารย์ให้คะแนนแต่ละทักษะด้วย **rating scale 1-9**

Mini-CEX

Please refer to www.rcp.ac.uk for guidance on this form and details of expected competences for F1

Mini-Clinical Evaluation Exercise (CEX) - F1 Version

Please complete the questions using a pencil. Please use black ink and CAPITAL LETTERS

Doctor's Surname: _____ Forename: _____

GMC Number: _____ **GMC NUMBER MUST BE COMPLETED**

Clinical setting: Ambulatory Outpatient Inpatient Acute Admission GP Surgery

Clinical problem category: Anaemia/ CVU Pain Psych/ Mental Neuro Gastro Other

Focus of clinical history: History Examination Management Investigation

Complexity of case: Low Average High

Number of times patient seen before by assessor: 1-4 5-9 10 11-20 21-30 31-40 41-50 51-60 61-70 71-80 81-90 91-100 Other

Assessor's Consultant GP Specialist Other

Number of previous mini-CEX observed for assessor with this patient: _____

Please grade the following areas using the scale below:

	Below expectations for this competence	Meets expectations for this competence	Exceeds expectations for this competence	Assess expectations for this competence	UAC*
1. History Taking	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
2. Physical Examination Skills	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
3. Communication Skills	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
4. Clinical Judgement	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
5. Professionalism	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
6. Organisation/Efficiency	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
7. Overall clinical case	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

*UAC: Please mark this if you have not observed the behaviour and therefore feel unable to comment.

Direct Observation of Procedural Skills (DOPS)

- ประเมินทักษะการทำหัตถการในขณะทำงานกับผู้ป่วยจริง
- อาจารย์สังเกตขั้นตอนการทำหัตถการแล้วให้คะแนนด้วย rating scale 1-6 คะแนนในแต่ละขั้นตอนที่ต้องการประเมิน
- แต่ละหัตถการทำการประเมินโดยอาจารย์หลายท่าน ในหลายบริบท
- แต่ละหัตถการใช้เวลาสังเกต 15 นาที และ feedback 5 นาที
- ตัวอย่างหัตถการ: endotracheal intubation, nasogastric tube insertion, IV injection, arterial blood sampling, etc.

DOPS

Please refer to www.rcp.ac.uk for guidance on this form and details of expected competences for F1

Direct Observation of Procedural Skills (DOPS) - F1 Version

Please complete the questions using a pencil. Please use black ink and CAPITAL LETTERS

Doctor's Surname: _____ Forename: _____

GMC Number: _____ **GMC NUMBER MUST BE COMPLETED**

Clinical setting: Ambulatory Outpatient Inpatient Acute Admission GP Surgery

Procedure Number: _____

Assessor's Consultant GP Specialist Other

Number of previous DOPS observed for assessor with this patient: _____

Number of times procedure performed (or observed): 1-4 5-9 10 11-20 21-30 31-40 41-50 51-60 61-70 71-80 81-90 91-100 Other

Please grade the following areas using the scale below:

	Below expectations for this competence	Meets expectations for this competence	Exceeds expectations for this competence	Assess expectations for this competence	UAC*
1. Demonstrate understanding of anatomy, medical history, history of procedure	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
2. Identify critical components	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
3. Demonstrate appropriate preparation arrangements	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
4. Adapt to changes or safe selection	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
5. Technical skills	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
6. Aseptic technique	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
7. Skills used where appropriate	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
8. Professionalism	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
9. Organisation of patient preparation	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
10. Overall ability to perform procedure	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

*UAC: Please mark this if you have not observed the behaviour and therefore feel unable to comment.

Case-based Discussion (CbD)

- นักศึกษาเลือกผู้ป่วย 2 รายที่ตนเคยดูแลนำเสนอให้อาจารย์
- อาจารย์ผู้ประเมินเลือก 1 ใน 2 ผู้ป่วยนั้นเพื่อทำการอภิปรายรายละเอียดของผู้ป่วย
 - Clinical assessment
 - Investigations
 - Treatment
 - Follow-up and future plan
- วัตถุประสงค์เพื่อประเมิน Clinical reasoning skills
- การอภิปรายผู้ป่วยแต่ละรายใช้เวลาไม่เกิน 20 นาที และ feedback 5 นาที

CbD

Please refer to www.rcp.ac.uk for guidance on this form and details of expected competences for F1 and F2

Case-based Discussion (CbD) - F2 Version

Please complete the questions using a pencil. Please use black ink and CAPITAL LETTERS

Doctor's Surname: _____ Forename: _____

GMC Number: _____ **GMC NUMBER MUST BE COMPLETED**

Clinical setting: Ambulatory Outpatient Inpatient Acute Admission GP Surgery

Clinical problem category: Pain Anaemia/ CVU Psych/ Mental Neuro Gastro Other

Focus of clinical history: Medical records keeping Clinical assessment Management Investigation

Complexity of case: Low Average High

Assessor's Consultant GP Specialist Other

Please grade the following areas using the scale below:

	Below expectations for this competence	Meets expectations for this competence	Exceeds expectations for this competence	Assess expectations for this competence	UAC*
1. Medical record keeping	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
2. Clinical assessment	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
3. Investigative and referrals	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
4. Treatment	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
5. Follow-up and future planning	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
6. Professionalism	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
7. Overall clinical judgement	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

*UAC: Please mark this if you have not observed the behaviour and therefore feel unable to comment.

Procedure-Based Assessment (PBA)

- A form of workplace-based assessment
- An assessor completes the form based on observation of a trainee performs a surgical procedure
- Six domains: consent, pre-operative planning, exposure and closure, intraoperative technique, postoperative management
- Two groups of items: general items, task-specific items
- Binary rating: satisfactory, unsatisfactory

Marriott J et al. Evaluation of procedure-based assessment for assessing trainees' skills in the operating theatre. *BJO 2011; 98: 450-7.*

WPBA Guidelines

- The purpose of WPBA must be clear to both trainers and trainees
 - Formative
 - Summative
- Transparent mapping of WPBA to the curriculum is essential

General Medical Council. Workplace based assessment: A guide for implementation, April 2010.

WPBA Guidelines (2)

- Setting up the WPBA
 - Environment: constructive environment, low ratings are acceptable
 - A framework to support trainees in planning WPBA
 - Multiple assessments by a range of assessors
- Roles of assessors
 - Training
 - Provide written records of feedback

WPBA Guidelines (3)

- Roles and responsibilities of trainees
 - Monitor their own progress
 - Pay attention to feedback
- Quality management
 - Constant monitoring of the implementation of WPBA

Summary

- Workplace-based assessment
 - Strengths and limitations
 - Examples of WPBA
 - WPBA Guidelines

*"I have failed many times,
and that's why I am a success."*

Michael Jordan

การประเมินผลในบริบทของการทำงาน (Workplace-based assessment)

รศ.นพ.เชิดศักดิ์ ไชยมณีรัตน์

ภาควิชาศัลยศาสตร์

การประเมินผลมีบทบาทในการส่งเสริมการเรียนรู้ของนักศึกษาได้จากหลายกลไก ซึ่งกลไกหนึ่งที่มีความสำคัญมากและอาจารย์แพทย์ควรมีการใช้มากขึ้นคือการใช้การประเมินผลในระหว่างเรียน (formative assessment) เพื่อให้ได้ข้อมูลว่านักศึกษามีระดับความรู้ ความสามารถมากน้อยเพียงใด ยังต้องพัฒนาในด้านใดบ้าง แล้วนำข้อมูลดังกล่าวให้แก่นักศึกษา (feedback) เพื่อให้ นักศึกษาได้พัฒนาตนเองให้ดีขึ้นก่อนที่จะถูกประเมินผลในตอนสิ้นสุดการเรียน (summative assessment) งานวิจัยหลายชิ้นแสดงให้เห็นว่านักศึกษาแพทย์และแพทย์ประจำบ้านได้รับการสังเกตและประเมินผลการทำงานในลักษณะของ formative assessment นี้ไม่เพียงพอ ซึ่งส่งผลให้นักศึกษาและแพทย์ประจำบ้านเหล่านี้ขาดโอกาสที่จะพัฒนาความรู้ และทักษะพื้นฐานในการดูแลผู้ป่วย อย่างมีประสิทธิภาพ

การประเมินผลในบริบทของการทำงาน (Workplace-based assessment) เป็นกลุ่มของวิธีการประเมินผลที่ถูกพัฒนาขึ้น เพื่อให้อาจารย์แพทย์ได้มีโอกาสประเมินความรู้ และทักษะต่างๆทางคลินิกของนักศึกษาในขณะที่ทำงานกับผู้ป่วยจริง และได้นำผลการประเมินนั้นมาชี้แนะแนวทางในการพัฒนาความรู้ และทักษะของนักศึกษา วิธีการประเมินผลในกลุ่มนี้มีลักษณะสำคัญต่างๆ ดังนี้

1. เป็นการประเมินผลที่ให้นักศึกษาเป็นผู้เริ่มต้น กล่าวคือ นักศึกษาไม่ต้องรอถึงวันที่กำหนดโดยอาจารย์ว่าจะทำการสอบในวันใด เมื่อไรที่นักศึกษาพบโอกาสเหมาะได้พบผู้ป่วยที่ตนสามารถแสดงระดับความรู้ และทักษะทางคลินิกของตนให้อาจารย์แพทย์ประเมินได้ก็ทำการขอให้อาจารย์ช่วยประเมินผลได้ทันที
2. นักศึกษาสามารถขอให้อาจารย์ประเมินได้ตลอดช่วงเวลาที่ทำการปฏิบัติงานในภาควิชา ไม่จำเป็นต้องรอถึงสิ้นสุดการปฏิบัติงาน การเปิดโอกาสให้ประเมินได้ตั้งแต่เริ่มปฏิบัติงานทำให้อาจารย์สามารถเห็นจุดบกพร่องของนักศึกษาแต่แรกในขณะที่ยังมีเวลาให้นักศึกษาได้ฝึกฝนพัฒนาตนเอง
3. นักศึกษาสามารถขอรับการประเมินทักษะเดิมซ้ำได้จนกว่านักศึกษาจะมีทักษะดังกล่าวดีเป็นที่น่าพอใจ โดยคะแนนที่จะนำไปตัดสินผลการศึกษาคือคะแนนครั้งที่นักศึกษาทำได้ดีที่สุด
4. จุดมุ่งหมายสำคัญของการประเมินผลคือการเปิดโอกาสให้อาจารย์ได้สังเกตนักศึกษาปฏิบัติงานกับผู้ป่วยจริงแล้วให้ข้อมูลย้อนกลับ (feedback) แก่นักศึกษา
5. นักศึกษาเป็นผู้รับผิดชอบในการดำเนินการรวบรวมคะแนนของตนเอง และตรวจสอบว่าตนยังต้องทำการประเมินทักษะใดอีกบ้าง

วิธีการประเมินผลกลุ่มนี้มีด้วยกันหลายวิธี เช่น

1. **Mini-clinical evaluation exercise (Mini-CEX)** เป็นการประเมินทักษะการตรวจรักษาผู้ป่วยที่แผนกผู้ป่วยนอก หรือในหอผู้ป่วยโดยอาจารย์แพทย์ให้เวลานักศึกษาซักประวัติ ตรวจร่างกายผู้ป่วยราว 15 นาทีแล้วจึงทำการอธิบายและทำการให้คะแนนร่วมกับให้ feedback แก่นักศึกษา

2. Direct Observation of Procedural Skills (DOPS) เป็นการประเมินทักษะการทำหัตถการพื้นฐาน โดยอาจารย์สังเกตนักศึกษาทำหัตถการดังกล่าวกับผู้ป่วย ซึ่งมักเป็นหัตถการที่ใช้เวลาทำไม่นานนัก เสร็จแล้วอาจารย์ให้คะแนนและให้ feedback แก่นักศึกษา
3. Case-based discussion (CbD) เป็นการประเมินทักษะการคิดวิเคราะห์แก้ปัญหาทางคลินิกของนักศึกษา โดยนักศึกษาเลือกผู้ป่วยสองรายที่ตนเคยดูแลเสนอให้อาจารย์ผู้ประเมิน อาจารย์ทำการเลือกหนึ่งในสองรายของผู้ป่วยให้นักศึกษานำเสนอประวัติและแนวทางในการตรวจวินิจฉัย และการรักษา ร่วมกับการที่อาจารย์ซักถามเพื่อประเมินความเข้าใจในผู้ป่วยของนักศึกษา เมื่ออภิปรายผู้ป่วยดังกล่าวเสร็จแล้ว อาจารย์ให้คะแนนและให้ feedback แก่นักศึกษา
4. Multisource feedback (MSF) เป็นการเก็บรวบรวมใบประเมินการปฏิบัติงานที่กรอกโดยบุคลากรที่ทำงานร่วมกับนักศึกษาที่หลากหลาย ได้แก่ อาจารย์ พยาบาล แพทย์ประจำบ้าน เพื่อนนักศึกษา แล้วนำผลการประเมินที่รวบรวมได้มาสรุปแล้วนำไปให้ feedback แก่นักศึกษาให้เห็นข้อมูลในภาพรวมว่าการทำงานของเขาในสายตาของเพื่อนร่วมงานนั้นมีประสิทธิภาพเพียงใด มีสิ่งใดที่นักศึกษาควรปรับปรุงบ้าง

หากอาจารย์มีความต้องการจะนำผลการประเมินผลในบริบทของการทำงานมาใช้มีข้อเสนอแนะแนวปฏิบัติบางประการเพื่อช่วยให้การประเมินผลนี้บรรลุตามวัตถุประสงค์

1. ต้องทำการชี้แจงให้ทั้งนักศึกษาและอาจารย์ที่เกี่ยวข้องในการประเมินผลทุกท่านเข้าใจถึงวัตถุประสงค์ของการประเมินนี้ที่ชัดเจนว่ามุ่งเน้นที่ formative assessment
2. ต้องทำการสร้างบรรยากาศที่เอื้อให้เกิดการประเมินผลในรูปแบบนี้ โดยทำให้ทุกคนเข้าใจว่าเป็นการประเมินเพื่อให้เกิดการพัฒนาในตัวนักศึกษา ดังนั้นไม่จำเป็นต้องคาดหวังว่าจะต้องได้คะแนนสูงในการประเมินทุกครั้ง และต้องสร้างกระบวนการช่วยกระตุ้นให้นักศึกษาล้ำที่ จะเชิญชวนให้อาจารย์ทำการประเมินนักศึกษาเมื่อมีโอกาส และให้นักศึกษาสามารถเข้าหาอาจารย์ได้หลากหลายท่าน
3. ต้องมีการกำหนดบทบาทที่ชัดเจนของอาจารย์ว่าในการประเมินรูปแบบนี้หน้าที่ของอาจารย์ไม่เพียงแต่ให้คะแนนตามใบประเมินแล้ว อาจารย์ยังต้องมีหน้าที่สอน และให้ข้อมูลย้อนกลับ (feedback) แก่นักศึกษาด้วย
4. ต้องสร้างความเข้าใจในกลุ่มนักศึกษว่าตัวนักศึกษาเองมีความรับผิดชอบในการหาโอกาสที่จะรับการประเมินด้วยตนเอง สนใจรับฟัง feedback จากอาจารย์เพื่อนำไปปรับปรุงตนเองให้ดีขึ้น และตรวจสอบว่าตนเองจะต้องประเมินทักษะใดบ้าง
5. การดำเนินการให้ประสบความสำเร็จต้องอาศัยความช่วยเหลือจากเจ้าหน้าที่สายสนับสนุนทางการศึกษาในการรวบรวมข้อมูลการประเมินผล และตรวจสอบว่ามีปัญหาใดเกิดขึ้นในกระบวนการประเมินผลหรือไม่ เช่น จำนวนผู้ป่วยที่เหมาะสมสำหรับการประเมินนักศึกษามีเพียงพอหรือไม่ มีอาจารย์ท่านใดไม่มีส่วนร่วมในการประเมินนักศึกษาหรือไม่ มีอุปสรรคใดทำให้นักศึกษาไม่สามารถรับการประเมินได้หรือไม่ เป็นต้น

หากอาจารย์สามารถนำรูปแบบการประเมินผลในบริบทของการทำงานไปประยุกต์ใช้ในการพัฒนาการเรียนการสอนในภาควิชาต่างๆได้ น่าจะทำให้เพิ่ม constructive feedback ให้แก่นักศึกษาและก่อให้เกิดการพัฒนาความรู้ และทักษะพื้นฐานทางการแพทย์ในนักศึกษาแพทย์ได้อย่างมีประสิทธิภาพมากขึ้น

AMEE GUIDE

Workplace-based assessment as an educational tool: AMEE Guide No. 31

JOHN NORCINI¹ & VANESSA BURCH²¹Foundation for Advancement of International Medical Education and Research, Philadelphia, USA, ²University of Cape Town, South Africa

Abstract

Background: There has been concern that trainees are seldom observed, assessed, and given feedback during their workplace-based education. This has led to an increasing interest in a variety of formative assessment methods that require observation and offer the opportunity for feedback.

Aims: To review some of the literature on the efficacy and prevalence of formative feedback, describe the common formative assessment methods, characterize the nature of feedback, examine the effect of faculty development on its quality, and summarize the challenges still faced.

Results: The research literature on formative assessment and feedback suggests that it is a powerful means for changing the behaviour of trainees. Several methods for assessing it have been developed and there is preliminary evidence of their reliability and validity. A variety of factors enhance the efficacy of workplace-based assessment including the provision of feedback that is consistent with the needs of the learner and focused on important aspects of the performance. Faculty plays a critical role and successful implementation requires that they receive training.

Conclusions: There is a need for formative assessment which offers trainees the opportunity for feedback. Several good methods exist and feedback has been shown to have a major influence on learning. The critical role of faculty is highlighted, as is the need for strategies to enhance their participation and training.

Introduction

For just over two decades leading educationists, including medical educators, have highlighted the intimate relationship between learning and assessment. Indeed, in an educational context it is now argued that learning is the key purpose of assessment (van der Vleuten 1996; Gronlund 1998, Shepard 2000). At the same time as this important connection was being stressed in the education literature; there were increasing concerns about the workplace-based training of doctors. A study by Day et al. (1990) in the United States documented that the vast majority of first-year trainees in internal medicine were not observed more than once by a faculty member in a patient encounter where they were taking a history or doing a physical examination. Without this observation, there was no opportunity for the assessment of basic clinical skills and, more importantly, the provision of feedback to improve performance.

As one step in encouraging the observation of performance by faculty, the American Board of Internal Medicine proposed the use of the mini-Clinical Evaluation Exercise (mini-CEX) (Norcini et al. 1995). In the mini-CEX, a faculty member observes a trainee as he/she interacts with a patient around a focused clinical task. Afterwards, the faculty member assesses the performance and provides the trainee feedback. It was expected that trainees would be assessed several times throughout the year of training with different faculty and in different clinical situations.

Practice points

- The research literature on work-based formative assessment and feedback suggests that it is a powerful means for changing the behaviour of learners.
- Several formative assessment methods have been developed for use in the workplace and there is preliminary data evidence of their reliability and validity.
- The efficacy of feedback is enhanced if it is consistent with the needs of the learner, focuses on important aspects of the performance in the work-place, and has characteristics such as being timely and specific.
- Faculty development is critical to the quality and effectiveness of formative assessment.
- Strategies to encourage the participation of faculty are critical to the successful implementation of formative assessment.

An advantage of the mini-CEX and other workplace-based methods is that they fulfil the three basic requirements for assessment techniques that facilitate learning (Frederiksen 1984; Crooks 1988; Swanson et al. 1995; Shepard 2000):(1) The content of the training programme, the competencies expected as outcomes, and the assessment practices are aligned (2) Trainee feedback is provided during and/or after assessment

Correspondence: John Norcini, Foundation for Advancement of International Medical Education and Research (FAIMER) 4th Floor 3624 Market St, Philadelphia PA 19104, USA. Tel: 1 215 823 2170; fax: 1 215 386 2321; email: JNorcini@faimer.org

events;(3) Assessment events are used strategically to steer trainee learning towards the desired outcomes. Over the past several years there has been growing interest in workplace-based assessment and additional methods have been (re)introduced to the setting of clinical training (National Health Service 2007).

Previous publications have focused on the advantages and disadvantages of workplace-based methods from the perspective of assessment alone (Norcini 2007). In this role, the methods are best thought of as analogous to classroom tests and they have much strength from this perspective. However, it is difficult to assure equivalence across institutions and the observations of faculty may be influenced by the stakes and their relationships with trainees. Consequently, their use faces challenges as national high stakes assessment devices.

Perhaps more importantly, workplace-based assessment can be instrumental in the provision of feedback to trainees to improve their performance and steer their learning towards desired outcomes. This paper focuses on the use of the methods for this purpose and it is divided into five sections. The first section briefly reviews the literature on the efficacy and prevalence of formative assessment and feedback. This is followed by a section that describes some of the more common methods of work-based assessment. The third section concentrates on feedback and it is explored from the perspective of the learner, its focus, and which characteristics make it effective in the context of formative assessment. Faculty play a key role in the successful implementation of formative assessment, so the fourth section describes strategies to encourage their participation and training to improve their performance. In the closing section we draw attention to the challenges faced by medical educators implementing formative assessment strategies in routine clinical teaching practice.

Efficacy and prevalence of formative assessment and feedback

The purpose of formative assessment and feedback

Formative assessment is not merely intended to assign grades to trainee performance at designated points in the curriculum; rather it is designed to be an ongoing part of the instructional process and to support and enhance learning (Shepard 2000). Clearly, feedback is a core component of formative assessment (Sadler 1989), central to learning, and at '*the heart of medical education*' (Branch & Paranjape 2002). In fact, it is useful to consider feedback as part of an ongoing programme of assessment and instruction rather than a separate educational entity (Hattie & Timperley 2007).

Feedback promotes student learning in three ways (Gipps 1999, Shepard 2000):

- it informs trainees of their progress or lack thereof;
- it advises trainees regarding observed learning needs and resources available to facilitate their learning; and
- it motivates trainees to engage in appropriate learning activities.

Efficacy of feedback

Given these presumed benefits, it is appropriate to ask whether there is a body of research supporting the efficacy of feedback in changing trainees' behaviour. Most compelling is a synthesis of information on classroom education by Hattie which included over 500 meta-analyses involving 1,800 studies and approximately 25 million students (Hattie 1999). He demonstrated that the typical effect size (ES) of schooling on overall student achievement is about 0.40 (i.e. it increases the mean on an achievement test by 0.4 of a standard deviation). Using this as a benchmark or 'gold standard' on which to judge the various factors that affect performance, Hattie summarized the results of 12 meta-analyses that specifically included the influence of feedback. The feedback effect size was 0.79, which is certainly very powerful, and among the four biggest influences on achievement. Hattie also found considerable variability based on the type of feedback, with the largest effect being generated by the provision of information around a specific task.

Data to answer the question about the efficacy of feedback are much more limited in the domain of medical education but a recent meta-analysis by Veloski and colleagues looked at its effect on clinical performance (Veloski et al. 2006). Of the 41 studies meeting the criteria for inclusion, 74% demonstrated a positive effect for feedback alone. When combined with other educational interventions, feedback had a positive effect in 106 of the 132 (77%) studies reviewed.

A recent paper by Burch and colleagues reports on the impact of a formative assessment strategy implemented in a 4th year undergraduate medical clerkship programme (Burch et al. 2006). In this paper, students who engaged in an average of 6 directly observed clinical encounters during a 14-week clerkship reported that they more frequently undertook blinded patient encounters (McLeod & Meagher 2001) in which they did not consult the patient records before interviewing and examining the patient. Prior to implementing the formative assessment programme, students traditionally interviewed and examined patients only after consulting patient records. In addition they reported that they read more frequently on topics only relevant to patients clerked in the ward. While this paper provides information on self-reported learning behaviour changes, it does suggest that formative assessment may have the potential to strategically direct student learning by reinforcing desirable learning behaviour (Gibbs 1999).

A recent publication by Driessen and van der Vleuten (2000) support the findings reported by Burch. In their study they introduced a portfolio of learning assignments as an educational tool in a legal skills training programme comprising tutorials which were poorly attended and for which students did not adequately complete the required pre-tutorial work. The portfolio assignments, such as writing a legal contract or drafting a legislative document, were reviewed by peers and the tutor prior to being used as the teaching basis for subsequent skills training sessions. This educational intervention resulted in a twofold increase in time spent preparing for skills training sessions.

Prevalence of feedback

It is clear from these data that formative assessment and feedback have a powerful influence on trainee performance. However, there is a significant gap between what should be done and 'on the ground' practice. Lack of assessment and feedback, based on observation of performance in the workplace, is one of the most serious deficiencies in current medical education practice (Holmboe et al. 2004; Kassebaum & Eaglen 1999). Indeed, direct observation of trainee performance appears to be the exception rather than the rule.

In a survey of 97 United States medical schools, accredited between 1993 and 1998, it was found that structured, observed assessments of students' clinical abilities were done across clinical clerkships for only 7.4% to 23.1% of medical students (Kassebaum and Eaglen 1999). A more recent survey of medical graduates found that during any given core clerkship, 17% to 39% of student were not observed performing a clinical examination (Association of American Medical Colleges 2004). Likewise, Kogan & Hauer (2006) found that only 28% of Internal Medicine clerkships included an in-course formative assessment strategy involving observation of student performance in the workplace setting. Outside the US, Daelmans et al. (2004) reported that over a 6-month period, observation of trainee performance occurred in less than 35% of educational events in which observation and the provision of feedback could have taken place.

Unfortunately the situation is no better in postgraduate training programmes. In one study, 82% of residents reported that they engaged in only one directly observed clinical encounter in their first year of training; far fewer (32%) engaged in more than one encounter (Day et al. 1990). In another survey of postgraduate trainees 80% reported never or only infrequently receiving feedback based on directly observed performance (Isaacson et al. 1995).

Not only is assessment of directly observed performance infrequently done as part of routine educational practice, but the quality of feedback, when given, may be poor. Holmboe colleagues evaluated the type of feedback given to residents after mini-CEX encounters and observed that while 61% of feedback sessions included a response from the trainee to the feedback, only 34% elicited any form of self-evaluation by the trainee. Of greatest concern, however, was the finding that only 8% of mini-CEX encounters translated into a plan of action (Holmboe et al. 2004a). The paper by Holmboe and colleagues suggests that there are key reasons why clinician-educators fail to give trainees effective feedback (see Box1):

In addition to finding that trainee observation and feedback is infrequently given and often of limited value, it has also been noted that the faculties' assessment of trainee performance may be less than completely accurate. Noel and colleagues found that faculty failed to detect 68% of errors committed by postgraduate trainees when observing a videotape scripted to depict marginal competence (Noel et al. 1992). The use of checklists prompting faculty to look for specific skills increased error detection from 32% to 64%. It was, however, noted that this did not improve the accuracy of assessors. Approximately two thirds of faculty still scored the overall performance of marginal postgraduate trainees as

Box 1. Key reasons why clinician-educators fail to give trainees effective feedback.

- Current in-vivo assessment strategies such as the mini-CEX may be focusing on assessment of performance at the expense of providing adequate feedback.
- The scoring sheets currently used for in-vivo assessment events provide only limited space for recording comments thereby limiting feedback given.
- Clinician-educators do not fully appreciate the role of feedback as a fundamental clinical teaching tool.
- Clinician-educators may not be skilled in the process of providing high quality feedback.

satisfactory or superior. Similar observations attesting to the poor accuracy of faculty observations have been made elsewhere (Herbers et al. 1989; Kalet et al. 1992).

Based on the infrequency with which trainees are observed and problems with the quality of the feedback they receive, it is fair to ask whether observation of trainee performance is an outdated approach to medical training and assessment. The critical question, therefore, is whether clinical interviewing and examination skills are still relevant to clinical practice such that faculty should be trained to properly observe performance and provide effective, useful feedback.

Feedback in relation to history and physical examination

Despite major technological advances, the ability to competently interview and examine patients remains one of the mainstays of clinical practice (Holmboe et al. 2004). Data gathered over the past 30 years highlight the critical importance of these skills. In 1975 Hampton and colleagues demonstrated that a good medical history produced the final clinical diagnosis in 82% of 80 patients interviewed and examined. In only one of 80 cases did laboratory tests provide the final diagnosis not made by history or physical examination (Hampton et al. 1975).

Technological advances over the past two decades have not made the findings of this study irrelevant. In 1992 Peterson and colleagues showed that among 80 patients presenting for the first time to a primary care clinic, the patient's history provided the correct final diagnosis in 76% of cases (Peterson et al. 1992). Even more recently, an autopsy study of 400 cases showed that the combination of a history and physical examination produced the correct diagnosis in 70% of cases. Diagnostic imaging studies successfully indicated the correct diagnosis in only 35% of cases (Kirch & Schafii 1996).

Beyond diagnostic accuracy, physician-patient communication is a key component of health care. In a review of the literature, Beck et al. (2002) found that both verbal behaviours (e.g., empathy, reassurance and support) and nonverbal behaviours (e.g., nodding, forward lean) were positively associated with patient outcomes. Likewise, a study by Little et al. (2001) found that the patients of doctors who took a patient-centred approach were more satisfied, more enabled, had greater symptom relief, and had lower rates of referral.

The ability to competently interview a patient and perform a physical examination thus remains the cornerstone

of clinical practice. The ability of faculty to accurately observe trainees performing these tasks and provide effective feedback is therefore one of the most important aspects of medical training. Although methods such as standardised patients certainly provide complementary assessment and feedback information, they cannot replace the central role of observation by faculty.

Formative assessment methods

A number of assessment methods, suitable for providing feedback based on observation of trainee performance in the workplace, have been developed or regained prominence over the past decade. This section provides a brief description of the essential features of some of them including:

- Mini-Clinical Evaluation Exercise (mini-CEX);
- Clinical Encounter Cards (CEC);
- Clinical Work Sampling (CWS);
- Blinded Patient Encounters (BPE);
- Direct Observation of Procedural Skills (DOPS);
- Case-based Discussion (CbD);
- MultiSource Feedback (MSF).

Mini-clinical evaluation exercise (mini-CEX)

As described above, the mini-CEX (Figure 1, Source: www.hcat.nhs.uk) is an assessment method developed in the United States (US) that is now in use in a number of institutions around the world. It requires trainees to engage in authentic workplace-based patient encounters while being observed by faculty members (Norcini et al. 1995). Trainees perform clinical tasks, such as taking a focused history or performing relevant aspects of the physical examination, after which they provide a summary of the patient encounter along with next steps (e.g., a clinical diagnosis and a management plan).

These encounters can take place in a variety of workplace settings including inpatient, outpatient, and emergency departments. Patients presenting for the first time as well as those returning for follow up visits are suitable encounters for the mini-CEX. Not surprisingly, the method lends itself to a wide range of clinical problems including: (1) presenting complaints such as chest pain, shortness of breath, abdominal pain, cough, dizziness, low back pain; or (2) clinical problems such as arthritis, chronic obstructive airways disease, angina, hypertension and diabetes mellitus (Norcini et al. 2003).

In the original work, each aspect of the clinical encounter is scored by a faculty member using a 9-point rating scale where 1–3 is unsatisfactory, 4–6 is satisfactory and 7–9 is superior. The parameters evaluated include: interviewing skill, physical examination, professionalism, clinical judgement, counselling, organization and efficiency, and overall competence. Different scales and different parameters have been used successfully in other settings (e.g., National Health Service).

The core purpose of the assessment method is to provide structured feedback based on observed performance. Each patient encounter takes roughly 15 minutes followed by 5–10 minutes of feedback. Trainees are expected to be evaluated

several times with different patients and by different faculty members during their training period.

This assessment tool has been shown to be a reliable way of assessing postgraduate trainee performance provided there is sufficient sampling. Roughly 4 encounters are sufficient to achieve a 95% confidence interval of less than 1 (on the 9-point scale) and approximately 12–14 are required for a reliability coefficient of 0.8 (Norcini et al. 1995, 2003; Holmboe et al. 2003).

In addition to the postgraduate setting, the mini-CEX has been successfully implemented in undergraduate medical training programmes (Hauer 2000; Kogan et al. 2003; Kogan & Hauer 2006). In this context, the period of observation and feedback is often longer, ranging from 30–45 minutes (Hauer 2000; Kogan et al. 2002).

There is a growing body of evidence supporting the validity of the mini-CEX. Kogan et al. (2002, 2003) found that mini-CEX performance was correlated with other assessments collected as part of undergraduate training. Faculty ratings of videotapes of student-standardized patient encounters, using the mini-CEX forms, were correlated with the checklist scores and standardized patient ratings of communication skills (Boulet et al. 2002). In postgraduate training, mini-CEX performance was correlated with a written in-training examination and routine faculty ratings (Durning et al. 2002). Holmboe et al. (2004) found that, using the mini-CEX form, they could differentiate amongst videos, scripted to represent different levels of ability. Finally, et al. (2006) found that mini-CEX scores were correlated with the results of a Royal College oral examination.

Clinical encounter cards (CEC)

The CEC system, developed at McMaster University in Canada (Hatala & Norman 1999) and subsequently implemented in other centres (Paukert et al. 2002), is similar to the mini-CEX. The basic purpose of this assessment strategy is also to score trainee performance based on direct observation of a patient encounter. The encounter card system scores the following dimensions of observed clinical practice: history-taking, physical examination, professional behaviour, technical skill, case presentation, problem formulation (diagnosis) and problem solving (therapy). Each dimension is scored using a 6-point rating scale describing performance as 1: unsatisfactory, 2: below the expected level of student performance, 3: at the expected level of student performance, 4: above the expected level of student performance, 5: outstanding student performance, and 6: performance at the level of a medical graduate.

In addition to capturing the quality of the performance, the 4 × 6 inch score cards also provide space for assessors to record the feedback given to the trainee at the end of the encounter.

This system has been shown to be a feasible, valid, and reliable measure of clinical competence, provided that a sufficient number of encounters (approximately 8 encounters for a reliability coefficient of 0.8 or more) are collected (Hatala & Norman 1999). Moreover, introduction of the system was found to increase student satisfaction with the feedback

Please refer to www.hcat.nhs.uk for guidance on this form and details of expected competencies for F1

Mini-Clinical Evaluation Exercise (CEX) - F1 Version

Please complete the questions using a cross: ☒ Please use black ink and CAPITAL LETTERS

Doctor's Surname:

Forename:

GMC Number: **GMC NUMBER MUST BE COMPLETED**

Clinical setting: A&E OPD In-patient Acute Admission GP Surgery

Clinical problem category: Airway/Breathing CVS/Circulation Gastro Neuro Pain Psych/Behav Other

New or FU: New FU Focus of clinical encounter: History Diagnosis Management Explanation

Number of times patient seen before by trainee: 0 1-4 5-9 >10 Complexity of case: Low Average High

Assessor's position: Consultant GP SpR SASG SHO Other

Number of previous mini-CEXs observed by assessor with any trainee: 0 1 2 3 4 5-9 >9

Please grade the following areas using the scale below:	Below expectations for F1 completion	Borderline for F1 completion	Meets expectations for F1 completion	Above expectations for F1 completion	U/C*
	1. History Taking	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
2. Physical Examination Skills	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
3. Communication Skills	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
4. Clinical Judgement	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
5. Professionalism	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
6. Organisation/Efficiency	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
7. Overall clinical care	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

*U/C Please mark this if you have not observed the behaviour and therefore feel unable to comment.

Anything especially good? **Suggestions for development**

Agreed action:

Have you had training in the use of this assessment tool?: Face-to-Face Have Read Guidelines Web/CD rom

Assessor's Signature:

Date (mm/yy): /

Time taken for observation: (in minutes)

Time taken for feedback: (in minutes)

Assessor's Surname:

Assessor's registration number:

Please note: Failure of return of all completed forms to your administrator is a probity issue

Acknowledgements: Adapted with permission from American Board of Internal Medicine




Figure 1. Mini-clinical evaluation exercise form. Source: www.hcat.nhs.uk.

process (Paukert et al. 2002) and to have modest correlations with other forms of assessment (Richards et al. 2007).

Clinical work sampling (CWS)

This assessment method, developed in Canada, is also based on direct observation of clinical performance in the workplace (Turnbull et al. 2000). The method requires collection of data concerning specific patient encounters for a number of different domains either at the time of admission (admission rating form) or during the hospital stay (ward rating form). These forms are completed by faculty members directly observing trainee performance. The domains assessed by faculty include: communication skills, physical examination skills, diagnostic acumen, consultation skills, management skills, interpersonal behaviour, continued learning skills and health advocacy skills. Not all skills are evaluated on each occasion.

Trainees are also assessed by ward nursing staff (using the multidisciplinary team rating form) and the patients (using the patient rating form) who are in the care of the trainees. These rating forms, also completed on the basis of directly observed behaviour, require a global assessment and ratings of the following domains: therapeutic strategies, communications skills, consultation with other health care professionals, management of resources, discharge planning, interpersonal relations, collaboration skills, and health advocacy skills and professionalism.

All rating forms use a 5-point rating scale ranging from unsatisfactory to excellent performance. This assessment method has also been shown to be valid and reliable provided a sufficient number (approximately 7 encounters for a reliability coefficient of 0.7) of encounters are observed (Turnbull et al. 2000).

A later study found that the CWS strategy could be adapted to radiology residency using a handheld computerised device (Finlay et al. 2006). Compliance with voluntary participation was not as great as expected but this evaluation format included the opportunity to discuss performance at the time of data entry, rather than at the end of rotation. The investigators found the method less useful for summative purposes although the sample size was small ($N=14$).

Blinded patient encounters

This formative assessment method is based on the same principle as the three assessment methods already mentioned. It is unique, however, in that it forms part of undergraduate bedside teaching sessions. (Burch et al. 2006). Students, in groups of 4-5, participate in a bedside tutorial. It starts with a period of direct observation in which one of the students in the group is observed performing a focused interview or physical examination as instructed by the clinician educator conducting the teaching session. Thereafter the student is expected to provide a diagnosis, including a differential diagnosis, based on the clinical findings.

The patient is unknown to the student, hence the term 'blinded' patient encounter (McLeod & Meagher 2001). This type of patient encounter has the advantage of safely allowing the trainee to practice information gathering, hypothesis

generation, and problem solving without access to the workup by more senior doctors.

After the presentation, the session focuses on demonstrating the important clinical features of the case as well as discussing various issues, for example appropriate investigation and treatment relevant to the patient's presenting clinical problem. It concludes with a feedback session in which the student receives personal private advice about his/her performance.

Feedback is provided using a 9-point rating scale for assessment of clinical interviewing and examination skills as well as clinical reasoning skills. The rating scale ranges from 1-3 for poor performance, 4-6 for adequate performance and 7-9 for good performance. Space is provided on the score sheet to add other written comments. Students keep the score sheets which are only used for feedback purposes.

Direct observation of procedural skills (DOPS)

This assessment method (Figure 2, Source: www.hcat.nhs.uk), developed in the UK, focuses on evaluating the procedural skills of postgraduate trainees by observing them in the workplace setting (Wragg et al. 2003). Just as in CWS and the Encounter Card Assessment systems, trainees' performance is scored using a 6-point rating scale where 1-2 is below the expected level of competency, 3 reflects a borderline level of competency, 4 meets the expected level of competency and 5-6 are above the expected level of competency. The assessment procedure is generally expected to require 15 minutes of observation time and 5 minutes dedicated to feedback.

Trainees are provided with a list of commonly performed procedures for which they are expected to demonstrate competence such as endotracheal intubation, nasogastric tube insertion, administration of intravenous medication, venepuncture, peripheral venous cannulation and arterial blood sampling. They are assessed by multiple clinicians on multiple occasions throughout the training period.

This method of procedural skills assessment is not limited to postgraduate training programmes. Paukert and colleagues have included basic surgical skills to be mastered by undergraduate students in their clinical encounter card system (Paukert et al. 2002).

Although DOPS is similar to procedural skills log books, the purpose and nature of these methods differ significantly. The recording of procedures is common to both of them, but log books are usually designed to ensure that trainees have simply performed the minimum number required to be considered competent. The provision of structured feedback based on observation of a performance is not necessarily part of the log book process. Moreover, the procedure is not necessarily performed under direct observation and little feedback, if any, is expected to be given. In contrast, DOPS ensures that trainees are given specific feedback based on direct observation so as to improve their procedural skills.

Case-based discussion (CbD)

This assessment method is an anglicised version of Chart-Stimulated Recall (CSR) developed for use by the American

Please refer to www.hcat.nhs.uk for guidance on this form and details of expected competencies for F1

Direct Observation of Procedural Skills (DOPS) - F1 Version

Please complete the questions using a cross: ☒ Please use black ink and CAPITAL LETTERS

Doctor's Surname:

Forename:

GMC Number: **GMC NUMBER MUST BE COMPLETED**

Clinical setting: A&E OPD In-patient Acute Admission GP Surgery

Procedure Number: Other:

Assessor's position: Consultant GP SpR SASG AHP Nurse Specialist Nurse
 Other (please specify)

Number of previous DOPS observed by assessor with any trainee: 0 1 2 3 4 5-9 >9

Number of times procedure performed by trainee: 0 1-4 5-9 >10 Difficulty of procedure: Low Average High

Please grade the following areas using the scale below:	Below expectations for F1 completion		Borderline for F1 completion	Meets expectations for F1 completion	Above expectations for F1 completion		U/C*
	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
1. Demonstrates understanding of indications, relevant anatomy, technique of procedure	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
2. Obtains informed consent	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
3. Demonstrates appropriate preparation pre-procedure	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
4. Appropriate analgesia or safe sedation	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
5. Technical ability	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
6. Aseptic technique	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
7. Seeks help where appropriate	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
8. Post procedure management	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
9. Communication skills	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
10. Consideration of patient/professionalism	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
11. Overall ability to perform procedure	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

*U/C Please mark this if you have not observed the behaviour and therefore feel unable to comment.

Please use this space to record areas of strength or any suggestions for development.

Have you had training in the use of this assessment tool?: Face-to-Face HaveReadGuidelines Web/CDrom

Assessor's Signature:

Date (mm/yy): /

Time taken for observation: (in minutes)

Time taken for feedback: (in minutes)

Assessor's Surname:

Assessor's registration number:

Please note: Failure of return of all completed forms to your administrator is a probity issue




Figure 2. Directly observed procedural skills form. Source: www.hcat.nhs.uk.

Board of Emergency Medicine (Maatsch et al. 1983). It is currently part of the Foundation Programme implemented for postgraduate training in the UK National Health Service. In CbD, the trainee selects two case records of patients in which they had made notes and presents them to an assessor. The assessor selects one of the two for discussion and explores one or more aspects of the case, including: clinical assessment, investigation and referral of the patient, treatment, follow-up and future planning, and professionalism (Figure 3, Source: www.mmc.nhs.uk). Since the case record is available at the time of assessment, medical record keeping can also be assessed by the examiner.

This type of performance assessment focuses on evaluating the clinical reasoning of trainees so as to understand the rationale behind decisions made in authentic clinical practice. As with other assessment methods described, each encounter is expected to last no more than 20 minutes, including 5 minutes of feedback. Trainees are expected to engage in multiple encounters with multiple different examiners during the training period.

There are several studies supporting the validity of this measure. Maatsch et al. (1983) collected several assessments for a group of practicing doctors eligible for recertification in Emergency Medicine. They found that CbD correlated with a number of the other measures, including chart audit. The score distribution and pass-fail results were consistent with scores on initial certification, ten years earlier. As importantly, CbD was considered the most valid of the measures by the practicing doctors participating in the study.

A study by Norman and colleagues compared a volunteer group of doctors to those referred for practice difficulties (Norman et al. 1989). CbD was highly correlated with a standardised patient examination and with an oral examination. More importantly, it was able to separate the volunteer group from the doctors who were referred. Likewise, Solomon et al. (1990) collected data from several different assessments on practicing doctors eligible for recertification. CbD was correlated with the oral examination as well as written and oral exams administered 10 years earlier.

MultiSource feedback (MSF)

More commonly referred to as 360-degree assessment, this method represents a systematic collection of performance data and feedback for an individual trainee, using structured questionnaires completed by a number of stakeholders. The assessments are all based on directly observed behaviour (Wragg et al. 2003) but they differ from the methods presented above in that they reflect routine performance, rather than performance during a specific patient encounter.

Although there are a number of different ways of conducting this form of assessment, the mini-peer assessment tool (mini-PAT) that has been selected for use in the Foundation Programme in the UK is a good example. Trainees nominate 8 assessors including senior consultants, junior specialists, nurses and allied health service professionals. Each of the nominated assessors receives a structured questionnaire (Figure 4) which is completed and returned to a central location for processing. Trainees also complete self-assessments, using the same

questionnaires, and submit these for processing. The categories of assessment include: good clinical care, maintaining good clinical practice, teaching and training, relationships with patients, working with colleagues and an overall assessment.

The questionnaires are collated and individual feedback is prepared for trainees. Data are provided in a graphic form which depicts the mean ratings of the assessors and the national mean rating. All comments are included verbatim, but they remain anonymous. Trainees review this feedback with their supervisor and together work on developing an action plan. This process is repeated twice yearly during the training period.

This method is widely used in industry and business, but has also been found to be useful in medicine. Applied to practicing doctors, it was able to distinguish certified from non-certified internists and the results were associated with performance on a written examination (Ramsey et al. 1989; Wenrich et al. 1993). In a follow-up study, two subscales were identified—one focused on technical/cognitive skills and the other focused on professionalism (Ramsey et al. 1993). Written examination performance was correlated with the former but not the latter.

Multisource feedback has been applied to postgraduate trainees as well as practicing doctors. The Sheffield Peer Review Assessment Tool, which is the full scale version of mini-PAT as shown in Figure 4 (Source: www.mmc.nhs.uk), was studied with paediatricians and found to be feasible and reliable (Archer et al. 2005). It also separated doctors by grade and tended to be insensitive to potential biasing factors such as the length of the working relationship. Whitehouse et al. (2002) also applied multisource feedback to postgraduate trainees with reasonable results.

Finally, this form of assessment has also been used successfully with medical students (Arnold et al. 1981, Small et al. 1993). Both positive and negative reports from peers have influenced academic actions.

Overall, reasonably reliable results can be achieved with the assessments of 8 to 12 peers.

Nature of the feedback

For the purpose of this discussion, feedback can be conceptualised as ‘*information provided by an agent (teacher, peer, self, etc.) regarding aspects of one’s performance or understanding*’ (Hattie & Timperley 2007). This information can be used by the learner to ‘*confirm, add to, overwrite, tune or restructure information in memory, whether that information is domain knowledge, meta-cognitive knowledge, belief about self and tasks or cognitive tactics and strategies*’ (Winnie & Butler 1994). The main purpose of feedback is, therefore, to reduce the discrepancy between current practices or understandings and desired practices or understandings (Hattie & Timperley 2007).

Perspective of the learner

In order for feedback to fulfil this purpose, it needs to address three fundamental questions for the learner:

- Where am I going?
- How am I going?
- Where to next?

Please refer to curriculum at www.mmc.nhs.uk for details of expected competencies for F1 and F2
mini-PAT (Peer Assessment Tool) - F1 Version

Please complete the questions using a cross: ☒ Please use black ink and CAPITAL LETTERS

Doctor's Surname

Forename

GMC Number:

How do you rate this Doctor in their:	Below expectations for F1 completion		Borderline for F1 completion	Meets expectations for F1 completion	Above expectations for F1 completion		U/C*
	1	2	3	4	5	6	
Good Clinical Care							
1 Ability to diagnose patient problems	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
2 Ability to formulate appropriate management plans	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
3 Awareness of their own limitations	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
4 Ability to respond to psychosocial aspects of illness	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
5 Appropriate utilisation of resources e.g. ordering investigations	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Maintaining good medical practice							
6 Ability to manage time effectively / prioritise	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
7 Technical skills (appropriate to current practice)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Teaching and Training, Appraising and Assessing							
8 Willingness and effectiveness when teaching/training colleagues	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Relationship with Patients							
9 Communication with patients	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
10 Communication with carers and/or family	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
11 Respect for patients and their right to confidentiality	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Working with colleagues							
12 Verbal communication with colleagues	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
13 Written communication with colleagues	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
14 Ability to recognise and value the contribution of others	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
15 Accessibility/Reliability	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
16 Overall, how do you rate this doctor compared to a doctor ready to complete F1 training?	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Do you have any concerns about this doctor's probity or health? Yes No
 If yes please state your concerns:

*U/C Please mark this if you have not observed the behaviour and therefore feel unable to comment. 6927534062

Figure 4. Mini-peer assessment questionnaire. Source: www.mmc.nhs.uk.

Anything especially good?

Please describe any behaviour that has raised concerns or should be a particular focus for development:

Please continue your comments on a separate sheet if required

Your Gender: Male Female

Your ethnic group:

<input type="checkbox"/> British <input type="checkbox"/> Irish <input type="checkbox"/> Other White Background <input type="checkbox"/> Caribbean <input type="checkbox"/> African <input type="checkbox"/> Any other Black background <input type="checkbox"/> Indian <input type="checkbox"/> Pakistani	<input type="checkbox"/> Bangladeshi <input type="checkbox"/> Other Asian Background <input type="checkbox"/> White and Black Caribbean <input type="checkbox"/> White and Black African <input type="checkbox"/> White and Asian <input type="checkbox"/> Any other mixed background <input type="checkbox"/> Chinese <input type="checkbox"/> Any other ethnic group
---	---

Which environment have you primarily observed the doctor in?
(Please choose one answer only)

<input type="checkbox"/> Inpatients	<input type="checkbox"/> Intensive Care
<input type="checkbox"/> Outpatients	<input type="checkbox"/> Theatre
<input type="checkbox"/> Both In and Out-patients	<input type="checkbox"/> General Practice
<input type="checkbox"/> A&E/Admissions	<input type="checkbox"/> Other (Please specify)
<input type="checkbox"/> Community Speciality	<input style="width: 100%;" type="text"/>
<input type="checkbox"/> Laboratory/Research	

Your position:

<input type="checkbox"/> Consultant	<input type="checkbox"/> SASG	<input type="checkbox"/> SpR	<input type="checkbox"/> Foundation/PRHO
<input type="checkbox"/> Nurse	<input type="checkbox"/> SHO	<input type="checkbox"/> Allied Health Professional	
<input type="checkbox"/> GP	<input type="checkbox"/> Other (Please specify) <input style="width: 100%;" type="text"/>		

If you are a Nurse or AHP how long have you been qualified?: years Length of working relationship: months

What training have you had in the use of this assessment tool?: Face-to-Face Have Read Guidelines Web/CD rom

How long has it taken you to complete this form (in minutes)?:

Your Signature:

Date: / /

Your Surname:

Your GMC Number: (Doctors only)

Acknowledgements: mini-PAT is derived from SPRAT (Sheffield Peer Review Assessment Tool) 5563534067

Figure 4. Continued.

To address the first question, it is critical that there be clearly defined learning goals. If the goals are not clearly articulated then *'the gap between current learning and intended learning is unlikely to be sufficiently clear for students to see a need to reduce it'* (Hattie & Timperley 2007). Goals can be wide ranging and variable, but without them students are less likely to engage in properly directed action, persist at tasks in the face of difficulties, or resume the task if disrupted (Bargh et al. 2001). The existence of goals is also more likely to lead students to seek and receive feedback, especially if they have a shared commitment to achieving them (Locke & Latham 1990). So, medical trainees need to have a clear understanding of desired practice or competence in order to seek feedback and stay focused on the task of achieving competence in the domain of interest.

The second question focuses on the provision of concrete information, derived from an assessment of the performance, relative to a task or goal. To do so well requires criteria that provide clear indicators of whether the task has been completed properly. The answer to this question addresses the traditional, restricted definition of feedback. Nonetheless, it is critical to the provision of effective feedback. Ironically, it is precisely this aspect of feedback which is usually poorly done. Clinician-educators are often reluctant to provide honest feedback, particularly in the face of poor performance. Having a set of clearly defined criteria makes it somewhat easier to provide guidance based strictly on observed performance, rather than interpretations of the trainee's intentions.

The final important question from the perspective of the trainee is what actions need to be taken in order to close the gap between actual performance and desired performance. Trainees need an action plan; specific information about how to proceed in order to achieve desired learning outcomes. As indicated previously, without honest feedback regarding actual performance, trainees are unlikely to seek advice about how to proceed in order to close the learning gap.

The interrelatedness of these questions becomes apparent when attempting to address this final question. Indeed, without clearly defined learning outcomes, including criteria which make achievement of the learning goals explicit, and honest feedback about observed performance, planning aimed at improving performance will not take place. Closing the gap between where trainees are and where they need to be is both the purpose of feedback and the source of its influence (Sadler 1989).

Focus of feedback

How effectively feedback addresses the three questions for learners is dependent in part on what aspects of the performance are addressed. Specifically, there are four foci for feedback (Hattie & Timperley 2007):

- feedback about the task;
- feedback about the process of the task;
- feedback about self-regulation;
- feedback about the self as a person.

The most basic focus of feedback addresses the quality of the task performed. Using well defined criteria, trainees are given specific information about whether they achieved the

required level of performance. This type of feedback is easiest to give, and is consequently the most frequently provided. It is most helpful when it concentrates on the performance, rather than the knowledge required for the task. The latter is best dealt with by providing direct instruction and it is not regarded as feedback (Hattie & Timperley 2007).

One of the limitations of providing feedback focused only on the task is that it is necessarily context-specific or task-specific. Consequently, it does not generalise readily to other tasks (Thompson 1998). On the other hand, providing feedback that focuses on the process can be of more value because it encourages a deeper appreciation of the performance. This involves giving feedback that enhances an understanding of relationships (the construction of meaning), cognitive processes, and transfer to different or novel situations (Marton et al. 1993). This focus for feedback is also more likely to promote deep learning (Balzer et al. 1989).

A major component of this type of feedback is the provision of strategies for error detection and correction, in other words developing the trainee's ability to provide self-feedback (Hattie & Timperley 2007). Feedback about the process underlying the task can also serve as a cueing mechanism leading to more effective information search strategies. Cueing is most useful when it assists trainees in detecting faulty hypotheses and provides direction for further searching and strategising (Harackiewicz 1979).

Feedback that focuses on self-regulation addresses the interplay between commitment, control, and confidence. It concentrates on the way trainees monitor, direct, and regulate their actions relative to the learning goal. It implies a measure of autonomy, self-control, self-direction, and self-discipline (Hattie & Timperley 2007). Effective learners are able to generate internal feedback and cognitive routines while engaged in a task (Butler & Winnie 1995).

Students who are able to self-appraise and self-manage are able to seek and receive feedback from others. At the other end of the spectrum are less effective learners who, having minimal self-regulation strategies, are more dependent on external factors, such as teachers, to provide feedback. For these learners, feedback is more effective if it directs attention back to the task and enhances feelings of self-efficacy such that trainees are likely to invest more time and become more committed to mastering the task (Kluger & DeNisi 1996).

Trainees' attributions of success and failure can have more impact than actual success or failure. Feelings of self-efficacy can be adversely affected if students are unable to relate feedback to the cause of their poor performance. In other words, feedback that does not specify the grounds on which students have achieved success or not, is likely to engender personal uncertainties and may ultimately lead to poorer performance (Thompson 1998). On the other hand, feedback that attributes performance to effort or ability is likely to increase engagement and task performance (Craven et al. 1991). Thus, when giving feedback it is critical that the assessor clearly directs the feedback to observed performance, while being aware of the impact feedback has on the self-efficacy of the trainee.

The final focus of feedback is discussed not because of its educational value but rather because it often has

adverse consequences. This feedback is typically concentrated on the personal attributes of the trainee and seldom contains task-related information, strategies to improve commitment to the task, or a better understanding of self or the task itself (Hattie & Timperley 2007). This focus for feedback is generally not effective, its impact is unpredictable, and it can have an adverse effect on learning. This is particularly true of negative feedback directed at a personal level.

Characteristics of effective feedback in the context of formative assessment

Formative assessment strategies are thought to best prompt change when they are integral to the learning process, performance assessment criteria are clearly articulated, feedback is provided immediately after the assessment event, and trainees engage in multiple assessment opportunities (Crooks 1988; Gibbs & Simpson 2004). In addition to these features, Ende (1983) suggested that specific conditions could make feedback more conducive to learning as described in Box 2.

In addition to the strategies suggested by Ende, it has also been suggested that the efficacy of feedback may be further improved by promoting trainee 'ownership' of feedback (Holmboe et al. 2004). Strategies to achieve this include:

- encouraging trainees to engage in a process of self-assessment prior to receiving external feedback;
- permitting trainees to respond to feedback;
- ensuring that feedback translates into a plan of action for the trainee.

Box 2. Specific conditions to make feedback more conducive to learning.

- Set an appropriate time and place for feedback.
- Provide feedback regarding specific behaviours, not general performance.
- Give feedback on decisions and actions, not one's interpretation of the trainees motives or intentions.
- Give feedback in small digestible quantities.
- Use language that is non-evaluative and non-judgemental.

Based on a large qualitative study, including 83 academics involved in education, Hewson & Little (1998) validated many of these literature-based recommendations. They developed a useful list of bipolar descriptors outlining feedback techniques to be adopted and avoided (Box 3).

As already mentioned, formulating an action plan at the end of a feedback session is critical to the success of formative assessment. If a plan addressing the deficiencies is not formulated, it results in failure to close the 'learning loop' and correct the identified problems (Holmboe et al. 2004). Indeed, formulation of an action plan may constitute the most critical step in providing feedback.

Beyond these actions, it is becoming increasingly recognised that ongoing coaching or mentoring improves the efficacy of feedback. This is particularly true of 360-degree feedback strategies (Luthans & Peterson 2004). Current literature in the business world reports that the role of the workplace managers has been reconceptualised such that they are seen to be facilitators of learning, creativity, and innovation rather than directors or controllers of activity. Furthermore, learning leaders or managers should foster interconnections between people and systems so as to create collective learning networks (Walker 2001). While this research has not been replicated in the medical workplace setting, the emerging success of these strategies in business suggests that similar methods merit further consideration in clinical training settings.

Faculty development

Faculty participation

From the preceding discussion it is clear that there is a need to increase the frequency of observation of trainee performance in order to provide feedback aimed at improving the quality of the services they later render in clinical practice. To this end a number of strategies have recently been implemented, but the studies of their efficacy are limited in number and they report variable success.

Holmboe and colleagues examined the impact of a scoring sheet specifically designed to remind faculty both of the dimensions of feedback and that its main purpose is to provide

Box 3. Feedback techniques to be avoided and adopted.

Feedback techniques to be avoided

Creating a disrespectful, unfriendly, closed, threatening climate
 Not eliciting thoughts or feelings before giving feedback
 Being judgemental
 Focusing on personality
 Basing feedback on hearsay
 Basing feedback on generalizations
 Giving too much/too little feedback
 Not suggesting ideas for improvement
 Basing feedback on unknown, non-negotiated goals

Feedback techniques to be adopted

Creating a respectful, open minded, non-threatening climate
 Eliciting thoughts and feelings before giving feedback
 Being non-judgemental
 Focusing on behaviours
 Basing feedback on observed facts
 Basing feedback on specifics
 Giving the right amount of feedback
 Suggesting ideas for improvement
 Basing feedback on well-defined, negotiated goals

Taken from Hewson & Little, 1998.

trainees with information about their performance aimed at improving it (Holmboe et al. 2001). In the study, the faculty control group did not receive any instruction regarding the use of the score sheet, while the intervention group received 20 minutes of instruction at the start of the clinical rotation. This information session outlined the characteristics of effective feedback and stressed the importance of direct observation of trainees to evaluate clinical competence. Results of the study indicated that while the intervention group did not provide more frequent feedback, their trainees were more satisfied with the quality of feedback they received.

Two recent studies in the Netherlands have produced similar findings. In one of the studies an undergraduate surgical clerkship was restructured in an attempt to increase the observation of trainee performance and the provision of feedback by senior faculty members (van der Hem-Stokroos et al. 2004). Restructuring of the clerkship included the introduction of a log book, a form documenting observation of skill performance, and individual appraisal by senior staff. Faculty was informed of the changes but they were not given formal instruction in trainee observation and how to provide feedback. The results indicated no significant increase in trainee observation or the provision of feedback. The authors suggest that the lack of impact of the intervention may be partly attributed to the limited input received by faculty involved in the study, particularly limited involvement in the process of restructuring the clerkship.

In the other study, Daelmans et al. (2005) introduced in-training assessment in an undergraduate medical clerkship programme. Senior clinical staff was informed about the introduction at a meeting held at the beginning of the clerkship. They also received a letter outlining the in-training assessment programme. The findings indicated that despite implementing this new programme, students were not more frequently observed performing clinical interviews and examinations in the workplace. In their discussion of the results they suggest that observation and feedback regarding student performance may have been improved if faculty members had been more frequently reminded of the programme, for example daily meetings could have been used to alert faculty to the importance and potential educational value of the programme.

In contrast to these studies, Turnbull et al. (2000) describe a strategy using clinical work sampling in which students received feedback based on directly observed patient encounters an average of eight times during a 4-week clerkship rotation. In this study, faculty members observing students in the workplace attended a 2-hour workshop outlining the assessment and feedback strategy. In addition, they received monthly communications reminding them of the project. Students were also oriented to the project before it started, and met with the research associate on a weekly basis during the clerkship rotation. Results indicated that the ongoing collection of performance data was feasible.

In another study using the clinical encounter card system, students engaged in a directly observed assessment event an average of 35 times during a 12-week surgery clerkship (Paukert et al. 2002). As in the other study, evaluators involved in the project were briefed about the project in a number of

short 15-minute meetings outlining the purpose and importance of the intervention implemented. These information sessions formed part of other meetings routinely held in the department, for example morbidity and mortality meetings. At each of these information sessions, faculty were asked to raise any issues or concerns they had regarding the project. They also received a letter explaining the assessment and feedback system prior to implementation. At the end of the clerkship, students were more satisfied with the feedback they received.

Based on these studies it is clear that a number of strategies need to be employed to successfully implement an assessment process in which trainees receive feedback based on directly observed performance in the workplace. First, it is apparent that involvement of faculty in planning an in-course formative assessment strategy is likely to enhance their engagement in the process. Second, faculty need to be thoroughly briefed about the purpose and process of the observation and feedback strategy implemented. Third, students need to be properly informed about the purpose and format of the assessment method used. In particular, it is critical that the potential learning benefits of the system are emphasized rather than the assessment aspects of the methods being used. Finally, faculty and students need to be regularly reminded of the benefit of formative assessment and the importance of keeping the assessment strategy active in the workplace.

Faculty training

While successfully implementing a formative assessment strategy in the workplace is an achievement in its own right, it is important to ensure that the quality of the observations made by attending faculty are accurate and that the feedback received by students is effective. As was highlighted earlier, faculty observations of student performance may not be sufficiently accurate to identify errors in student performance. While the use of checklists has been shown to improve the ability of assessors to detect errors in performance (Noel et al. 1992), they have not been shown to improve the overall accuracy of assessors. This is an issue that requires further research; effective strategies to address this problem clearly need to be found.

While the accuracy of examiners remains an issue needing further work, the stringency of examiners can be improved with training. A recent paper by Boulet et al. (2002) examined the stringency of examiners using the mini-CEX to evaluate directly observed trainee performance. They reported significant variability among the examiners even when they were observing the same event. Holmboe and colleagues have shown that assessor training can address this issue. In their paper, study participants engaged in a one-day video-based training session aimed at reducing variability among faculty when providing assessments and feedback on observed performance. Participants engaged in performance dimension training and frame-of-reference training (Holmboe et al. 2004). The former was accomplished by getting faculty to discuss and define key components of competence for specific clinical skills and develop criteria for satisfactory performance. The latter was addressed by giving individual faculty members the opportunity to score real-time trainee performance using

standardised patients and standardised trainees. While one faculty member scored the performance of the trainee and provided feedback, other faculty members scored the trainee's performance by watching the interview and examination on a video monitor. The encounter ended with a group discussion of how each member of the group rated the performance and reasons for the scores allocated. Finally the facilitator described what type of trainee performance the case scenario was scripted to depict.

Eight months after this faculty development effort, a set of video recordings of scripted patient encounters were again used to compare the performance of trained faculty as compared to a cohort of untrained faculty. Trained faculty were more stringent than untrained faculty members and they also reported feeling more comfortable providing trainee feedback. This study is one of the first demonstrating the beneficial impact of faculty training for the purpose of scoring performance with the intention of providing trainee feedback.

Challenges

In this closing section of the paper we wish to highlight areas where further work is needed to address some pivotal questions regarding workplace-based formative assessment and feedback. First and foremost, we need to develop strategies that will ensure successful and sustainable implementation of formative assessment in the workplace. Most of what has been done to date has been research-based, short term projects. We need studies that identify the determinants of successful, sustainable assessment and feedback strategies so that we can better understand factors that promote trainee feedback as a routine feature of training programmes rather than a unique feature of selected programmes only. Long term use may require further modification and simplification of existing methods so as to make them more user-friendly in busy clinical settings where patient care is the first priority and trainee assessment of less importance.

Based on current literature it is apparent that poor faculty participation in formative assessment and feedback strategies is probably the most significant limiting factor currently identified. Why faculty do not routinely engage in trainee assessment and feedback needs to be better understood if we wish to improve the situation. One strategy that may be of benefit would be a reward structure for busy clinicians that appropriately recognises their educational contributions and/or provides them protected time to engage in teaching activities. Another strategy would be to identify a core group of faculty whose only educational job is assessment and formative feedback. Other strategies clearly need to be identified. In any event, these realities need to be addressed before formative assessment is likely to be a routine feature of workplace-based training programmes.

Second, we need to improve the quality of the assessments and feedback given to trainees through a concerted faculty development effort. Current work indicates that feedback rarely results in the formulation of an action plan, a critical component of effective feedback, and only sometimes involves self-assessment by the trainee. Both these issues need to be addressed if feedback is to be owned by the trainee

and remedial action undertaken to improve performance. In addition, the accuracy and stringency of feedback need to be improved. Innovative strategies to address this important aspect of formative assessment need to be developed.

Finally, the impact of feedback on trainee learning behaviour and performance needs to be determined. To date there is very little information about the strategic use of formative assessment in the workplace context to drive the learning of medical trainees. The need for such data is apparent. Not only do we need to determine the impact of feedback on learning behaviour, but we also need to know what the performance-in-the-workplace benefits can be expected to be achieved by successful formative assessment strategies.

Summary

In the context of the workplace-based education of doctors, there has been concern that trainees are seldom observed, assessed, and given feedback. This has led to increasing interest in a variety of formative assessment methods that require observation and offer the opportunity for feedback, including the mini-clinical evaluation exercise, clinical encounter cards, clinical work sampling, blinded patient encounters, direct observation of procedural skills, case-based discussion, and multisource feedback. The research literature on formative assessment and feedback suggests that it is a powerful means for changing the behaviour of students and trainees.

To enhance the efficacy of the methods of workplace-based assessment, it is critical that the feedback which is provided be consistent with the needs of the learner, focus on important aspects of the performance (while avoiding personal issues), and have a series of characteristics which make it maximally effective. Since faculty play a key role in the successful implementation of formative assessment, strategies to provide training and encourage their participation are critical.

Notes on contributors

JOHN J. NORCINI, PhD has been President and CEO of the Foundation for Advancement of International Medical Education and Research (FAIMER®) since May 2002. For the 25 years before joining the Foundation, Dr. Norcini held a number of senior positions at the American Board of Internal Medicine. His principal academic interest is in the area of the assessment of physician performance.

VANESSA C. BURCH, MBChB, PhD is Associate Professor of Medicine at the University of Cape Town, South Africa. She convenes the undergraduate medical degree programme in the Faculty of Health Sciences and is also actively involved in postgraduate education in the Faculty. Her main academic interests are in the assessment of clinical competence and innovative methods of medical education in resource-constrained educational environments typical of developing countries.

References

- Arnold L, Willoughby L, Calkins V, Eberhart G. 1981. Use of peer evaluation in the assessment of medical students. *Med Educ* 56:35-41.
- Archer JC, Norcini JJ, Davies HA. 2005. Peer review of paediatricians in training using SPRAT. *Br Med J* 330:1251-1253.

- Association Of American Medical Colleges. 2004. Medical school graduation questionnaire: all schools report. Available at: URL: <http://www.aamc.org/data/gq/allschoolsreport/2004.pdf> (accessed on 11 April 2007).
- Balzer WK, Doherty ME, O'connor R Jr. 1989. Effects of cognitive feedback on performance. *Psychol Bull* 106:410-433.
- Bargh JA, Gollwitzer PM, Lee-Chai A, Barndollar K, Trötschel R. 2001. The automated will: Nonconscious activation and pursuit of behavioural goals. *J Personality Social Psychol* 81:1014-1027.
- Beck RS, Daughtridge R, Sloane PD. 2002. Physician-patient communication in the primary care office: a systematic review. *J Am Board Fam Pract* 15:25-38.
- Boulet JR, Mckinley DW, Norcini JJ, Whelan GP. 2002. Assessing the comparability of standardized patient and physician evaluations of clinical skills. *Ad Health Sci Educ* 7:85-97.
- Branch WT, Paranjape A. 2002. Feedback and reflection: teaching methods for clinical settings. *Acad Med* 77:1185-1188.
- Burch VC, Seggie JL, Gary NE. 2006. Formative assessment promotes learning in undergraduate clinical clerkships. *S Af Med* 96:430-433.
- Butler DL, Winnie PH. 1995. Feedback and self-regulated learning: a theoretical synthesis. *Rev Educ Res* 65:245-274.
- Craven RG, Marsh HW, Debus RL. 1991. Effects of internally focused feedback and attributional feedback on enhancement of academic self-concept. *J Educ Psychol* 83:17-27.
- Crooks TJ. 1988. The impact of classroom evaluation practices on students. *Rev Educ Res* 58:438-481.
- Daelmans HE, Overmeer RM, van der Hem-Stokroos HH. 2005. Reliability of the clinical teaching effectiveness instrument. *Med Educ* 39:904-910.
- Day SC, Grosso LG, Norcini JJ, Blank LL, Swanson DB, Horne MH. 1990. Residents' perceptions of evaluation procedures used by their training program. *J Gen Inter Med* 5:421-426.
- Daelmans HE, Hoogenboom RJ, Donker AJ, Scherpier AJ, Stehouwer CD, Van Der Vleuten CP. 2004. Effectiveness of clinical rotations as a learning environment for achieving competences. *Med Teach* 26:305-312.
- Driessen E, Van Der Vleuten C. 2000. Matching student assessment to problem-based learning: lessons from experience in a law faculty. *Stud Cont Educ* 22:235-248.
- Durning SJ, Cation LJ, Markert RJ, Pangaro LN. 2002. Assessing the reliability and validity of the mini-clinical evaluation exercise for internal medicine residency training. *Acad Med* 77:900-904.
- Ende J. 1983. Feedback in medical education. *J Am Med Assoc* 250:777-781.
- Finlay K, Norman GR, Stolberg H, Weaver B, Keane DR. 2006. In-training evaluation using hand-held computerized clinical work sampling strategies in radiology residency. *J Can Ass Radiol* 57:232-237.
- Frederiksen N. 1984. The real test bias. Influences on testing and teaching and learning. *Am Psychol* 39:193-202.
- Gibbs G. 1999. Using assessment strategically to change the way students learn, in: S. Brown (Ed.) *Assessment Matters in Higher Education. Choosing and using Diverse Approaches*, (Buckingham, Society for Research into Higher Education and Open University Press).
- Gibbs G, Simpson C. 2004-2005. Conditions under which assessment supports student learning. *Learn Teach Higher Educ* 1:3-31.
- Gipps C. 1999. Socio-cultural aspect of assessment. *Rev Educ Res* 24:355-392.
- Gronlund NE. 1998. *Assessment of Student Achievement*, 6th edn (Needham Heights, MA, Allyn and Bacon).
- Hampton JR, Harrison MJG, Mitchell JRA, Prichard JS, Seymour C. 1975. Relative contributions of history-taking, physical examination, and laboratory investigation to diagnosis and management of medical outpatients. *Br Med J* 2:486-489.
- Harackiewicz JM. 1979. The effect of reward contingency and performance feedback on intrinsic motivation. *J Pers Soc Psychol* 37:1352-1363.
- Hatala R, Ainslie M, Kassen BO, Mackie I, Roberts JM. 2006. Assessing the mini-Clinical Evaluation Exercise in comparison to a national specialty examination. *Med Educ* 40:950-956.
- Hatala R, Norman GR. 1999. In-training evaluation during an Internal Medicine clerkship. *Acad Med* 74:S118-S120.
- Hattie JA. 1999. *Influences on Student Learning*. Inaugural professorial address, University of Auckland, New Zealand. Available at: URL: <http://www.arts.auckland.ac.nz/staff/index.cfm?P=8650> (Accessed on 4 April 2007).
- Hattie J, Timperley H. 2007. The power of feedback. *Rev Educl Res* 77:81-112.
- Hauer KE. 2000. Enhancing feedback to students using the mini-CEX (clinical evaluation exercise). *Acad Med* 75:524.
- Herbers JE, Noel GL, Cooper GS, Harvey J, Pangaro LN, Weaver MJ. 1992. How accurate are faculty evaluations of clinical competence? *J Gen Inter Med* 4:202-208.
- Hewson MG, Little ML. 1998. Giving feedback in medical education. Verification of recommended techniques. *J Gen Inter Med* 13:111-116.
- Holmboe ES, Yepes M, Williams F, Huot SJ. 2004a. Feedback and the mini-clinical evaluation exercise. *J Gen Inter Med* 19:558-561.
- Holmboe ES, Hawkins RE, Huot SJ. 2004b. Direct observation of competence training: a randomized controlled trial. *Ann Inter Med* 140:874-881.
- Holmboe ES, Fiebach NH, Galaty LA, Huot S. 2001. Effectiveness of a focused educational intervention on resident evaluations from faculty: a randomized controlled trial. *J Gen Intern Med* 16:427-434.
- Holmboe ES, Huot S, Chung J, Norcini JJ, Hawkins RE. 2003. Construct validity of the miniClinical Evaluation Exercise (MiniCEX). *Acad Med* 78:826-830.
- Isaacson JH, Posk LK, Litaker DG, Halperin AK. 1995. Residents' perceptions of the evaluation process. *J Gen Inter Med* 10(suppl.):89.
- Kalet A, Earp JA, Kowlowitz V. 1992. How well do faculty evaluate the interviewing skills of medical students? *J Gen Inter Med* 7:499-505.
- Kassebaum DG, Eaglen RH. 1999. Shortcoming in the evaluation of students' clinical skills and behaviours in medical school. *Acad Med* 74:841-849.
- Kirch W, Schaffi C. 1996. Misdiagnosis at a university hospital in 4 medical eras. *Medicine (Baltimore)* 75:29-40.
- Kluger AN, DeNisi A. 1996. The effects of feedback interventions on performance: a historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychol Bull* 119:254-284.
- Kogan JR, Bellin LM, Shea JA. 2002. Implementation of the mini-CEX to evaluate medical students' clinical skills. *Acad Med* 77:1156-1157.
- Kogan JR, Hauer KE. 2006. Brief report: use of the mini-clinical evaluation exercise in Internal Medicine core clerkships. *J Gen Inter Med* 21:501-502.
- Little P, Everitt H, Williamson I, Warner G, Moore M, Gould C, Ferrier K, Payne S. 2001. Observational study of effect of patient centredness and positive approach on outcomes of general practice consultations. *Br Med J* 323:908-911.
- Locke EA, Latham GP. 1990. *A Theory of Goal Setting and Task Performance* (Englewood Cliffs, NJ, Prentice Hall).
- Luthans F, Peterson SJ. 2004. 360-degree feedback with systematic coaching: empirical analysis suggests a winning combination. *Hum Res Manag* 42:243-256.
- Maatsch JL, Huang R, Downing S, Barker B. 1983. Predictive validity of medical specialist examinations. *Final report for Grant HS 02038-04, National Center of Health Services Research*. Office of Medical Education Research and Development, Michigan State University, East Lansing, MI.
- Marton F, Dall'Alba G, Beaty E. 1993. Conceptions of learning. *Int. J Educ Res* 19:277-300.
- McLeod PJ, Meagher TW. 2001. Educational benefits of blinding students to information acquired and management plans generated by other physicians. *Med Teach* 23:83-85.
- National Health Service. 2007. *Modernising Medical Careers: Foundation Programmes*. Available at: URL: <http://www.mmc.nhs.uk/pages/foundation> (Accessed on 7 April 2007).
- Noel GL, Herbers JE, Caplow MP, Cooper GS, Pangaro LN, Harvey J. 1992. How well do Internal Medicine faculty members evaluate the clinical skills of residents? *J Gen Inter Med* 11:757-765.
- Norcini JJ, Blank LL, Arnold GK, Kimball HR. 1995. The mini-CEX (clinical evaluation exercise): A preliminary investigation. *Ann Inter Med* 123:795-799.
- Norcini JJ, Blank LL, Duffy FD, Fortna G. 2003. The mini-CEX: A method for assessing clinical skills. *Ann Inter Med* 138:476-481.

- Norcini JJ. 2007. Workplace-based assessment in clinical training, in: Swanwick T. (Ed.) *Understanding Medical Education series* (Edinburgh, UK: Association for the Study of Medical Education).
- Norman GR, Davis D, Painvin A, Lindsay E, Rath D, Ragbeer M. 1989. Comprehensive assessment of clinical competence of family/general physicians using multiple measures. *Proceedings of the Research in Medical Education Conference*, pp 75-79.
- Paukert JL, Richards ML, Olney C. 2002. An encounter card system for increasing feedback to students. *Am J Surg* 183:300-304.
- Peterson MC, Holbrook JH, Hales DV, Smith NL, Staker LV. 1992. Contributions of the history, physical examination and laboratory investigation in making medical diagnoses. *Wes J Med* 156:163-165.
- Ramsey P, Carline JD, Inui TS, Larson EB, LoGerfo JP, Wenrich MD. 1989. Predictive validity of certification by the American Board of Internal Medicine. *Ann Intern Med* 110:719-726.
- Ramsey P, Wenrich MD, Carline JD, Inui TS, Larson EB, LoGerfo JP. 1993. Use of peer ratings to evaluate physician performance. *J Am Med Ass* 269:1655-1660.
- Richards ML, Paukert JL, Downing SM, Bordage G. 2007. Reliability and usefulness of clinical encounter cards for a third-year surgical clerkship. *J Surg Res* 140:139-48.
- Sadler R. 1989. Formative assessment and the design of instructional systems. *Instruct Sci* 18:119-144.
- Shepard LA. 2000. The role of assessment in a learning culture. *Educ Res* 29:4-14.
- Small PA, Stevens B, Duerson MC. 1993. Issues in medical education: basic problems and potential solutions. *Acad Med* 68:S89-S98.
- Solomon DJ, Reinhart MA, Bridgham RG, Munger BS, Starnaman S. 1990. An assessment of an oral examination format for evaluating clinical competence in emergency medicine. *Acad Med* 65:S43-S44.
- Stillman PL, Haley H-L, Regan MB, Philbin MM. 1991. Positive effects of a clinical performance assessment programme. *Acad Med* 66:481-483.
- Swanson DB, Norman GR, Linn RL. 1995. Performance-based assessment: lessons from the health professions. *Educ Res* 24:5-11.
- Thomposn T. 1998. Metamemory accuracy: effects of feedback and the stability of individual differences. *Am J Psychol* 111:33-42.
- Turnbull J, MacFayden J, van Barneveld C, Norman G. 2000. Clinical works sampling. A new approach to the problem of in-training evaluation. *J Gen Inter Med* 15:556-561.
- van der Vleuten CPM. 1996. The assessment of professional competence: developments, research and practical implications. *Adv Health Sci Educ* 1:41-67.
- van der Hem-Stokroos NH, Daelmans HE, van der Vleuten CP, Haarman HJ, Scherpbier AL. 2004. The impact of multi-faceted educational structuring on learning effectiveness in a surgical clerkship. *Med Educ* 38:879-886.
- Veloski J, Boex JR, Grasberger J, Evans A, Wolfson DB. 2006. Systematic review of the literature on assessment, feedback, and physicians' clinical performance: BEME Guide No 7. *Med Teach* 28:117-128.
- Walker J. 2001. The managerial mentor-leading productive learning in the workplace: an integral view. University of Technology Sydney Research Centre Vocational Education & Training. *Productive Learning Seminar Series*, November, p.3. Available at: URL: http://www.oval.uts.edu.au/working_papers/2002WP/0209walker.pdf (Accessed on 4 April 2007).
- Wenrich MD, Carline J.D, Giles LM, Ramsey PG. 1993. Ratings of the performance of practicing internists by hospital-based registered nurses. *Acad Med* 68:680-687.
- Whitehouse A, Waltzman M, Wall D. 2002. Pilot study of 360° assessment of personal skills to inform record of in-training assessments for senior house officers. *Hosp Med* 63:172-175.
- Winnie PH, Butler DL., 1994. Student cognition in learning from teaching, in: T. Husen, & T. Postlewaite, (Eds.), *International Encyclopedia of Education*, pp. 5738-5745 (Oxford, UK: Pergamon).
- Wragg A, Wade W, Fuller G, Cowan G, Mills P. 2003. Assessing the performance of specialist registrars. *Clin Med* 3:131-134.

Copyright of *Medical Teacher* is the property of Taylor & Francis Ltd and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.

กระดาษบันทึก

กระดาษบันทึก

Question & Comments



sishee@mahidol.edu



mahidol.shee



SHEE FC



ศูนย์ความเป็นเลิศด้านการศึกษาวិทยาศาสตร์สุขภาพ (ศตว)
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 24123901



<http://shee.si.mahidol.ac.th>

