Mahidol University
Faculty of Medicine Siriraj Hospital

หน่วยพัฒนาแพทยศาสตรศึกษาและวิจัยการศึกษา
คณะแพทยศาสตร์ศิริราชพยาบาล มหาวิทยาลัยมหิดล

Assessment workshop for clinical teachers

การวัดและประเมินผลนักศึกษาชั้นคลินิก

ระหว่างวันที่ 15-17 มีนาคม 2560
ณ ห้องบรรยาย 3A01 อาคารศรีสวรินทิรา ชั้น 3A
คณะแพทยศาสตร์ศิริราชพยาบาล

เอกสารประกอบการอบรม

สอบถามเพิ่มเติม
คุณภัทรพร / คุณสุวรรณี โทร. 024199978 / 024196637
E-mail : merd.project@gmail.com

si-merd.com
MERD

(ร่าง) กำหนดการอบรมเชิงปฏิบัติ เรื่อง Assessment workshop for clinical teachers

ระหว่างวันที่ 15 – 17 มีนาคม พ.ศ.2560

ณ ห้องบรรยาย 3A01 อาคารศรีสวรินทิรา ชั้น 3A คณะแพทยศาสตร์ศิริราชพยาบาล

วันพุธที่ 15 มีนาคม พ.ศ.2560		**วิทยากรหลัก**	**วิทยากรร่วม**
08.00 – 08.30 น.	*ลงทะเบียน*		
08.30 – 10.00 น.	Basic principles of assessment	รศ. ดร.นพ.เชิดศักดิ์ ไอรมณีรัตน์	
10.15 – 11.00 น.	How to choose assessment methods	รศ. ดร.นพ.เชิดศักดิ์ ไอรมณีรัตน์	
11.00 – 12.00 น.	Multiple-choice questions item development	ศ. พญ.บุญมี สถาปัตยวงศ์	
12.00 – 13.00 น.	*รับประทานอาหารกลางวัน*		
13.00 – 14.45 น.	Multiple-choice questions item review	ศ. พญ.บุญมี สถาปัตยวงศ์	รศ. พญ.พรพรรณ กู้มานะชัย รศ. ดร.นพ.เชิดศักดิ์ ไอรมณีรัตน์ ผศ. นพ.สุประพัฒน์ สนใจพาณิชย์ ผศ. พญ.กษณา รักษมณี
15.00 – 16.00 น.	Multiple-choice questions item analysis	รศ. ดร.นพ.เชิดศักดิ์ ไอรมณีรัตน์	
วันพฤหัสบดีที่ 16 มีนาคม พ.ศ.2560		**วิทยากรหลัก**	**วิทยากรร่วม**
08.30 – 09.45 น.	Constructed response item development	ผศ. นพ.สุประพัฒน์ สนใจพาณิชย์	
10.00 – 11.15 น.	Constructed response item review	ผศ. นพ.สุประพัฒน์ สนใจพาณิชย์	รศ. พญ.พรพรรณ กู้มานะชัย รศ. ดร.นพ.เชิดศักดิ์ ไอรมณีรัตน์ ผศ. พญ.กษณา รักษมณี อ. นพ. อนุภพ จิตต์เมือง
11.15 – 12.00 น.	Long case examination	รศ. พญ.พรพรรณ กู้มานะชัย	
12.00 – 13.00 น.	*รับประทานอาหารกลางวัน*		
13.00 – 14.30 น.	Portfolio	ผศ. นพ.ตรีภพ เลิศบรรณพงษ์	
14.45 – 16.00 น.	OSCE item development	ผศ. พญ.กษณา รักษมณี	
วันศุกร์ที่ 17 มีนาคม พ.ศ.2560		**วิทยากรหลัก**	**วิทยากรร่วม**
08.30 – 10.15 น.	OSCE item review	ผศ. พญ.กษณา รักษมณี	รศ. พญ.พรพรรณ กู้มานะชัย รศ. ดร.นพ.เชิดศักดิ์ ไอรมณีรัตน์ ผศ. นพ.สุประพัฒน์ สนใจพาณิชย์ อ. นพ. อนุภพ จิตต์เมือง
10.30 – 12.00 น.	Clinical performance ratings	รศ. ดร.นพ.เชิดศักดิ์ ไอรมณีรัตน์	
12.00 – 13.00 น.	*รับประทานอาหารกลางวัน*		
13.00 – 14.00 น.	Workplace-based assessment	ผศ. พญ.กษณา รักษมณี	
14.00 – 15.15 น.	Standard setting	รศ. ดร.นพ.เชิดศักดิ์ ไอรมณีรัตน์	
15.30 – 16.00 น.	Summary	รศ. ดร.นพ.เชิดศักดิ์ ไอรมณีรัตน์	

หมายเหตุ: กำหนดการอาจมีการเปลี่ยนแปลงตามความเหมาะสม

รายชื่อผู้ร่วมอบรม

กลุ่มที่ 1

ลำดับ	คำนำหน้า	ชื่อ	สกุล	สังกัด	หน่วยงาน/ภาควิชา
1	อ. นพ.	วรุตม์	พงศาพิชญ์	คณะแพทยศาสตร์ศิริราชพยาบาล	ภาควิชาโสต นาสิก ลาริงซ์วิทยา
2	อ. พญ.	ปวีณา	พิทักษ์สุรชัย	คณะแพทยศาสตร์ศิริราชพยาบาล	ภาควิชาโสต นาสิก ลาริงซ์วิทยา
3	พญ.	ภคนันท์	ธนมิตรามณี	โรงพยาบาลพระนั่งเกล้า	กลุ่มงานโสต ศอ นาสิก
4	อ. พญ.	ศิวพร	เกียรติธนะบำรุง	คณะแพทยศาสตร์โรงพยาบาลรามาธิบดี มหาวิทยาลัยมหิดล	สาขา วิชาโสต ศอ นาสิกวิทยา
5	พญ.	กอบเกื้อ	เลาหพจนารถ	โรงพยาบาลพระนั่งเกล้า	กลุ่มงานจักษุวิทยา
6	ผศ. พญ.	นันทวัน	ปิยะภาณี	คณะแพทยศาสตร์ศิริราชพยาบาล	ภาควิชากุมารเวชศาสตร์
7	อ. พญ.	ธนพร	ไชยภักดิ์	คณะแพทยศาสตร์ศิริราชพยาบาล	ภาควิชากุมารเวชศาสตร์
8	อ. พญ.	ศศิธร	จันทรทิณ	คณะแพทยศาสตร์ศิริราชพยาบาล	ภาควิชากุมารเวชศาสตร์
9	นพ.	วสวัฒน์	ถิ่นพังงา	โรงพยาบาลหาดใหญ่	จิตเวช
10	รศ. พญ.	สุพร	อภินันทเวช	คณะแพทยศาสตร์ศิริราชพยาบาล	ภาควิชาจิตเวชศาสตร์
11	อ.	ธนยศ	สุมาลย์โรจน์	คณะแพทยศาสตร์ศิริราชพยาบาล	ภาควิชาจิตเวชศาสตร์

กลุ่มที่ 2

ลำดับ	คำนำหน้า	ชื่อ	สกุล	สังกัด	หน่วยงาน/ภาควิชา
1	ดร.	สมพงษ์	ศรีบุรี	คณะเทคนิคการแพทย์ มหาวิทยาลัยเชียงใหม่	ภาควิชารังสีเทคนิค
2	รศ. พญ.	จิตรลัดดา	วะศินรัตน์	คณะแพทยศาสตร์ศิริราชพยาบาล	ภาควิชารังสีวิทยา
3	อ. พญ.	วรปารี	สุวรรณฤกษ์	คณะแพทยศาสตร์ศิริราชพยาบาล	ภาควิชารังสีวิทยา
4	รศ. พญ.	โสภา	พงศ์พรทรัพย์	คณะแพทยศาสตร์ศิริราชพยาบาล	ภาควิชารังสีวิทยา
5	พญ.	ปิยาอร	นำไพศาล	สำนักวิชาแพทยศาสตร์ มหาวิทยาลัยเทคโนโลยีสุรนารี	สาขาศัลยศาสตร์ออร์โธปิดิกส์
6	นพ.	นัฏฐพล	ชิงชนะ	โรงพยาบาลสุรินทร์	ศัลยกรรมออร์โธปิดิกส์
7	พญ.	พัฒนา	แก้วประสิทธิ์	โรงพยาบาลพุทธชินราช พิษณุโลก	กลุ่มงานวิสัญญีวิทยา
8	พญ.	วชิราพร	ภูริภูมิ	โรงพยาบาลสุรินทร์	สาขา วิสัญญีวิทยา
9	รศ. พญ.	ศิริพร	ปิติมานะอารี	คณะแพทยศาสตร์ศิริราชพยาบาล	ภาควิชาวิสัญญีวิทยา
10	พญ.	จิรัฐคณา	จันทร์งาม	โรงพยาบาลกลาง	สาขา วิสัญญีวิทยา
11	รศ. นพ.	วิสูตร	ฟองศิริไพบูลย์	คณะแพทยศาสตร์ศิริราชพยาบาล	ภาควิชานิติเวชศาสตร์
12	นพ.	จรูญศักดิ์	นวลแจ่ม	คณะแพทยศาสตร์วชิรพยาบาล มหาวิทยาลัยนวมินทราธิราช	นิติเวชศาสตร์

กลุ่มที่ 3

ลำดับ	คำนำหน้า	ชื่อ	สกุล	สังกัด	หน่วยงาน/ภาควิชา
1	พญ.	ศรีรัตน์	มากมาย	โรงพยาบาลแพร่	กลุ่มงานเวชกรรมสังคม
2	นางสาว	ศิรินาฏ	ต้นสวรรค์	วิทยาลัยพยาบาลบรมราชชนนีนครพนม	พยาบาล
3	นาง	ปาริชาติ	เมืองขวา	วิทยาลัยพยาบาลบรมราชชนนีนครพนม	พยาบาล
4	ผศ.	จุรีรัตน์	กอเจริญยศ	วิทยาลัยพยาบาลบรมราชชนนีนครพนม	พยาบาล
5	ดร.	เพ็ญศิริ	ดำรงภคภากร	วิทยาลัยพยาบาลบรมราชชนนีนครพนม	พยาบาล
6	นางสาว	สมสมร	เรืองวรบูรณ์	วิทยาลัยพยาบาลบรมราชชนนีนครพนม	พยาบาล
7	นาง	ดวงใจ	บุญคง	วิทยาลัยพยาบาลบรมราชชนนีนครพนม	พยาบาล
8	ดร.	พิชาภรณ์	จันทนกุล	คณะพยาบาลศาสตร์ มหาวิทยาลัยสยาม	พยาบาล
9	นาง	ยุวดี	ลีลัคนาวีระ	คณะพยาบาลศาสตร์ มหาวิทยาลัยบูรพา	สาขาการพยาบาลชุมชน
10	นางสาว	วรรณภา	เพชรแก้วมณี	คณะแพทยศาสตร์ศิริราชพยาบาล	โรงเรียนผู้ช่วยพยาบาล
11	อ.	จตุพร	สุวรรณกิจ	คณะเภสัชศาสตร์ มหาวิทยาลัยพายัพ	บริบาลเภสัชกรรม
12	อ.	ภัณฑิรา	ปริญญารักษ์	คณะเภสัชศาสตร์ มหาวิทยาลัยพายัพ	เภสัชกร

กลุ่มที่ 4

ลำดับ	คำนำหน้า	ชื่อ	สกุล	สังกัด	หน่วยงาน/ภาควิชา
1	อ. นพ.	สุขสันต์	กนกศิลป์	สำนักวิชาแพทยศาสตร์ มหาวิทยาลัยเทคโนโลยีสุรนารี	สาขาวิชาศัลยศาสตร์
2	นพ.	คันธชาติ	ทัศคร	คณะแพทยศาสตร์ มหาวิทยาลัยนเรศวร	สาขา วิชาศัลยศาสตร์
3	นพ.	จิโรจน์	จิรานุกูล	คณะแพทยศาสตร์ มหาวิทยาลัยนเรศวร	สาขา วิชาศัลยศาสตร์
4	นพ.	สุรพล	ศรีวงค์พานิช	โรงพยาบาลพระนั่งเกล้า	กลุ่มงานสูติ-นรีเวชกรรม
5	พญ.	ลักขณา	ปรีชาสุข	คณะแพทยศาสตร์ศิริราชพยาบาล	ศูนย์เบาหวานศิริราช
6	พญ.	กมลทิพย์	เลิศชัยสถาพร	โรงพยาบาลจุฬาภรณ์	สาขาอายุรศาสตร์
7	พญ.	วรางคณา	พิชัยวงศ์	วิทยาลัยแพทยศาสตร์ มหาวิทยาลัยรังสิต	สาขา อายุรศาสตร์
8	ผศ. นพ.	สมคิด	อุ่นเสมาธรรม	วิทยาลัยแพทยศาสตร์ มหาวิทยาลัยรังสิต	สาขา วิชาอายุรศาสตร์
9	นพ.	ไกรรัตน์	คำดี	ศูนย์แพทยศาสตรศึกษาชั้นคลินิก โรงพยาบาลพะเยา	อายุรกรรม
10	ผศ. พญ.	สุภินดา	ศิริลักษณ์	คณะแพทยศาสตร์ มหาวิทยาลัยนเรศวร	สาข อายุรศาสตร์
11	พญ.	ศิดายุ	สุริยะ	คณะแพทยศาสตร์ มหาวิทยาลัยนเรศวร	สาข อายุรศาสตร์
12	ผศ. พญ.	สุภาวดี	มากะนัดถ์	คณะแพทยศาสตร์ มหาวิทยาลัยนเรศวร	อายุรศาสตร์ แผนกโรคระบบทางเดินหายใจและภาวะวิกฤต

กลุ่มที่ 5

ลำดับ	คำนำหน้า	ชื่อ	สกุล	สังกัด	หน่วยงาน/ภาควิชา
1	พญ.	นันทนีย์	หวังธีระนนท์	โรงพยาบาลพระนั่งเกล้า	กลุ่มงานเวชกรรมฟื้นฟู
2	พญ.	เจณณัฏฐญา	พลังแสงวิไล	โรงพยาบาลกลาง	สาขา เวชศาสตร์ฟื้นฟู
3	นางสาว	ปวีณา	สุดเดช	คณะแพทยศาสตร์ศิริราชพยาบาล	โรงเรียนกายอุปกรณ์สิรินธร
4	นางสาว	ศิรินทิพย์	แก้วทิพย์	คณะแพทยศาสตร์ศิริราชพยาบาล	โรงเรียนกายอุปกรณ์สิรินธร
5	นางสาว	สาวิตรี	ศรีทำบุญ	คณะแพทยศาสตร์ศิริราชพยาบาล	โรงเรียนกายอุปกรณ์สิรินธร
6	นางสาว	ปาริชาติ	พงษ์พานิช	คณะแพทยศาสตร์ศิริราชพยาบาล	สถานการแพทย์แผนไทยประยุกต์
7	นางสาว	ปิยาอร	สีรูปหมอก	คณะแพทยศาสตร์ศิริราชพยาบาล	สถานการแพทย์แผนไทยประยุกต์
8	นางสาว	กวินกานต์	ทองย้อย	คณะแพทยศาสตร์ศิริราชพยาบาล	สถานการแพทย์แผนไทยประยุกต์
9	อ.	ทัพพ์เทพ	ทิพยเจริญธัม	คณะแพทยศาสตร์ศิริราชพยาบาล	สถานการแพทย์แผนไทยประยุกต์
10	อ.	วัชรินทร์พร	พรหมพิทักษ์	คณะแพทยศาสตร์ศิริราชพยาบาล	สถานการแพทย์แผนไทยประยุกต์
11	นาย	อัศม์เดช	แหลงิ	คณะแพทยศาสตร์ศิริราชพยาบาล	สถานการแพทย์แผนไทยประยุกต์

15 March 2017

เอกสารประกอบการอบรม : Basic principles of assessment

Basic Principles of Assessment

นพ. เชิดศักดิ์ ไอรมณีรัตน์
ภาควิชาศัลยศาสตร์ คณะแพทยศาสตร์ศิริราชพยาบาล
มหาวิทยาลัยมหิดล

Assessment

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

www.si-merd.com

Outline

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

A Research Study

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

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

A Research Study

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

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

Cherdsak.ira@mahidol.ac.th

The Benefit of Testing

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

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

Assessment and Instructional Process

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

Four Ways that assessment can aid instruction

1. Student motivation
2. Retention and transfer of knowledge
3. Student self-assessment
4. Evaluating instructional effectiveness

Medical Council of Thailand Core Competencies (2012)

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

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

Cherdsak.ira@mahidol.ac.th

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

3. Equivalence

- การจัดสอบหัวข้อเดียวกันกับนักศึกษาระดับชั้นเรียนเดียวกัน ที่จัดสอบกันต่างเวลา ได้คะแนนที่เทียบเคียงกันได้

4. Feasibility

ความเป็นไปได้ของการจัดสอบ

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

- Time
- Money
- Expertise
- Administration

5. Educational Effect

- การประเมินผลนั้นกระตุ้นให้ผู้เรียนมีการเรียนรู้ในเรื่องที่ควรเรียนรู้ … educational benefit

www.si-merd.com March 2017

6. Catalytic Effect

- การประเมินผลก่อให้เกิดการนำผลของการสอบไปใช้ให้ feedback เพื่อสร้าง หรือส่งเสริม หรือสนับสนุนการเรียนรู้ของนักศึกษา

7. Acceptability

- ผู้เกี่ยวข้อง (stakeholders) ทั้งหมดเชื่อถือผลการประเมิน

Practical guidelines

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

Guidelines for Effective Assessment (1)

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

Guidelines for Effective Assessment (2)

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

Guidelines for Effective Assessment (3)

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

Cherdsak.ira@mahidol.ac.th

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

สิ่งไม่พึงประสงค์ในการสอบ

เชิดศักดิ์ ไอรมณีรัตน์

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

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

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

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

สำหรับการสอบในรูปแบบอื่นที่ต้องใช้กรรมการให้คะแนน เช่น OSCE การสอบข้อสอบบรรยาย หรือการสอบปากเปล่า นั้นจะมีปัจจัยที่เกี่ยวเนื่องกับกรรมการผู้ให้คะแนนเข้ามารบกวนการแปลผลคะแนนสอบได้ด้วย เช่น

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

ปัจจัยต่างๆ เหล่านี้ เป็นสิ่งที่ผู้จัดสอบต้องคำนึงถึงเสมอในการจัดสอบและตั้งมาตรการเพื่อควบคุมและกำจัดปัจจัยรบกวนเหล่านี้จากการสอบ เพื่อให้ได้ผลการสอบที่มีความเที่ยงตรง เป็นธรรม และสามารถใช้บอกความรู้ ความสามารถของนักเรียนได้ตามที่ต้องการ

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

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

ความแม่นยำของคะแนนสอบ (Reliability)

เชิดศักดิ์ ไอรมณีรัตน์

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

1. Reliability เป็นคุณสมบัติของคะแนนสอบ ไม่ใช่ตัวข้อสอบ ข้อสอบชุดหนึ่งทำการสอบกับนักเรียนกลุ่มหนึ่งพบว่ามีความแม่นยำสูง แต่เมื่อเอาข้อสอบชุดเดียวกันไปทำการสอบนักเรียนอีกกลุ่มหนึ่ง อาจมีความแม่นยำต่ำได้

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

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

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

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

อีกประเด็นหนึ่งที่มีความสำคัญคือ มีปัจจัยใดบ้างที่ส่งผลต่อค่า reliability coefficient สิ่งเหล่านี้มีความสำคัญมากเมื่อเราต้องการอธิบายว่าเหตุใดคะแนนสอบที่ได้จึงไม่แม่นยำ และเราต้องทำอย่างไรจึงจะทำให้คะแนนสอบมีความแม่นยำมากขึ้น โดยทั่วไปปัจจัยที่สำคัญที่ส่งผลต่อความแม่นยำของคะแนนสอบมีด้วยกัน 4 ปัจจัย คือ

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

15 March 2017

How to Choose Assessment Methods

เชิดศักดิ์ ไอรมณีรัตน์
ภาควิชาศัลยศาสตร์ คณะแพทยศาสตร์ศิริราชพยาบาล
มหาวิทยาลัย มหิดล

Assessment Approaches

Does

Shows how

Knows how

Knows

Miller's Pyramid

2 2

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

Multiple-Choice Questions

- Limitations
  - Cueing of correct answer
  - Random guessing
  - Testing of trivial knowledge
  - Difficulty of development of good MCQ items
  - Unable to assess psychomotor and other non-cognitive abilities

Constructed Response Items

- Constructed response items ask examinees to create responses rather than select answers from lists of possible answers.

Cherdsak.ira@mahidol.ac.th

1

www.si-merd.com March 2017

Comparison

	Selected Response	Constructed Response
Measured construct	Concrete knowledge, basic interpretation, some applications	Complex cognitive ability: problem solving, interpretation, decision making
Item construction	Simple	Complex
Cost of scoring	Low	Expensive
Type of scoring	Objective	Subjective
Rater effects	No effect	Significant factor
Reliability	High	Low

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

CR: Strengths

- Examinees' responses are non-cued: more authentic
- Able to measure higher-order cognitive tasks: application, analysis, synthesis, and evaluation
- Motivation for clinical learning

CR: Limitations

- Difficult to develop and score
- Inefficient exam format
- Expensive
- Subjectivity
- Low reliability
- Construct underrepresentation
- Cannot assess affective or psychomotor abilities

Long case Examination

An examination conducted by assigning a candidate to approach a patient under direct observation of an examiner. The candidate then collects information from history, physical exam and provides diagnosis, investigation, and treatment plan

Advantages

- Face validity
- Authentic
- Holistic
- Assessment of certain skills: decision making, medical professionalism, communication

Limitations

- Expensive
- Time consuming
- Limited content validity: case specificity
- Low reliability
  - Lack of objectivity
  - Small number of cases
  - Variation in case difficulty

Cherdsak.ira@mahidol.ac.th

2

www.si-merd.com March 2017

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

Assessment 13

OSCE

- Advantages
  - Can assess clinical skills, technical skills, communication skills
  - Standardization of cases, observations
  - Supporting research evidence

Medical Competencies

OSCE

- Limitations
  - Expensive
  - Time consuming
  - Difficult to administer
  - Many potential sources of CIV: SPs, raters, cases, scoring sheets
  - Construct underrepresentation

Medical Competencies

Performance Ratings

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

16

Performance Ratings

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

Performance Ratings

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

Cherdsak.ira@mahidol.ac.th

3

Portfolio

- A systematic collection of student work and related material that depicts a student's activities, accomplishments, and achievements in one or more school subjects. The collection should include evidence of student reflection and self-evaluation, guidelines for selecting the portfolio contents, and criteria for judging the quality of the work.

Venn JJ..Assessing students with special needs, 2nd ed.. Upper Saddle River, NJ: Merrill., 2000

Advantages of Portfolio

- Use multiple methods of assessment
- Take into account multiple assessors
- Integrate learning and assessment
- Facilitate reflection
- Promote creativity and problem solving
- Can be used for both formative and summative
- Can be used to assess attitudes and personal development
- Provide vital information for student diagnosis

Davis MH, Ponnamperuma GG. Portfolio assessment. JVME 2005; 32: 279 – 83.

Disadvantages of Portfolios

- For summative assessment, students may be reluctant to reveal weaknesses.
- Privacy and confidentiality of information on portfolio
- Difficulty in verification of the materials (plagiarism?)
- Workload (students, teachers)
- Low inter-rater reliability

Davis MH, Ponnamperuma GG. Portfolio assessment. JVME 2005; 32: 279 – 83.

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)
  - Case-based Discussion (CbD)
  - Mulitsource Feedback (MSF)

WPBA: Advantages

- 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

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

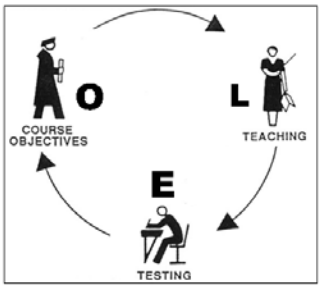
General Medical Council. Workplace based assessment: A guide for implementation

15 March 2017

MAHIDOL UNIVERSITY

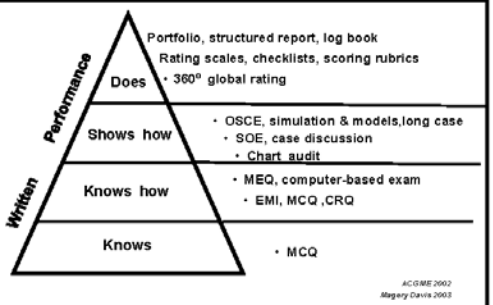
ASSESSMENT IN MEDICAL EDUCATION : MCQ

Boonmee Sathapatayavongs, MD
Prof. Channiwat Kasemsunt
Faculty of Medicine Ramathibodi Hospital
March 15, 2017



Cycle of Agreement Among Course Objectives, Teaching, and Testing

Assessment Methods According to Miller's Competency Pyramid



IS MCQ A GOOD TEST ?

IS MCQ A GOOD TEST ?

Validity
Reliability
Objectivity
Feasibility / practicability
Educational effect
Catalytic effect
Acceptability

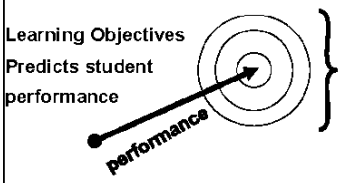
NONSENSE EXAMINATION

- An exercise to demonstrate how students who know nothing can still get good marks if we do not write good MCQ items (obtained from Professor DE Benor, Ben-Gurion University of the Negev, Beer Sheva, Israel)
- Please try your best

1

Learning Objectives and MCQs

Clear and concise learning objectives potentially deliver → Clear and concise MCQ



Boston University, Office of Med.Ed.,2005

Multiple Choices Formats (MCQ)

One-best-answer

- *Conventional (A type)*
- *Matching (B type)*
- *Extended matching (R- type)*

True / False

- *Complex or Multiple true / false (K- type)*
- *Simple true / false (X- type)*

Multiple True - False (K-Type)

A	B	C	D	E
1,2,3	1,3	2,4	4	1,2,3,4

Traumatic arteriovenous fistula produces → Stem

1. a wide pulse pressu
2. increased cardiac o
3. dilatation of the left ventricle

→ Correct Answer

4. pulmonary hypertension → Distracter

Multiple True- False (Simple, X-Type)

Serum electrolytes include ... Stem

Options

A. sodium
B. potassium
C. albumin
D. chloride
E. globulin

Answers

	Y	N
A	✓	
B	✓	
C		✓
D	✓	
E		✓

One – Best - Answer

Conventional (A type)

Extended matching (R – type)

Shape of a Well Constructed Question

Long Stem: consisted of a clinical case and all relevant facts

Lead-in: a focused question

A.
B
C. Short Options
D. (Responses)
E.

One - Best Answer (A- type) 2008

Stem :

A 2-year-old boy has a 1-week history of edema. Blood pressure is 100/60 mmHg, and there is generalized edema and ascites. Serum concentrations are: creatinine 0.4 mg/dL, albumin 14 g/L, and cholesterol 570 mg/dL. Urinalysis shows 4+ protein and no blood.

Lead-in :

Which of the following is the most likely diagnosis ?

Options:

A. Hemolytic-uremic syndrome
B. Minimal change nephrotic syndrome
C. Henoch-Schoenlein purpura with nephritis
D. Acute poststreptococcal glomerulonephritis
E. Focal and segmental glomerulosclerosis

Matching (B type)

Set 6-8

(A) Captopril
(B) Chlorthiazide
(C) Clonidine
(D) Guanethidine
(E) Propranolol

Adverse effects :

6. Postural hypotension
7. Bradycardia
8. Hypokalemia

Stem or Header
Composed of
5 individual options

An introductory phrase

Items or Trailers

Extended-Matching (R-type) Items

Theme: Fatigue

Options (6-24)

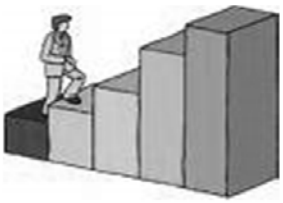
A. Acute leukemia
B. Anemia of chronic disease
C. Congestive heart failure
D. Depression
E. Epstein-Barr virus infection
F. Folate deficiency
G. Glucose 6-phosphate dehydrogenate deficiency
H. Hereditary spherocytosis
I. Hypothyroidism
J. Iron deficiency
K. Lyme disease
L. Microangiopathic hemolytic anemia
M. Miliary tuberculosis
N. Vitamin B_{12} (cyanocobalamin) deficiency

Lead-in: *For each patient with fatigue, select the most likely diagnosis.*

Stems:

1. A 19-year-old woman has had fatigue, fever, and sore throat for the past week. She has a temperature of 38.3 C (101 F), cervical lymphadenopathy, and splenomegaly. Initial laboratory studies show a leukocyte count of 5000/mm^3 (80% lymphocytes, with many lymphocytes exhibiting atypical features. Serum aspartate aminotransferase (AST, GOT) activity is 200 U/L. Serum bilirubin concentration and serum alkaline phosphatase activity are within normal limits.

Ans: E



Steps in Designing MCQ Items

Preparing Test Specification (Blue Print)

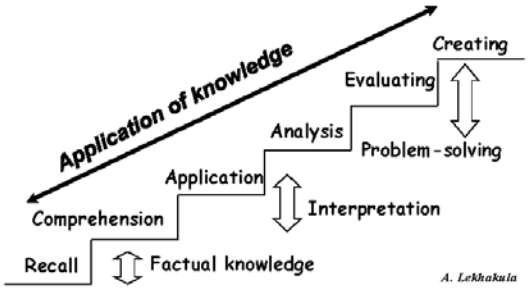
Table of Specification (Blueprint)

A two – way chart describes the sample of items to be included in the test

- Learning outcomes to be tested :
  - *Physician tasks / Level of performance*
- Selected subject matters :
  - *M.D. Curriculum*
  - *Thai Medical Council (2555)*

3

Cognitive Domain: Bloom's Taxonomy 2003



Example: Table of Specifications

	Bloom's Taxonomy			
Subject Content	Knowledge & Comprehension	Application	Analysis, Synthesis & Evaluation	Totals
Topic A	5%	10%	15%	30%
Topic B	10%	15%	45%	70%
Totals	15%	25%	60%	100%

Sampling Appropriate Contents

The content should be appropriate for the level of difficulty, and reflect the level of knowledge expected of the students

Revised Criteria for Assessment of Medical Graduates : Thai Medical Council 2555

Construct each item to assess single written objective

APPLICATION OF KNOWLEDGE

If an item requires an examinee to reach a conclusion, make a prediction, or select a course of action, etc. in a realistic situation, it is "application of knowledge"

This item helps us to assess the ability of examinees to recall the information and use them

Patient Vignettes / Scenarios

Patient vignettes should include:

- age, gender, site of care
- chief complaint
- duration
- pertinent history
- examination findings
- pertinent lab, initial treatment
- lead-in statement
- five-options set with answer

Key points for good applied MCQ writing :

- No medical term in presenting complaint
- No summaries of test or examination findings
  - Use data as full descriptions
- Tasks should model thinking process that physicians have to be able to perform
- Scenario-based questions are most useful

MCQ : Rote Memory

Which of the following has the use of carbamazepine been associated with ?

A. Hypothyroidism
B. SIADH
C. Nephrogenic diabetes insipidus
D. Leucocytosis
E. Tardive dyskinesia

MCQ : Application of Knowledge : compare & contrast

A female patient has been treated for partial complex seizure, then develops SIADH. Which of the following drugs is she most likely treated with?

A. Valproic acid
B. Clopromazine
C. Lithium
D. Carbamazepine
E. Clonazepam

MCQ : Problem solving

A 30-year-old woman presents with nausea, headache, dizziness,and confusion for the past 2 weeks. Lab.results are: serum Na 110, Cl 88 mEq/L], plasma osmolality 236, urine 420 mOsm/Kg

She has normal renal,adrenal,and thyroid function. Six weeks ago, she began drug therapy for a disorder characterized by partial complex seizures. Which of the following drugs is the most likely cause of her symptoms?

A. Valproic acid
B. Chlopromazine
C. Lithium
D. Carbamazepine
E. Clonazepam

An outbreak of food poisoning in a student cafeteria was investigated and the results are shown :

	Ate a particular food			Did not eat a particular food		
Food	Total	No ill	%	Total	No ill	%
Chicken	133	97	72.9	25	2	8
Salad	121	88	72.7	37	11	29.7
Sandwich	11	1	9.1	147	98	66.7
Soup	98	59	60.2	60	40	66.7

Which food was most probably contaminated ?

A. Chicken
B. Salad
C. Soup
D. Chicken and salad
E. Chicken, salad and soup

Following are 5 patients, all in emergency states. Whom will you send to the hospital FIRST if you have possibility to send only one? Consider yourself a proficient 6 th year student who is staying alone in a well equipped rural clinic. The driving time to the nearest hospital is 30 minutes.

A. 79-year-old man, unconscious due to accidental overdose of insulin
B. 15-year-old boy who fell from a roof and broke a thoracic spine
C. 50-year-old man with a 3 rd degree burn of face and neck
D. 28-year-old woman in labour
E. 25-year-old man with arterial hemorrhage from the groin due to work accident

Writing Good MCQ

Criteria for Good Stem

- Focuses on concepts than trivial facts
- Phrase stem as clearly as possible
- Include all information that you have to repeat in each answer option
- Sufficient information
- Avoid extra language
- No more lecture

Writing a Lead - in

- Using "Which of the following?"or "What"
- Better with complete question rather than "fill in the blank"
- Use "focus" lead – in
- Use a clear Lead- in

" Can examinee answer before reading the options ?"

Options (Distractors)

- Each option should be linked to the *ability* being measured by the lead – in
- Follow grammatically from the stem
- Be similar in grammar, length and complexity
- Are plausible but clearly incorrect
- Follow a logical order
- Be independent, mutually exclusive
- Avoid *none of the above* and *all of the above*
- Vary the position of the correct answer

Good Options (Responses)

Homogenous Responses

Are all responses similar eg, all drugs, all diseases, all laboratory tests ?

Can the examinee rank the responses on a single dimension ?

6

Criteria for Good Responses

Homogenous, parallel in content
Only one correct answer
Non-controversial
Responses similar in length, grammatical construction
No new information
Include most of the information in the stem NOT in the lengthily responses
Responses should rarely exceed one line

Avoid responses that may help unknowledgeable but test-wise examinees to select the correct responses

- long correct responses; grammatically different responses
- mutually exclusive response if one is correct
- "might", "may", "can", "usually", "rarely", "never" "always"
- double negative
- all of the above
- none of the above

Avoid tricks that may cause examinees to select incorrect responses

Vague terms
Negative terms, double negative
Reverse truths
Double options
Medical jargons
Popular slang
Abbreviations

Topic: โรค ตา หู จมูก และคอ
Keyword : ยาในรูปการเตรียมสารละลาย

โจทย์ : Rx. $NaHCO_3$ 5 g
Vehicle qs. to 100 mL

คำถาม : จากตำรับข้างต้น แพทย์ต้องการให้เภสัชกรเตรียมยาละลายชี้หูใน ปริมาณ 30 มล. แต่ห้องยามี 25 %$NaHCO_3$ แอมพูลละ 10 มล. เภสัชกรต้องเตรียมยาดังกล่าวอย่างไร

ก. ใช้ 25% $NaHCO_3$ 6 มล. และปรับปริมาตรให้ได้ 100 มล.
ข. ใช้ 25% $NaHCO_3$ 6 มล. และปรับปริมาตรให้ได้ 30 มล.
ค. ใช้ 25% $NaHCO_3$ 20 มล. และปรับปริมาตรให้ได้ 100 มล.
ง. ใช้ 25% $NaHCO_3$ 20 มล. และปรับปริมาตรให้ได้ 30 มล.
จ. ใช้ 25% $NaHCO_3$ 10 มล. และปรับปริมาตรให้ได้ 100 มล.

Topic: โรค ตา หู จมูก และคอ
Keyword : ยาในรูปการเตรียมสารละลาย

Comprehensiveness

โจทย์ : Rx. $NaHCO_3$ 5 g
Vehicle qs. to 100 mL

คำถาม : จากตำรับข้างต้น แพทย์ต้องการให้เภสัชกรเตรียมยาละลายชี้หูใน ปริมาณ 30 มล. แต่ห้องยามี 25 %$NaHCO_3$ แอมพูลละ 10 มล. เภสัชกรต้องเตรียมยาดังกล่าวอย่างไร

	25% $NaHCO_3$ (มล)	ปรับปริมาตรให้ได้ (มล)
ก.	6	100
ข	6	30
ค	20	100
ง.	20	30
จ.	10	100

ขณะที่ผู้ป่วยหญิง อายุ 63 ปี กำลังได้รับการรักษา acute myeloid leukemia ที่ต่างจังหวัด ด้วย cytotoxic chemotherapy ผู้ป่วยเกิด septicemia ขึ้นมา และได้รับการรักษาไปด้วย clindamycin และ gentamicin ต่อมาผู้ป่วยย้ายเข้ามารักษาต่อที่ท่านในกรุงเทพฯ ญาติผู้ป่วยแจ้งว่า 1 สัปดาห์หลังจากเริ่มมี septicemia ผู้ป่วยมีอาการท้องเสีย และท้องบวม อาเจียนเป็นน้ำวันละ 8 ครั้ง ไม่มีเลือดไม่มีมูกปน ผู้ป่วยมีประวัติใช้ยาปฏิชีวนะหลายตัว

PE. Temp. 38°C พบว่ามี ascites, P.R. nodular mucosa และไม่มี feces palpable
Lab. Hb 9.8 g/dL, WBC 14,000/cu mm / L , serum albumin 28 g/L
SGOT 25 IU/L (normal 5-30) SGPT 25 IU/L (normal 5-45)
sigmoidoscopy พบว่า mucosa มี reddened polypoid appearance with white exudate-membrane-like in places

ท่านคิดว่าอาการข้างเคียงของผู้ป่วยน่าจะเกิดจากโรคอะไรมากที่สุด

A. Feces impaction
B. Acute ulcerative colitis
C. Crohn's colitis
D. Pseudomembranous colitis
E. Vibrio cholera

Comprehensiveness
Excessive wording

Time for exercise !

Write 1MCQ each, for presentation & review

การออกข้อสอบให้ตรว.

รศ.พญ.วัลลี สัตยาศัย

การออกข้อสอบ license

- จำนวน 300 ข้อ : **Application ≥ 90%**

  Recall ≤ 10%

  ตั้งแต่ **2558** ภาษาอังกฤษ 100 **%**

*กำหนดตารางข้อสอบใหม่ตามเกณฑ์แพทยสภา 2555

ลักษณะข้อสอบที่ต้องการ

- เป็น **application** มีโจทย์และคำถามที่สัมพันธ์กับโจทย์ **NT 2** เป็นข้อมูลผู้ป่วย **NT 1** เป็นข้อมูลผู้ป่วยหรือเป็นโจทย์คนปกติก็ได้
- **content** ยึดตามเกณฑ์แพทยสภา (**NT2** กลุ่ม 3 ควรออกเพียงการวินิจฉัยและการดูแลรักษาเบื้องต้น)
- มี 5 ตัวเลือก

ลักษณะข้อสอบที่ต้องการ

- แต่ละข้อนิสิตนักศึกษาสามารถทำได้ทันภายใน 1 นาที ดังนั้นโจทย์ไม่ควรยาวเกินไป ควรมีเฉพาะข้อมูลที่เกี่ยวข้องกับคำถามและกลุ่มตัวเลือก
- ความยากง่ายเหมาะสมกับระดับพ.บ.
- เป็นภาษาอังกฤษ

การสร้างโจทย์

1. ให้ข้อมูลเฉพาะที่จำเป็น สัมพันธ์กับคำถามและกลุ่มตัวเลือก
2. ใช้คำน้อยที่สุดที่อ่านแล้วได้ใจความ ไม่เยิ่นเย้อ
3. ภาษาที่ใช้ควรชัดเจน ตรงจุดที่ต้องการถาม
4. ไม่ใช้คำย่อที่ไม่เป็นสากล

8

การสร้างคำถาม

1. เป็นประโยคคำถามที่สมบูรณ์
2. แต่ละข้อควรถามเพียงประเด็นเดียว
3. ไม่ใช้คำถามที่เป็น **negative** เช่น **wrong, incorrect, except, false**

การสร้างตัวเลือก

1. ขอให้มีตัวเลือกแบบ 5 ข้อ
2. มีคำตอบที่ถูกต้องเพียงคำตอบเดียว
3. คำตอบที่ถูกต้องและตัวลวง ควรสร้างให้มีความละม้ายคล้ายคลึงกัน
4. เรียงตามลำดับสั้นยาว หรือตามอักษร
5. คำตอบที่ถูกต้องไม่ควรโดดเด่นกว่าตัวลวงอื่นๆ เช่น สั้นกว่ามาก หรือยาวกว่ามาก และไม่ควรลอกคำมาโดยตรงจากตำรา
6. ตัวลวงไม่ควรมีความหมายตรงข้ามกับตัวเลือกที่ถูกต้อง
7. ไม่ใช้ **all of the above** และ **none of the above**

ตัวอย่างข้อสอบขั้นตอนที่ 1

A 10-year-old boy is stung by a bee. Five minute later the lesion is swollen and redness about 2 cm in diameter.

Which of the following is the most likely pathophysiology ?

A. Hemorrhage
B. Vasodilatation
C. Foreign body reaction
D. Neutrophilic migration
E. Lymphocytic infiltration

ตัวอย่างข้อสอบขั้นตอนที่ 2

A 25-year-old female presents with fever and sore throat for 2days. She was diagnosed with Graves' disease 3 weeks ago and has been treated with PTU. PE: T 39^{0}C, injected pharynx and tonsils. CBC: Hct 35%, WBC 2,000/cu mm (N 20, L 70, M 10 %) platelet 150,000/cu.mm.

What is the most appropriate antibiotic?

A. PGS
B. Imipenem
C. Ceftazidime
D. Azithromycin
E. Cotrimoxazole

การออกข้อสอบให้ศ.ร.ว.

- เข้า website www.si.mahidol.ac.th/issuer
- ส่งได้ตลอดเวลา

ประเด็นจริยธรรม

- ไม่ใช้ข้อสอบที่ออกให้ศรว.ไปใช้ในการสอบรายวิชา หรือ การสอบ **comprehensive** ของคณะ

A checklist for constructing one-best answer MCQs

Essential Features		
Basic Structure	Consists of stem, lead-in, and 5 options (the stem and lead-in may become combined in a question assessing factual knowledge)	
	A single best answer is included among the 5 options	
	4 distracters are included in the 5 options	
	Options are labelled A - E	
Characteristics of the stem	Based on realistic clinical vignettes / scenarios	
	Good grammatical structure	
	Clearly worded	
	Longer than the options	
Characteristics of the lead-in	Clearly worded	
	A question is asked	
Effective combination of the stem and the lead-in	Question can be answered by reading the stem and the lead-in alone, without reading the options	
Characteristics of the options	All options grammatically follow the stem	
	All options are homogenous; i.e. all belong to the same category or group	
	All options are approximately of same length	
	Numerical options are arranged in ascending or descending order	
	All options are plausible	
	Positioning of the correct answer is random	
Curriculum area assessed	Test at least one content area	
	Test at least one outcome	
Cognitive level tested	Factual recall	
	Application	
	Evaluation	

What to avoid	
Incomplete Sentences	
Use of negative terms e.g. except	
Use of "none of the above" or "all of the above" as options	
Use of absolute terms e.g. always, never	
Use of vague terms e.g. may	
Use of frequency terms e.g. rarely, often, frequently	
Testing trivial areas (this does not mean that all rare conditions should be avoided, as they may be clinically important)	

ตารางข้อสอบสำหรับการประเมินความรู้ความสามารถในการประกอบวิชาชีพเวชกรรม

ขั้นตอนที่ 1 : วิทยาศาสตร์การแพทย์พื้นฐาน

ลักษณะข้อสอบ เป็นข้อสอบปรนัยแบบ One best response เนื้อหาข้อสอบอิงตามประกาศแพทยสภาที่ 12/2555 เรื่อง เกณฑ์ความรู้ความสามารถในการประเมินเพื่อรับใบอนุญาตเป็นผู้ประกอบวิชาชีพเวชกรรม พ.ศ. 2555 ส่วนที่ 1 ก.

จำนวนข้อสอบ มี 300 ข้อ แบ่งตามหมวดที่ 1 หลักการทั่วไป และ หมวดที่ 2 การจำแนกตามระบบอวัยวะ ตามเกณฑ์ข้างต้นของแพทยสภา ดังนี้

สาระ		น้ำหนัก (รวม 100%)	จำนวนข้อสอบ (รวม 300 ข้อ)
B1 General Principles		**33%**	**100**
B1.1	Biochemistry and molecular biology		10
B1.2	Biology of cells		10
B1.3	Human development and genetics		5
B1.4	Normal immune responses		12
B1.5	Pathogenesis, pathophysiology, basic pathological process and laboratory investigation		25
B1.6	Gender, ethnic, and behavioral considerations affecting disease treatment and prevention, including psychosocial, cultural, occupational, and environment		10
B1.7	Multisystem processes		6
B1.8	General pharmacology		12
B1.9	Quantitative methods		10
B2 Hematopoietic and Lymphoreticular Systems		**7.4%**	**22**
B2.1	Normal processes		8
B2.2	Abnormal processes		10
B2.3	Principles of therapeutics		2

สาระ		น้ำหนัก (รวม 100%)	จำนวนข้อสอบ (รวม 300 ข้อ)
B2.4	Gender, ethnic, and behavioral considerations affecting disease treatment and prevention, including psychosocial, cultural, occupational, and environmental		2
B3 Central and Peripheral Nervous Systems		**7.4%**	**22**
B3.1	Normal processes		8
B3.2	Abnormal processes		10
B3.3	Principles of therapeutics		2
B3.4	Gender, ethnic, and behavioral considerations affecting disease treatment and prevention, including psychosocial, cultural, occupational, and environmental		2
B4 Skin and Related Connective Tissue		**4.0%**	**12**
B4.1	Normal processes		4
B4.2	Abnormal processes		6
B4.3	Principles of therapeutics		1
B4.4	Gender, ethnic, and behavioral considerations		1
B5 Musculoskeletal System		**4.0%**	**12**
B5.1	Normal processes		4
B5.2	Abnormal processes		6
B5.3	Principles of therapeutics		1
B5.4	Gender, ethnic, and behavioral considerations affecting disease treatment and prevention, including psychosocial, cultural, occupational, and environmental		1

สาระ		น้ำหนัก (รวม 100%)	จำนวนข้อสอบ (รวม 300 ข้อ)
B6 Respiratory System		**7.4%**	**22**
B6.1	Normal processes		8
B6.2	Abnormal processes		10
B6.3	Principles of therapeutics		2
B6.4	Gender, ethnic, and behavioral considerations affecting disease treatment and prevention, including psychosocial, cultural, occupational, and environmental		2
B7 Cardiovascular System		**7.4%**	**22**
B7.1	Normal processes		8
B7.2	Abnormal processes		10
B7.3	Principles of therapeutics		2
B7.4	Gender, ethnic, and behavioral considerations affecting disease treatment and prevention, including psychosocial, cultural, occupational, and environmental		2
B8 Gastrointestinal System		**7.4%**	**22**
B8.1	Normal processes		8
B8.2	Abnormal processes		10
B8.3	Principles of therapeutics		2
B8.4	Gender, ethnic, and behavioral considerations		2
B9 Renal/Urinary System		**7.4%**	**22**
B9.1	Normal processes		8
B9.2	Abnormal processes		10
B9.3	Principles of therapeutics		2

สาระ		น้ำหนัก (รวม 100%)	จำนวนข้อสอบ (รวม 300 ข้อ)
B9.4	Gender, ethnic, and behavioral considerations affecting disease treatment and prevention, including psychosocial, cultural, occupational, and environmental		2
B10 Reproductive System		**7.4%**	**22**
B10.1	Normal processes		8
B10.2	Abnormal processes		10
B10.3	Principles of therapeutics		2
B10.4	Gender, ethnic, and behavioral considerations affecting disease treatment and prevention, including psychosocial, cultural, occupational, and environmental		2
B11 Endocrine System		**7.4%**	**22**
B11.1	Normal processes		8
B11.2	Abnormal processes		10
B11.3	Principles of therapeutics		2
B11.4	Gender, ethnic, and behavioral considerations affecting disease treatment and prevention, including psychosocial, cultural, occupational, and environmental		2

ตารางข้อสอบสำหรับการประเมินความรู้ความสามารถในการประกอบวิชาชีพเวชกรรม
ขั้นตอนที่ 2 : ความรู้วิทยาศาสตร์การแพทย์คลินิก

ลักษณะข้อสอบ เป็นข้อสอบปรนัยแบบ One best response เนื้อหาข้อสอบอิงตามประกาศแพทยสภาที่ 12/2555 เรื่อง เกณฑ์ความรู้ความสามารถในการประเมินเพื่อรับใบอนุญาตเป็นผู้ประกอบวิชาชีพเวชกรรม พ.ศ.2555 ส่วนที่ 2-5 ข. ค. ง. และ จ.

จำนวนข้อสอบ มี 300 ข้อ แบ่งตามหมวดและกลุ่มของ Competencies ของเกณฑ์มาตรฐานฯ ของแพทยสภา ภาคผนวกที่ ดังนี้

<u>หมวดที่</u> 1 ภาวะปกติและหลักการดูแลทั่วไป จำนวนข้อสอบ 30 ข้อ

หัวข้อ	จำนวนข้อสอบ
1.1 การสร้างเสริมสุขภาพ และระบบบริบาลสุขภาพในสุขภาพของบุคคล ชุมชน และประชาชน (ดูรายละเอียดในส่วนที่ 3 ค. สุขภาพและการสร้างเสริมสุขภาพ)	4
1.2 การรวบรวมข้อมูล และประเมินปัญหาทางสุขภาพของบุคคล ครอบครัว และชุมชนในความรับผิดชอบโดยใช้ วิธีการทางระบาดวิทยาพื้นฐาน	3
1.3 การประเมินสุขภาพ และให้คำแนะนำที่เหมาะสมเพื่อความมีสุขภาพดี แก่บุคคลตามวัยและสภาวะต่างๆ ตั้งแต่ทารกในครรภ์ ทารกแรกเกิด วัยก่อนเข้าเรียน วัยเรียน วัยรุ่น วัยหนุ่มสาว ผู้ใหญ่ หญิงมีครรภ์ วัยสูงอายุ ผู้พิการ และผู้ทุพพลภาพ	3
1.4 การเชื่อมโยงความสัมพันธ์ของสุขภาพบุคคลกับสุขภาพครอบครัว ประเมินพัฒนาการและปัญหาสุขภาพของครอบครัว รวมทั้งให้คำปรึกษาและดูแลปัญหาสุขภาพเบื้องต้นแก่ผู้ป่วยและครอบครัว	2
1.5 การตรวจสุขภาพ ตรวจคัดกรองโรค และออกความเห็นหรือหนังสือรับรองความเห็นได้อย่างเหมาะสม	4
1.6 การตรวจและให้ความเห็นหรือทำหนังสือรับรองเกี่ยวกับผู้ป่วย ผู้พิการและทุพพลภาพ ผู้เสียหาย ผู้ต้องหา หรือจำเลย ตามความที่กฎหมายกำหนดให้พนักงานสอบสวน องค์กร หรือศาลในกิจการต่าง ๆ ได้ เช่น หนังสือรับรองสุขภาพ หนังสือรับรองความพิการทุพพลภาพ หนังสือรับรองการตาย การเป็นพยานต่อพนักงานสอบสวนและศาล	4
1.7 การชันสูตรพลิกศพ เก็บวัตถุพยานจากศพ ร่วมกับพนักงานสอบสวน ตามที่หมายกำหนดได้ สามารถออกรายงานการชันสูตรพลิกศพ ให้ถ้อยคำเป็นพยานในชั้นสอบสวนและชั้นศาลได้	5
1.8 การระบุปัญหา วิเคราะห์ และให้แนวทางปฏิบัติเชิงเวชจริยศาสตร์ (ดูรายละเอียดใน ส่วนที่ 4 ง. เวชจริยศาสตร์) และกฎหมายที่เกี่ยวข้อง (ดูรายละเอียดใน ส่วนที่ 5 จ. กฎหมายที่เกี่ยวข้องกับการประกอบวิชาชีพเวชกรรม)	5

<u>หมวดที่</u> 2 ภาวะผิดปกติจำแนกตามระบบอวัยวะ จำนวนข้อสอบ 270 ข้อ
จำแนกออกเป็น 3 กลุ่ม ดังนี้

กลุ่มที่ 1	กลุ่มที่ 2	กลุ่มที่ 3
45 ข้อ	225 ข้อ	
- กลไกการเกิดโรค - การวินิจฉัย**เบื้องต้น** - การบำบัดรักษา**ได้อย่างทันท่วงที** - รู้ข้อจำกัดของตนเองและปรึกษาผู้เชี่ยวชาญหรือผู้มีประสบการณ์มากกว่าได้อย่างเหมาะสม	- กลไกการเกิดโรค - การวินิจฉัย**ด้วยตนเอง** - การบำบัดรักษา**ได้ด้วยตนเอง** - การฟื้นฟูสภาพ การสร้างเสริมสุขภาพและการป้องกันโรค - การส่งต่อเพื่อปรึกษาผู้เชี่ยวชาญกรณีโรครุนแรงหรือซับซ้อนเกินความสามารถ	- กลไกการเกิดโรค - การวินิจฉัย**แยกโรค** - การแก้ไขปัญหา**เฉพาะหน้า**/หลักในการดูแลรักษา - การฟื้นฟูสภาพ การสร้างเสริมสุขภาพ และการป้องกันโรค - การตัดสินใจส่งต่อเพื่อปรึกษาผู้เชี่ยวชาญ

จำแนกตามกลุ่มโรค ดังนี้

หัวข้อ	จำนวนข้อสอบ
2.1 อาการ / ปัญหาสำคัญ	ร่วมกับโรคกลุ่มที่ 1 และโรคตามระบบ
2.2 โรคกลุ่มอาการ/ภาวะฉุกเฉิน (กลุ่มที่ 1)	45
2.3 โรคตามระบบ	225
I. INFECTIOUS AND PARASITIC DISEASES	20
II. NEOPLASM	6
III. DISEASES OF BLOOD & BLOOD FORMING ORGANS AND DISORDERS INVOLVING THE IMMUNE MECHANISM	14
IV. ENDOCRINE, NUTRITIONAL, AND METABOLIC DISEASE	15
V. MENTAL & BEHAVIORAL DISORDERS	15
VI. DISORDERS OF THE NERVOUS SYSTEM	14
VII. DISORDERS OF THE EYE AND ADNEXA	6
VIII. DISORDERS OF THE EAR & MASTOID PROCESS	6
IX. DISORDERS OF THE CIRCULATORY SYSTEM	14
X. DISORDERS OF THE RESPIRATORY SYSTEM	14
XI. DISORDERS OF THE DIGESTIVE SYSTEM	14
XII. DISORDERS OF SKIN & SUBCUTANEOUS TISSUE	7
XIII. DISORDERS OF THE MUSCULOSKELETAL SYSTEM AND CONNECTIVE TISSUE	13
XIV. DISORDERS OF THE GENITO-URINARY SYSTEM	14
XV. PREGNANCY, CHILDBIRTH, AND THE PUERPERIUM	18
XVI. CERTAIN CONDITIONS ORIGINATING IN THE PERINATAL PERIOD	8
XVII. CONGENITAL MALFORMATIONS, DEFORMATIONS AND CHROMOSOMAL ABNORMALITIES	5
XIX. INJURY, POISONING AND CONSEQUENCES OF EXTERNAL CAUSES	12
XX. EXTERNAL CAUSES OF MORBIDITY & MORTALITY	10
รวม	225

การสร้างข้อสอบปรนัย

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

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

รูปแบบพื้นฐานของข้อสอบปรนัย

ข้อสอบปรนัยคือข้อสอบชนิดที่มีคำถามแล้วมีตัวเลือกให้ผู้สอบเลือกตัวเลือกที่เหมาะสมเพื่อตอบคำถามดังกล่าว ข้อสอบปรนัยสามารถแบ่งออกได้เป็น ๒ รูปแบบ[๕] ได้แก่

๑. ข้อสอบถูกผิด (True/false item)

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

ข้อสอบชนิดถูกผิดนี้เคยเป็นที่นิยมมากในวงการแพทยศาสตรศึกษาอยู่ระยะหนึ่งเนื่องจากสามารถทดสอบความรู้ได้ปริมาณมาก แต่ข้อสอบชนิดนี้มีข้อจำกัดที่สำคัญคือสามารถใช้ได้เฉพาะกับเนื้อหาที่มีความถูกผิดชัดเจนเท่านั้น ซึ่งการตัดสินใจทางการแพทย์ส่วนมากไม่เป็นเช่นนั้น การตัดสินใจในการวินิจฉัย การตรวจค้นเพิ่มเติม หรือการรักษาผู้ป่วยส่วนใหญ่นั้นแพทย์ตัดสินใจเลือกระหว่างทางเลือกที่แตกต่างกันสามสี่อย่างซึ่งทุกทางเลือกมีความเป็นไปได้ มีส่วนถูก หรือมีความเหมาะสมในบางด้าน

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

๒. ข้อสอบเลือกคำตอบที่ถูกที่สุด (one best response item)

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

เนื่องจากข้อสอบชนิดที่มีใช้กันแพร่หลายในวงการแพทยศาสตรศึกษาในประเทศไทยในปัจจุบันคือข้อสอบประเภทที่มีตัวเลือก ๔-๕ ตัวเลือก (A-type) ผู้นิพนธ์จะขอเน้นหลักการสำหรับการออกข้อสอบประเภทนี้เป็นสำคัญ

องค์ประกอบของข้อสอบปรนัยชนิดเลือกคำตอบที่ถูกที่สุด

ข้อสอบปรนัยแต่ละข้อมีส่วนประกอบสำคัญ ๒ ส่วนด้วยกันคือ

๑. โจทย์ (stem) เป็นข้อมูลของโรค หรือภาวะหรือผู้ป่วยตามด้วยคำถาม หรือเว้นช่องว่างสำหรับเติมคำหรือข้อความที่เหมาะสมลงไป

๒. ตัวเลือก (options) คือคำ หรือข้อความที่ผู้ออกข้อสอบนำเสนอตามหลังจากโจทย์เพื่อให้ผู้สอบเลือกไปใช้ตอบคำถาม หรือเติมลงในช่องว่างในโจทย์

๒.๑ ตัวเลือกที่ถูกต้อง (correct option) เป็นคำตอบที่ถูกต้องมีเพียงตัวเลือกเดียวต่อข้อสอบข้อหนึ่ง

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

ข้อแนะนำพื้นฐานของการเขียนข้อสอบปรนัย

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

๑. เนื้อหาข้อสอบ

๑.๑ ข้อสอบหนึ่งข้อควรมุ่งเน้นประเมินความรู้เพียงเรื่องเดียว

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

๑.๒ หลีกเลี่ยงการถามความรู้ในรายละเอียดปลีกย่อยที่ไม่มีที่ใช้ทางคลินิก (trivial content)

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

๑.๓ หลีกเลี่ยงการถามความรู้ในเรื่องที่ยังมีความขัดแย้งกันในแนวทางปฏิบัติ (controversy)

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

๑.๔ หลีกเลี่ยงการลอกประโยคหรือข้อความจากตำราโดยตรง

ดังได้กล่าวแล้วว่าข้อสอบที่ดีควรมุ่งเน้นการประเมินความเข้าใจ หรือ การประยุกต์ใช้ความรู้ ไม่ควรออกข้อสอบที่ประเมินความสามารถในการจำรายละเอียดปลีกย่อย การออกข้อสอบโดยวิธีการเปิดตำราแล้วคัดลอกประโยคจากตำราโดยตรงมักจะลงเอยด้วยข้อสอบที่ทดสอบความจำว่าผู้สอบท่องเนื้อหาในตำราตรงส่วนนั้นได้หรือไม่ ข้อสอบที่ดีควรได้จากการดูผู้ป่วย โจทย์ที่ดีควรเป็นปัญหาของผู้ป่วยที่พบในการทำงานนั่นเอง ตัวเลือกก็ได้จากข้อผิดพลาดที่นักศึกษาหรือแพทย์ประจำบ้านมักปฏิบัติกับผู้ป่วยแล้วทำให้ผลการรักษาไม่ดีนั่นเอง

๑.๕ หลีกเลี่ยงการนำเสนอข้อสอบที่ประเมินความรู้ในเรื่องเดียวกันสองข้อในข้อสอบชุดเดียวกัน

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

๒. การจัดรูปแบบข้อสอบ

๒.๑ เลือกใช้คำศัพท์หรือรูปประโยคที่ง่ายต่อการทำความเข้าใจ

อาจารย์ผู้ออกข้อสอบต้องระลึกไว้เสมอว่าข้อสอบที่อาจารย์ออกเพื่อใช้ในการประเมินผลนักศึกษาแพทย์หรือแพทย์ประจำบ้านนั้นมีวัตถุประสงค์เพื่อทดสอบความรู้ทางการแพทย์เป็นสำคัญ มิใช่การประเมินความรู้ทางภาษาศาสตร์ ดังนั้นการเขียนข้อสอบของอาจารย์ควรเลือกใช้รูปแบบประโยคที่ง่ายต่อการทำความเข้าใจ อย่าเขียนประโยคซับซ้อนที่มีความยาวประโยคละหลายบรรทัด มุ่งเน้นให้ภาษาเป็นสื่อในการนำเสนอความคิดของอาจารย์ผู้ออกข้อสอบไปยังผู้สอบ อย่าให้

ภาษาเป็นอุปสรรคในการสื่อสาร การจะเลือกใช้ภาษาใดในการเขียนข้อสอบนั้นให้พิจารณาตามข้อกำหนดขององค์กรหรือหน่วยงานที่ควบคุมการสอบที่อาจารย์ส่งข้อสอบไปให้ใช้ ข้อสอบที่ใช้ในระดับการศึกษาหลักสูตรแพทยศาสตรบัณฑิตทั้งในระดับคณะ หรือข้อสอบที่ใช้ในการสอบระดับประเทศในปัจจุบันยังนิยมใช้ข้อสอบที่เขียนด้วยภาษาไทยโดยมีการใช้ศัพท์เทคนิคเป็นภาษาอังกฤษเหมือนดังภาษาที่แพทย์ใช้สื่อสารกันในการทำงานปกติ ส่วนข้อสอบในระดับหลังปริญญามีหลายการสอบที่ภาควิชา หรือราชวิทยาลัยที่เกี่ยวข้องกำหนดให้ใช้ภาษาอังกฤษทั้งหมด ก่อนที่อาจารย์จะสร้างข้อสอบต้องมีการศึกษาข้อกำหนดของแต่ละการสอบให้ดี

๒.๒ หลีกเลี่ยงการนำเสนอข้อมูลที่ไม่เกี่ยวข้องกับการแก้ปัญหาของโจทย์ข้อนั้น

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

๒.๓ จัดให้มีการตรวจสอบเนื้อหา คำศัพท์ และรูปประโยคที่ใช้ในข้อสอบแต่ละข้อก่อนนำไปใช้

ถึงแม้ว่าอาจารย์ผู้เขียนข้อสอบจะได้มีการอ่านทวนสิ่งที่ตนเองเขียนแล้วเข้าใจเนื้อหาได้ดีและคิดว่าข้อสอบอยู่ในรูปแบบที่สามารถนำไปใช้ได้แล้ว ก็ไม่ควรนำข้อสอบข้อนั้นไปใช้สอบเลย ควรให้มีคณะกรรมการข้อสอบซึ่งประกอบไปด้วยอาจารย์หลายท่านช่วยกันตรวจสอบและพิจารณาปรับแก้ข้อสอบทุกข้อก่อนนำไปใช้จริงเสมอ เนื่องจากผู้เขียนข้อสอบย่อมเข้าใจสิ่งที่ตนเขียนเสมอ แต่เมื่อผู้อื่นอ่านแล้วอาจพบว่ามีเนื้อหาที่กำกวมหรือเข้าใจโจทย์ต่างออกไปได้ การปรับแก้เนื้อหาที่มีความกำกวม หรือเฉลยซึ่งอาจารย์บางท่านอาจไม่เห็นด้วยให้ได้ข้อสอบที่มีความชัดเจน และอาจารย์ทุกท่านยอมรับในคำเฉลยได้ก่อนจะนำข้อสอบไปทำการสอบจริงย่อมเป็นสิ่งที่ดีกว่าการตรวจพบปัญหาหลังจากสอบเสร็จแล้วซึ่งต้องมาตัดสินใจกันอีกว่าจะทำอย่างไรกับการคิดคะแนนของข้อสอบข้อดังกล่าว

๓. การเขียนโจทย์

๓.๑ เขียนโจทย์ให้มีความชัดเจน ผู้สอบทุกคนอ่านแล้วมีความเข้าใจตรงกัน

ข้อแนะนำนี้อาจดูเหมือนตรงไปตรงมา แต่กลับเป็นปัญหาที่พบบ่อยมากในการพัฒนาข้อสอบปรนัย ประเด็นสำคัญคือโจทย์ที่ดีนั้นต้องมีความสมบูรณ์ในตัวเองโดยไม่ต้องอาศัยตัวเลือก โจทย์ข้อสอบที่ดีนั้นเมื่ออ่านโจทย์เสร็จแล้ว หากผู้สอบมีความรู้ในเรื่องที่ทำการประเมินนั้น เขาจะบอกคำตอบได้โดยไม่จำเป็นต้องอ่านตัวเลือกเลย ดังนั้นเมื่ออาจารย์เขียนข้อสอบเสร็จแล้วแนะนำให้ลองปิดตัวเลือกแล้วอ่านเฉพาะโจทย์ดู หากอาจารย์อ่านแล้วบอกได้ว่าโจทย์ถามอะไรและบอกได้ว่าควรตอบอะไรโดยไม่ต้องอ่านตัวเลือกจัดว่าข้อสอบข้อดังกล่าวมีโจทย์ที่มีความชัดเจน

๓.๒ เรียบเรียงเนื้อหาให้ใจความสำคัญของข้อสอบอยู่ในโจทย์

เนื่องจากข้อสอบปรนัยมีตัวเลือกที่อาจารย์ต้องสร้างขึ้นหลายตัวเลือก บางครั้งอาจารย์ผู้พัฒนาข้อสอบอาจเผลอเรอนำเอาใจความสำคัญไปใส่ไว้ในตัวเลือกซึ่งทำให้เนื้อหาในโจทย์ขาดสาระสำคัญ อ่านโจทย์แล้วไม่เข้าใจว่าผู้ออกข้อสอบต้องการถามความรู้เรื่องอะไร ตัวอย่างข้อสอบที่ไม่เป็นไปตามข้อแนะนำนี้คือข้อสอบที่ถามว่า ข้อใดต่อไปนี้ถูกต้อง หรือข้อใดต่อไปนี้ไม่ถูกต้อง แล้วเขียนรายละเอียดเกี่ยวกับโรค หรือการรักษาบางอย่างในตัวเลือกแต่ละข้อ ข้อสอบในลักษณะนี้มักทำให้

ผู้สอบต้องอ่านข้อสอบย้อนไปมาหลายรอบกว่าจะเข้าใจจุดประสงค์ของข้อสอบ แล้วจึงตัดสินใจเลือกคำตอบ โดยทั่วไปแนะนำให้อาจารย์นำเสนอรายละเอียดต่าง ๆ ไว้ในตัวโจทย์ให้มากที่สุด ส่วนตัวเลือกเขียนเป็นคำหรือข้อความสั้น ๆ

๓.๓ หลีกเลี่ยงการเขียนโจทย์ที่มีรูปประโยคเป็นเชิงปฏิเสธ

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

๔. การเขียนตัวเลือก

๔.๑ เขียนตัวเลือกที่มีประสิทธิภาพให้มีจำนวนมากที่สุดเท่าที่เหมาะสมกับบริบท

เรื่องจำนวนตัวเลือกที่เหมาะสมนี้เป็นเรื่องที่ผู้เชี่ยวชาญด้านการประเมินผลจำนวนมากสนใจ มีงานวิจัยเกี่ยวกับเรื่องจำนวนตัวเลือกที่เหมาะสมในข้อสอบปรนัยอยู่มากมาย[๙] อาจารย์ผู้ออกข้อสอบส่วนมากจะคุ้นเคยกับข้อสอบปรนัยชนิดที่มีห้าตัวเลือก บ่อยครั้งที่อาจารย์ออกข้อสอบแล้วนึกตัวเลือกได้เพียงสามหรือสี่ตัว จึงเกิดคำถามว่าจำเป็นต้องมีตัวเลือกครบห้าตัวเลือกหรือไม่ งานวิจัยบางชิ้นพบว่าการลดจำนวนตัวเลือกลงทำให้ข้อสอบง่ายขึ้น[๑๓-๑๔] แต่งานวิจัยบางชิ้นพบว่าการลดจำนวนตัวเลือกลงทำให้ได้ข้อสอบยากขึ้น[๑๕-๑๖] ผู้เชี่ยวชาญในการประเมินผลเสนอว่าข้อสอบปรนัยที่มีตัวเลือกเพียงสามตัวเลือกก็สามารถทดสอบความรู้ได้อย่างมีประสิทธิภาพ[๙, ๑๐, ๑๗] แต่มีอาจารย์จำนวนไม่น้อยที่ไม่สบายใจที่มีตัวเลือกในข้อสอบแต่ละข้อน้อยกว่าห้าตัวเลือกด้วยกังวลว่าจะทำให้มีโอกาสสูงที่ผู้สอบที่ไม่มีความรู้จะเดาสุ่มได้คำตอบที่ถูกต้อง แต่จากข้อมูลที่ปรากฏในปัจจุบันพบว่าผู้สอบในการสอบในระดับสูงนั้นพฤติกรรมการเดาสุ่มโดยที่ผู้สอบปราศจากความรู้นั้นน่าจะมีบทบาทน้อยมาก ผู้สอบส่วนใหญ่มักพอมีความรู้บ้างและสามารถตัดตัวเลือกที่ไม่สมเหตุสมผลอย่างชัดเจนได้[๙] ในการศึกษาข้อสอบปรนัยส่วนใหญ่พบตัวเลือกที่ไม่ทำงานเป็นจำนวนไม่น้อย[๙๕] ข้อมูลที่ได้จากการวิเคราะห์ข้อสอบปรนัยที่ใช้ในทางแพทยศาสตรศึกษาในประเทศไทยหลายครั้งก็สอดคล้องกับงานวิจัยในต่างประเทศที่พบว่าข้อสอบส่วนใหญ่มักมีตัวเลือกที่ทำงานจริงราวสามหรือสี่ตัวเลือก มีข้อสอบน้อยข้อมากที่ตัวเลือกทั้งห้าตัวเลือกทำงานอย่างมีประสิทธิภาพ

ด้วยข้อมูลจากการศึกษาต่าง ๆ ข้อแนะนำในการออกข้อสอบปรนัยในปัจจุบันคือให้อาจารย์เขียนจำนวนตัวเลือกมากที่สุดที่มีความเหมาะสมกับเนื้อหาโจทย์ ไม่จำเป็นต้องเขียนตัวเลือก ๕ ตัวเลือกเสมอไป เนื่องจากตัวเลือกที่ห้าที่เขียนขึ้นเพื่อเติมเต็มโดยไม่สมเหตุสมผลนั้นมักไม่ค่อยมีคนเลือก หากเนื้อหาที่อาจารย์นำมาสอบมีตัวเลือกที่เหมาะสมเพียงสามหรือสี่ตัวเลือกก็เขียนจำนวนตัวเลือกเพียงสามหรือสี่ตัวเลือก[๙๐] แต่อย่างไรก็ตามให้อาจารย์ศึกษาข้อกำหนดของแต่ละการสอบที่อาจารย์เกี่ยวข้องด้วย เนื่องจากนโยบายของแต่ละการสอบแตกต่างกันไป องค์กรที่จัดสอบทางแพทยศาสตรศึกษาจำนวนไม่น้อยยังคงตั้งข้อกำหนดให้ใช้ข้อสอบ ๕ ตัวเลือกเสมอ ซึ่งหากอาจารย์ไม่ทำตามข้อกำหนดดังกล่าวข้อสอบที่ออกไปอาจไม่ได้รับการพิจารณาได้

๔.๒ จัดให้ตัวเลือกที่ถูกต้องมีการกระจายตำแหน่งไปให้มีจำนวนพอ ๆ กันในทุกตัวเลือก

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

๔.๓ เขียนตัวเลือกแต่ละข้อให้เป็นอิสระไม่ขึ้นต่อกัน

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

๔.๔ เขียนตัวเลือกให้ทุกตัวเลือกมีความเป็นเนื้อเดียวกัน (homogeneous)

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

ในบางบริบทของการดูแลรักษาผู้ป่วย สิ่งที่แพทย์ต้องตัดสินใจเลือกอาจมีทั้งการเลือกที่จะให้การรักษาเลยหรือจะส่งตรวจค้นเพิ่มเติมก่อน ในกรณีนี้อาจารย์สามารถเขียนตัวเลือกที่มีการรักษาและการตรวจเพิ่มเติมปะปนกันได้ แต่การเขียนรูปประโยคคำถามต้องไม่เป็นการบอกใบ้ว่าจะไปทิศทางใด แต่ต้องเลือกใช้คำถามที่เป็นกลาง เช่น ท่านจะปฏิบัติต่อผู้ป่วยอย่างไร, ท่านจะดำเนินการอย่างไรต่อไป เป็นต้น

๔.๕ เขียนตัวเลือกแต่ละข้อให้มีความยาวพอ ๆ กัน

จากการสังเกตข้อสอบปรนัยจำนวนมากจะพบว่าตัวเลือกที่ถูกต้องมักมีความยาวมากกว่าตัวเลือกอื่น ซึ่งข้อสังเกตนี้ผู้สอบจำนวนไม่น้อยก็ทราบดี และผู้สอบส่วนมากเมื่อไม่ทราบคำตอบก็มักเลือกตัวเลือกที่มีความยาวมากที่สุด ดังนั้นอาจารย์ผู้ออกข้อสอบควรระมัดระวังไม่ให้ตัวเลือกตัวใดตัวหนึ่งมีความยาวแตกต่างไปจากตัวเลือกอื่นชัดเจน เพราะจะทำให้ผู้สอบเดาคำตอบที่ถูกได้ง่าย

๔.๖ หลีกเลี่ยงการใช้ตัวเลือก “ถูกทุกข้อ” หรือ “ไม่มีข้อใดถูก”

ตัวเลือก “ถูกทุกข้อ” เป็นตัวเลือกที่ผู้เชี่ยวชาญในการประเมินผลส่วนใหญ่เห็นสอดคล้องกันว่าไม่ควรใช้เนื่องจากมักช่วยใบ้ตัวเลือกที่ถูกต้องให้กับผู้สอบ ทำให้ผู้สอบส่วนหนึ่งตอบถูกโดยไม่ต้องอาศัยองค์ความรู้ที่สมบูรณ์ในเรื่องที่ทดสอบ งานวิจัยพบว่าข้อสอบที่มีตัวเลือกชนิดนี้จะมีผลให้ค่าความเที่ยงของคะแนนสอบลดลง[๑๐] จึงแนะนำให้หลีกเลี่ยงการใช้

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

การใช้ตัวเลือก “ไม่มีข้อใดถูก” จะยิ่งเป็นปัญหามากขึ้นในข้อสอบที่ถามถึงสิ่งที่ไม่ควรทำ เช่นยาใดไม่ควรใช้ในผู้ป่วย ซึ่งนอกจากยาที่นำเสนอในตัวเลือกแล้วย่อมมียาชนิดอื่นอีกมากมายในบัญชียาที่ไม่เหมาะสม ซึ่งไม่มี

ทางที่ใครจะรู้ได้ว่าการที่ผู้สอบเลือกตอบ "ไม่มีข้อใดถูก" นั้นเขาคิดถึงยาใด และยานั้นไม่เหมาะสมมากไปกว่ายาที่มีอยู่ในตัวเลือกหรือไม่ งานวิจัยทั้งหมดที่ศึกษาถึงตัวเลือกชนิดนี้ได้ข้อสรุปที่ตรงกันว่าข้อสอบที่ใช้ตัวเลือกประเภทนี้เพิ่มระดับความยากให้ข้อสอบ[๖] โดยทั่วไปแล้วจึงไม่แนะนำให้ใช้ตัวเลือกประเภทนี้ในการสอบทางแพทยศาสตรศึกษาซึ่งทางเลือกสำหรับสถานการณ์ที่นำเสนอมีได้มากและการตัดสินใจเลือกคำตอบต้องอาศัยการเปรียบเทียบข้อดีข้อเสียของแต่ละตัวเลือก

สรุป

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

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

เอกสารอ้างอิง

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

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

ข้อผิดพลาดที่ควรระวังในการสร้างข้อสอบปรนัย

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

ข้อผิดพลาดที่ควรระวังในการสร้างข้อสอบปรนัย

ข้อสอบปรนัย (multiple-choice question) เป็นรูปแบบการประเมินผลที่นิยมใช้กันอย่างแพร่หลาย ในวงการแพทยศาสตรศึกษา ข้อสอบชนิดนี้เป็นที่ชื่นชอบของนักศึกษาผู้เข้าสอบจำนวนมากเนื่องจากมีคำตอบให้เลือก หากไม่มีความรู้ก็สามารถเดาได้ ซึ่งต่างไปจากข้อสอบประเภทอัตนัยซึ่งผู้สอบต้องเขียนคำตอบจากความคิดของตนเอง[๑] ดังนั้นข้อสอบปรนัยจึงเป็นข้อสอบที่ผู้สอบทำได้ง่าย แต่ในทางตรงข้ามข้อสอบปรนัยเป็นข้อสอบที่สร้างปัญหาให้กับอาจารย์ผู้สร้างข้อสอบไม่น้อย เนื่องจากในกระบวนการเขียนข้อสอบปรนัยแต่ละข้อนั้นต้องใช้ทักษะอย่างมาก ต้องใช้ทั้งศาสตร์และศิลป์ และบ่อยครั้งอาจารย์ผู้สร้างข้อสอบก็ถูกขอให้ทำการปรับแก้ข้อสอบเนื่องจากคณะกรรมการพิจารณาข้อสอบมีความเห็นว่ารายละเอียดในข้อสอบไม่เหมาะสม มีการศึกษาวิจัยพบว่าคุณภาพของข้อสอบปรนัยที่พัฒนาขึ้นในโรงเรียนแพทย์หลายแห่งนั้นไม่สู้ดีนัก มีข้อสอบที่มีลักษณะไม่เหมาะสมอยู่จำนวนไม่น้อย[๒-๓] ข้อสอบปรนัยที่มีลักษณะไม่เหมาะสมเหล่านี้ส่งผลเสียต่อการสอบได้หลายประการ เช่น ทำให้ข้อสอบยากขึ้น สร้างความสับสนให้ผู้สอบ ทำให้ผู้สอบบางกลุ่มเสียเปรียบ และทำให้การตัดสินผลสอบผิดพลาด เป็นต้น[๓-๕] ดังนั้นการออกข้อสอบปรนัยที่มีคุณภาพดีจึงเป็นงานที่มีความสำคัญและท้าทายความสามารถ

การสร้างข้อสอบปรนัยที่มีคุณภาพดีนั้นควรเริ่มต้นจากการมีองค์ความรู้พื้นฐานในการสร้างข้อสอบแล้วเกิดการฝึกฝนทักษะ สั่งสมประสบการณ์ในการออกข้อสอบจนเกิดความชำนาญ ปัญหาที่พบบ่อยในโรงเรียนแพทย์หลายแห่งคือมีอาจารย์จำนวนไม่น้อยที่ได้รับมอบหมายให้ออกข้อสอบปรนัย โดยไม่ได้มีการพัฒนาองค์ความรู้พื้นฐานที่เหมาะสมก่อน ซึ่งเป็นเหตุให้มีข้อสอบปรนัยที่มีลักษณะไม่เหมาะสมตามหลักการออกข้อสอบปะปนมาในข้อสอบที่ให้นักศึกษาแพทย์และแพทย์ประจำบ้านทำอยู่บ้าง ผู้นิพนธ์จึงเห็นความสำคัญของการเผยแพร่องค์ความรู้พื้นฐานของการออกข้อสอบปรนัย องค์ความรู้พื้นฐานในการสร้างข้อสอบปรนัยนั้นมีสองส่วน ส่วนแรกเป็นหลักการของการสร้างข้อสอบทั่วไปซึ่งได้มีผู้รวบรวมเป็นข้อแนะนำดีพิมพ์ในตำราและวารสารทางวิชาการอยู่บ้าง[๑,๕-๗] ส่วนที่สองเป็นข้อผิดพลาดในการสร้างข้อสอบที่อาจารย์ผู้ออกข้อสอบพึงหลีกเลี่ยง ในบทความนี้ผู้นิพนธ์จะมุ่งเน้นในส่วนที่สองนี้โดยจะรวบรวมข้อผิดพลาดในการสร้างข้อสอบปรนัย ที่อาจเป็นตัวบอกใบ้ให้ผู้สอบที่ไม่มีความรู้ในเรื่องที่ทำการทดสอบสามารถเลือกคำตอบที่ถูกต้องได้ ดังนั้นการที่อาจารย์ผู้ออกข้อสอบทราบถึงสิ่งเหล่านี้และหลีกเลี่ยงเสียจะส่งผลให้ข้อสอบปรนัยที่สร้างขึ้นสามารถใช้วัดองค์ความรู้ทางการแพทย์ได้จริง โดยปราศจากปัจจัยรบกวนจากการสังเกตพบสิ่งบอกใบ้คำตอบ

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

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

๑. ข้อผิดพลาดในไวยากรณ์

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

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

A. lumbar puncture

B. computerized scan of the brain

C. glucose with Thiamine

D. Sodium bicarbonate

ในตัวอย่างที่ ๑ นี้โจทย์ให้ผู้สอบเลือกตัวเลือกไปเติมในช่องว่าง ซึ่งสิ่งที่เติมลงในช่องว่างได้นั้นต้องเป็นยาที่สามารถให้ทางหลอดเลือดดำได้ ผู้สอบที่มีทักษะการทำข้อสอบดีจะตัดตัวเลือก A และ B ได้โดยไม่ต้องใช้ความรู้ทางการแพทย์

ตัวอย่างที่ ๒. Which organism is the cause of syphilis?

A. *Neisseria gonorrheae*

B. *Chlamydia trachomatis and Giardia lamblia*

C. *Treponema pallidum*

D. *Ureaplasma urealyticum and Mycoplasma genitalium*

ในตัวอย่างที่ ๒ นี้โจทย์ถามหาเชื้อก่อโรค โดยใช้รูปประโยคถามหาคำตอบที่เป็นเอกพจน์ ดังนั้นคำตอบที่ถูกต้องย่อมมีเชื้อก่อโรคตัวเดียว ผู้สอบที่มีทักษะการทำข้อสอบดีจะตัดตัวเลือก B และ D ได้โดยไม่ต้องใช้ความรู้ทางการแพทย์

๒. การใบ้คำตอบด้วยหลักตรรกะ

ในการเขียนตัวเลือก อาจารย์ผู้ออกข้อสอบต้องระมัดระวังไม่ให้ผู้สอบสามารถตัดตัวเลือกได้ด้วยหลักตรรกศาสตร์ เนื่องจากผู้สอบที่มีทักษะการทำข้อสอบดีจะสามารถพิจารณาความเป็นไปได้ของตัวเลือกต่าง ๆ และตัดตัวลวงที่ไม่มีทางเป็นไปได้ตามหลักของเหตุและผลออกไปได้โดยไม่ต้องอาศัยความรู้เรื่องที่อาจารย์ตั้งเป้าหมายว่าจะทดสอบ

ตัวอย่างที่ ๓. ภาวะไส้เลื่อนบริเวณขาหนีบ (inguinal hernia) …….

A. พบในผู้ชายบ่อยกว่าผู้หญิง

B. พบในผู้หญิงบ่อยกว่าผู้ชาย

C. พบเกิดขึ้นในผู้หญิงและผู้ชายในอัตราเท่ากัน

D. พบบ่อยในผู้ที่มีเศรษฐานะยากจน

E. พบในผู้ที่มีภูมิลำเนาในทวีปเอเชีย มากกว่าผู้ที่มีภูมิลำเนาในทวีปยุโรป

ในตัวอย่างที่ ๓ นี้อาจารย์ผู้ออกข้อสอบต้องการวัดความรู้เรื่องอุบัติการณ์ของไส้เลื่อนขาหนีบ แต่หาก

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

๓. การใช้คำคุณศัพท์บอกระดับของความแน่ชัด

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

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

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

B. Patients with Marfan syndrome always developed inguinal hernia.

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

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

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

๔. ความยาวของตัวเลือก

มีการตั้งข้อสังเกตว่าอาจารย์แพทย์มักชอบสอนและอธิบายแม้กระทั่งในการสอบอาจารย์แพทย์หลายท่านก็ติดนิสัยรักการสอนนี้มาด้วย ทำให้อาจารย์มักเขียนตัวเลือกที่ถูกต้องที่มีคำอธิบายประกอบอย่างครบถ้วนทำให้ตัวเลือกที่ถูกมักมีความยาวมากกว่าตัวลวง นักศึกษาผู้เข้าสอบจำนวนไม่น้อยรู้ถึงความจริงข้อนี้และมักเลือกตัวเลือกที่มีความยาวมากที่สุด หากเขาไม่สามารถหาคำตอบได้ด้วยความรู้ทางการแพทย์ที่เขามี

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

A. ให้เปลี่ยนไปใช้การใส่ห่วงอนามัย

B. ให้ใช้ถุงยางอนามัย

C. ให้กินยาคุมกำเนิดต่อได้เนื่องจากมีการศึกษาแล้วว่ายาคุมกำเนิดชนิดกินไม่ส่งผลให้เกิดการเพิ่มขึ้นของน้ำหนักตัว

D. ให้รับประทานยาลดความอ้วน

ในตัวอย่างที่ ๕ นี้จะสังเกตเห็นว่าตัวเลือก C มีการอธิบายเหตุผลประกอบส่งผลให้มีความยาวมากกว่าตัวเลือกอื่นชัดเจน ลักษณะเช่นนี้จะเป็นการบอกใบ้ให้นักศึกษาเลือกตัวเลือกนี้

๕. การใช้คำซ้ำในโจทย์และตัวเลือก

การใช้คำเดียวกัน หรือคำที่มีความหมายเหมือนกันในโจทย์และตัวเลือก มักเป็นการบอกใบ้ว่าตัวเลือกดังกล่าวเป็นตัวเลือกที่ถูกต้อง

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

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

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

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

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

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

๖. การเข้าพวกของคำ หรือข้อความที่ปรากฏในตัวเลือก

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

ตัวอย่างที่ ๗. โรคที่แพทย์วินิจฉัยผิดว่าเป็นไส้ติ่งอักเสบบ่อยที่สุดเรียงลำดับจากมากไปน้อยคือ

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

B. acute mesenteric lymphadenitis, Meckel diverticulitis, acute cholecystitis

C. Meckel diverticulitis, twisted ovarian cyst, sigmoid diverticulitis

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

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

การเข้าพวกของตัวเลือกที่ถูกนั้นไม่จำเป็นต้องเป็นลักษณะของการมีจำนวน หรือความถี่ของคำมากที่สุดเพียงเท่านั้น อาจหมายรวมถึงการมีรูปร่างลักษณะ หรือความหมายคล้ายคลึงกันได้ด้วย

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

A. Amoxycillin PO

B. Ceftazidime IV + Amikacin IV

C. Amphotericin B IV + Ceftazidime IV

D. Cloxacillin IV + Metronidazole IV

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

๗. การบอกใบ้คำตอบโดยโจทย์ข้ออื่น

ข้อผิดพลาดนี้เป็นข้อผิดพลาดที่ตัวผู้เขียนข้อสอบไม่ค่อยรู้ แต่ผู้ที่จะตรวจพบข้อผิดพลาดนี้คืออาจารย์ผู้เลือกข้อสอบไปใช้ เนื่องจากในการสอบแต่ละครั้งใช้ข้อสอบจำนวนมาก หากเลือกข้อสอบโดยไม่ระมัดระวังอาจมีข้อสอบสองข้อที่ถามเกี่ยวกับโรคหรือกลุ่มอาการเดียวกัน ซึ่งข้อมูลจากโจทย์ในข้อหนึ่งอาจเป็นตัวบอกใบ้คำตอบของข้อสอบอีกข้อได้ ดังนั้นเมื่อทำการเลือกข้อสอบเสร็จแล้วจัดหน้ากระดาษเข้ารูปเล่มข้อสอบแล้วอาจารย์ควรอ่านข้อสอบฉบับสมบูรณ์นี้อีกหนึ่งหรือสองรอบก่อนส่ง

ไปพิมพ์ ซึ่งการอ่านทวนในขั้นตอนนี้อาจทำให้ตรวจพบข้อสอบที่มีเนื้อหาซ้ำซ้อนกันได้

ตัวอย่างที่ ๙. ผู้ป่วย febrile neutropenia มักมีไข้ขึ้นหลังจากได้รับยาเคมีบำบัดเป็นเวลากี่วัน

A. 2 - 4 วัน

B. 3 - 5 วัน

C. 5 - 7 วัน

D. 10 - 14 วัน

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

สรุป

ผู้นิพนธ์ได้รวบรวมข้อผิดพลาดในการสร้างข้อสอบปรนัยที่ผู้สอบอาจใช้เป็นแนวทางในการเลือกคำตอบที่ถูกได้โดยไม่ต้องอาศัยความรู้ทางการแพทย์ที่อาจารย์ต้องการประเมินผล โดยเรียบเรียงเป็นเจ็ดกลุ่มข้อผิดพลาดด้วยกัน ผู้อ่านทุกท่านพึงตระหนักว่าสิ่งเหล่านี้ไม่ใช่หลักการทางวิทยาศาสตร์ที่ชัดเจนดังกฎทางคณิตศาสตร์หรือฟิสิกส์ หากแต่เป็นการรวบรวมข้อสังเกตและคำแนะนำของผู้เชี่ยวชาญทางการวัดและประเมินผล จึงเป็นเพียงแนวทางเบื้องต้นในการพิจารณาตรวจสอบเนื้อหาของข้อสอบเท่านั้น การประยุกต์ใช้องค์ความรู้นี้คงต้องอาศัยศิลปะพอสมควรเพื่อที่จะได้ข้อสอบที่ดีสามารถวัดองค์ความรู้ทางการแพทย์ของนักศึกษาหรือแพทย์ประจำบ้านที่เข้าสอบได้ตามวัตถุประสงค์ของการสอบ

เอกสารอ้างอิง

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

15 March 2017

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.

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

Item Difficulty

- Proportion of examinees answering an item correctly (p)

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

C = number of examinees with a correct answer

I = number of examinees with incorrect answers

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

Item Discrimination

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

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

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

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

Cherdsak.ira@mahidol.ac.th

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

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

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

Applications

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

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.

Cherdsak.ira@mahidol.ac.th

การวิเคราะห์ข้อสอบปรนัย

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

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

ความรู้พื้นฐานเกี่ยวกับข้อสอบปรนัย

ก่อนที่จะกล่าวถึงรายละเอียดในการวิเคราะห์ข้อสอบ ผู้นิพนธ์ก็จะขอทบทวนความรู้พื้นฐานเกี่ยวกับข้อสอบปรนัยก่อน โดยทั่วไปข้อสอบปรนัยแต่ละข้อมีส่วนประกอบสำคัญ ๒ ส่วนด้วยกันคือ

๑. โจทย์ (stem) เป็นข้อมูลของโรค หรือภาวะหรือผู้ป่วยตามด้วยคำถาม หรือเว้นช่องว่างสำหรับเติมคำหรือข้อความที่เหมาะสมลงไป

๒. ตัวเลือก (options) คือคำ หรือข้อความที่ผู้ออกข้อสอบนำเสนอตามหลังจากโจทย์เพื่อให้ผู้สอบเลือกไปใช้ตอบคำถาม หรือเติมลงในช่องว่างในโจทย์

๒.๑ ตัวเลือกที่ถูกต้อง (correct option) เป็นคำตอบที่ถูกต้องมีเพียงตัวเลือกเดียวต่อข้อสอบข้อหนึ่ง

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

ทฤษฎีที่ใช้ในการวิเคราะห์ข้อสอบ

ทฤษฎีที่ใช้ในการวิเคราะห์ข้อสอบในปัจจุบันนั้นมี ๒ ทฤษฎีด้วยกัน ได้แก่ทฤษฎีการสอบแบบดั้งเดิม

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

เนื่องจากการวิเคราะห์ข้อสอบในวงการแพทยศาสตรศึกษาในประเทศไทยทั้งหมดในปัจจุบันยังใช้เทคนิคต่าง ๆ ตามทฤษฎีการสอบแบบดั้งเดิมอยู่ ดังนั้นผู้นิพนธ์จะขอกล่าวถึงเทคนิคการวิเคราะห์ข้อสอบตามทฤษฎีการสอบแบบดั้งเดิมเท่านั้น เพราะจะเป็นสิ่งที่อาจารย์แพทย์ทุกท่านจะได้พบและใช้งานเป็นประจำ

การวิเคราะห์ข้อสอบตามทฤษฎีการสอบแบบดั้งเดิม

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

๑. การวิเคราะห์ข้อสอบรายข้อ (item analysis)

การวิเคราะห์ข้อสอบแต่ละข้อให้อาจารย์พิจารณา ๓ ปัจจัย คือ

๑.๑ ความยากง่ายของข้อสอบ (item difficulty, p)

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

๑.๒ ความสามารถในการจำแนกผู้สอบตามระดับความสามารถ (item discrimination, r)

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

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

เมื่อ Mp = คะแนนรวมเฉลี่ยของผู้สอบที่ตอบข้อสอบถูก

Mq = คะแนนรวมเฉลี่ยของผู้สอบที่ตอบข้อสอบผิด

SD = ค่าเบี่ยงเบนมาตรฐาน (standard deviation) ของคะแนนสอบ

p = สัดส่วนของผู้สอบที่ตอบข้อสอบถูกต่อผู้สอบทั้งหมด

q = สัดส่วนของผู้สอบที่ตอบข้อสอบผิดต่อผู้สอบทั้งหมด

ค่า point-biserial correlation ที่คำนวณได้นี้มีค่าอยู่ในช่วง -๑ ถึง ๑ โดยค่าที่ติดลบหมายถึง ข้อสอบข้อนั้นผู้ที่ตอบถูกมักสอบได้คะแนนรวมต่ำ แต่ผู้ที่ตอบผิดมักสอบได้คะแนนรวมสูง ในทางตรงข้าม หากค่า point-biserial ยิ่งสูง แสดงถึงข้อสอบที่มีความสามารถในการแยกแยะดี ผู้ที่ตอบข้อสอบข้อนั้นถูกมักทำคะแนนรวมได้สูง ข้อสอบที่ดีควรมีค่า point-biserial สูงกว่า ๐.๒๐, ข้อสอบที่พอใช้ได้ควรมีค่า point-biserial อยู่ในช่วง ๐.๑ – ๐.๑๙, ข้อสอบที่มีค่า point-biserial ต่ำกว่า ๐.๑ เป็นข้อสอบที่ไม่สู้ดีนัก โดยเฉพาะอย่างยิ่งข้อสอบที่มีค่า point-biserial ต่ำกว่า ๐ ไม่ควรนำมาคิดคะแนน[๕-๖] (โดยทั่วไปแล้วข้อสอบที่มีค่า point-biserial ติดลบ ให้สงสัยว่าจะเฉลยผิด)

๑.๓ ประสิทธิภาพของตัวลวง (distractor functionality)

ตัวลวงที่มีประสิทธิภาพนั้นมีคุณสมบัติ ๒ ประการคือ[๕]

(๑) มีผู้สอบเลือกตัวลวงนั้นไม่ต่ำกว่าร้อยละ ๕ ของจำนวนผู้สอบทั้งหมด

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

ตัวลวงใดที่มีผู้สอบเลือกน้อย หรือลวงให้ผู้ที่มีความรู้ดีมาเลือกจัดเป็นตัวลวงที่ไม่ดี สมควรพิจารณาตัดทิ้งหรือปรับเปลี่ยน

๒. การวิเคราะห์ข้อสอบโดยรวม (test analysis)

การวิเคราะห์ข้อสอบโดยรวมเป็นการพิจารณาว่าเมื่อข้อสอบทั้งชุดทำงานร่วมกันแล้วผลสอบที่ได้ออกมาเป็นอย่างไร มีระดับความยากง่ายเป็นอย่างไร มีการกระจายตัวของคะแนนเป็นอย่างไร มีความน่าเชื่อถือของคะแนนสอบมากน้อยเพียงใด ดัชนีต่าง ๆ ที่ต้องพิจารณาได้แก่

๒.๑ ความเที่ยงตรงของคะแนนสอบ (internal consistency reliability)

การประเมินความเที่ยงตรงของคะแนนสอบเป็นการตรวจสอบว่าคะแนนที่ได้ออกมานั้นมีความน่าเชื่อถือเพียงใด เป็นการตอบคำถามว่าหากนำผู้สอบมาสอบใหม่ในสภาวการณ์เดิม ด้วยข้อสอบที่มีระดับความยากง่ายเท่าเดิม และผู้สอบมีความรู้เท่าเดิมไม่ได้ไปศึกษาหาความรู้เพิ่มเติม จะได้คะแนนสอบเท่าเดิมหรือไม่[๕,๑๑]

ดัชนีชี้วัดความเที่ยงตรงของคะแนนสอบที่นิยมใช้ในการรายงานผลสอบด้วยข้อสอบปรนัยคือ ค่าสัมประสิทธิ์ อัลฟ่า (Coefficient Alpha) ซึ่งสามารถคำนวณได้จากสูตร[๑๑]

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

เมื่อ α = สัมประสิทธิ์ อัลฟ่า (Coefficient Alpha)

n = จำนวนชุดย่อยของข้อสอบที่ทำการแบ่งออกเพื่อหาความเที่ยง

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

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

ค่าสัมประสิทธิ์อัลฟ่านี้มีค่าอยู่ในช่วง ๐ – ๑ ค่าต่ำแสดงว่าคะแนนที่ได้มีความเชื่อถือได้น้อย ไม่แตกต่างไปจากการเดาสุ่ม ค่าสูงแสดงว่าคะแนนที่ได้นั้นมีความน่าเชื่อถือมาก หากทำการทดสอบซ้ำคะแนนที่ได้ก็จะใกล้เคียงเดิม โดยทั่วไประดับของความเที่ยงตรง

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

ประเด็นสำคัญที่ต้องพิจารณาคือเมื่อได้คะแนนสอบที่มีค่าสัมประสิทธิ์ อัลฟาต่ำ จะต้องดำเนินการอย่างไรเพื่อพัฒนาให้การสอบครั้งต่อไปไม่ประสบปัญหาเรื่องความไม่น่าเชื่อถือของคะแนนสอบอีก ปัจจัยหลักที่จะช่วยเพิ่มความเที่ยงตรงของคะแนนสอบปรนัยมี ๓ ปัจจัยด้วยกัน[๙] คือ

(๑) เพิ่มจำนวนข้อสอบให้มากขึ้น ยิ่งมีข้อสอบมากข้อคะแนนที่ได้ก็จะมีความเที่ยงตรงเพิ่มมากขึ้น

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

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

๒.๒ การกระจายตัวของคะแนน และคะแนนเฉลี่ย (standard deviation and mean score)

การตรวจดูลักษณะพื้นฐานของคะแนนสอบนี้จะช่วยบอกได้คร่าว ๆ ว่าการเรียนการสอนมีประสิทธิภาพเพียงใด หากอาจารย์สอนได้ดี นักเรียนทั้งชั้นเรียนเข้าใจเนื้อหาดี คะแนนสอบที่ได้ออกมาก็ควรจะไม่กระจายตัวมากนัก (คะแนนเกาะกลุ่มกัน) และคะแนนเฉลี่ยก็ควรจะค่อนข้างสูงเมื่อเทียบกับนักเรียนรุ่นอื่น ๆ หากคะแนนสอบของนักเรียนมีการกระจายตัวมากผิดปกติ แสดงว่าอาจมีปัญหาบางประการในการเรียนการสอนทำให้นักเรียนบางคนมีความรู้ความเข้าใจดี แต่มีนักเรียนบางกลุ่มที่ไม่ค่อยรู้เรื่อง[๙]

๒.๓ ค่าความยากง่ายเฉลี่ยของข้อสอบ (average difficulty)

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

๒.๔ ค่าความสามารถในการแยกแยะผู้สอบเฉลี่ย (average discrimination)

การนำค่า point-biserial correlation ของข้อสอบทั้งชุดมาหาค่าเฉลี่ย เป็นการบอกคร่าว ๆ ว่าโดยรวมแล้วข้อสอบชุดนี้มีความสามารถในการแยกแยะผู้สอบตามระดับความสามารถเพียงใด ยิ่งได้ค่าสูงก็ยิ่งดี แต่มีข้อควรระวังในการแปลผลในกรณีที่การเรียนการสอนเป็นไปได้ดี และผู้สอบทั้งหมด หรือเกือบทั้งหมดทำคะแนนได้สูง ค่า point-biserial correlation เฉลี่ยของข้อสอบทั้งชุดจะไม่สูงแต่ไม่ได้แปลว่าข้อสอบที่ใช้มีคุณภาพไม่ดี[๙]

การนำผลการวิเคราะห์ข้อสอบไปใช้

ผลการวิเคราะห์ข้อสอบด้วยดัชนีชี้วัดต่าง ๆ ดังกล่าวข้างต้นสามารถนำไปใช้ประโยชน์ได้หลายประการ เช่น

๑. ใช้เป็นประโยชน์ในการปรับแก้คะแนนสอบ

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

๒. ใช้เป็นประโยชน์ในการปรับปรุงคุณภาพข้อสอบ

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

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

๓. ใช้เป็นประโยชน์ในการบริหารคลังข้อสอบ

ข้อสอบแต่ละข้อนั้นได้มาด้วยความยากลำบาก อาจารย์แต่ละท่านต้องใช้เวลาและความคิดอย่างมากเพื่อพัฒนาข้อสอบที่ดีขึ้นมาใช้ ดังนั้นเมื่อนำข้อสอบมาใช้แล้วผลการวิเคราะห์ข้อสอบแสดงว่าข้อสอบข้อใดเป็นข้อสอบที่ดี มีระดับความยากง่ายเหมาะสม มีความสามารถในการจำแนกผู้สอบที่ดีก็ควรจะพิจารณาเลือกเก็บข้อสอบดังกล่าวไว้ในคลังข้อสอบเพื่อที่จะได้นำกลับมาใช้ใหม่ในอนาคต ในการเก็บข้อสอบเข้าในคลังข้อสอบก็ต้องมีการแนบข้อมูลเกี่ยวกับประวัติการใช้งานและผลการวิเคราะห์ข้อสอบในแต่ละครั้งไว้คู่กันด้วย เพื่อที่จะได้เป็นประโยชน์ในการเลือกข้อสอบมาใช้งาน หากอาจารย์ต้องการข้อสอบที่มีระดับความยากง่าย หรือความสามารถในการจำแนกผู้สอบมากน้อยเพียงใดจะได้ดึงเอาข้อสอบที่มีคุณลักษณะตามต้องการออกมาใช้ได้ตามต้องการ

๔. ใช้เป็นประโยชน์ในการพัฒนาคุณภาพการสอน

การพิจารณาผลการวิเคราะห์ข้อสอบโดยละเอียดในหัวข้อที่อาจารย์ท่านใดท่านหนึ่งรับผิดชอบ

ในการสอนนักเรียนหรือแพทย์ประจำบ้านอยู่นั้นจะทำให้ได้ข้อมูลที่เป็นประโยชน์ในการพัฒนาการเรียนการสอนได้ กล่าวคืออาจารย์สามารถตรวจสอบดูได้ว่านักเรียนหรือแพทย์ประจำบ้านมีความเข้าใจที่ถูกต้องในเรื่องดังกล่าวหรือไม่ ประเด็นใดที่มีผู้เข้าใจผิดอยู่มากก็สมควรที่อาจารย์จะทำการเน้นย้ำในบรรดานักเรียนหรือแพทย์ประจำบ้านในการสอนครั้งต่อ ๆ ไปเพื่อแก้ไขความเข้าใจผิดดังกล่าว ประเด็นใดที่นักเรียนหรือแพทย์ประจำบ้านมีความเข้าใจดีมากอยู่แล้ว อาจารย์อาจไม่ต้องใช้เวลามากนักในการสอนเรื่องดังกล่าว แต่เอาเวลามาใช้สอนในเรื่องที่นักเรียนหรือแพทย์ประจำบ้านยังไม่ค่อยเข้าใจให้มากขึ้นได้

ข้อจำกัดของการวิเคราะห์ข้อสอบ

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

พื้นฐานสำคัญที่เป็นข้อจำกัดของผลการวิเคราะห์ข้อสอบด้วยทฤษฎีการสอบแบบดั้งเดิมคือค่าต่าง ๆ ที่ได้มาจากการวิเคราะห์นั้นขึ้นอยู่กับกลุ่มตัวอย่างที่ใช้ในการเก็บข้อมูล[๑๓,๑๕] หากได้ข้อมูลมาจากกลุ่มตัวอย่างที่มีขนาดใหญ่พอและมีการกระจายตัวของระดับความสามารถของผู้สอบที่เหมาะสม ค่าต่าง ๆ ที่ได้ (p, r, coefficient alpha) จะค่อนข้างเที่ยงตรง ปัญหาที่สำคัญในการวิเคราะห์ข้อสอบในโรงเรียนแพทย์คือการสอบจำนวนมากจัดในนักศึกษากลุ่มเล็ก และนักศึกษาแต่ละกลุ่มก็มีการกระจายตัวของระดับความสามารถแตกต่างกัน นักศึกษาบางกลุ่มมีความสามารถสูงกว่านักศึกษากลุ่มอื่น ดังนั้นผลการวิเคราะห์ข้อสอบไม่ว่าจะเป็นค่า p, r, coefficient alpha, mean, หรือ standard deviation อาจจะเปลี่ยนแปลงไปในแต่ละกลุ่มของนักศึกษา ดังนั้นการนำผลการวิเคราะห์ข้อสอบไปใช้ในทางปฏิบัติจึงมีข้อควรระวังดังต่อไปนี้

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

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

ค่าสัมประสิทธิ์อัลฟ่า เป็นค่าที่มีความเฉพาะเจาะจงกับการสอบของนักเรียนกลุ่มใดกลุ่มหนึ่งเท่านั้น หาใช่เป็นคุณสมบัติติดตัวข้อสอบแต่ละข้อไม่ หากข้อสอบชุดหนึ่งทำการสอบกับนักเรียนกลุ่มหนึ่งแล้วพบว่าคะแนนสอบที่ได้มานั้นมีค่าสัมประสิทธิ์อัลฟ่าสูงในระดับที่ต้องการก็ไม่ได้เป็นตัวรับประกันว่าหากนำข้อสอบชุดเดิมนั้นไปทำการสอบกับนักเรียนกลุ่มอื่นจะได้ค่าสัมประสิทธิ์อัลฟ่าที่สูงเช่นเดียวกัน นอกจากนี้ค่าสัมประสิทธิ์อัลฟ่าที่สูงไม่ได้เป็นตัวบอกถึงคุณภาพของข้อสอบรายข้อแต่อย่างใด

ค่าสัมประสิทธิ์อัลฟ่าที่สูงช่วยบอกแค่เพียงว่า

คะแนนสอบในข้อสอบข้อหนึ่งมีความผันแปรไปในทิศทางเดียวกันกับคะแนนสอบในข้อสอบข้ออื่นในการสอบชุดเดียวกัน นั่นคือในข้อสอบชุดที่มีค่าสัมประสิทธิ์อัลฟ่าสูงก็อาจประกอบไปด้วยข้อสอบที่ดี และข้อสอบที่ไม่ดีรวมกันอยู่ ต้องไปตรวจสอบดัชนีชี้วัดคุณภาพของข้อสอบตัวอื่น ๆ ในแต่ละข้ออีกครั้ง

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

เอกสารอ้างอิง

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

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

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

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

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

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

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

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

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

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

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

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

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

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

โปรแกรมวิเคราะห์ข้อสอบ

รุ่น 2.0

การสอบ : SIID 521 (Basic Sciences)

วันที่ : 22 ธันวาคม 2555

จำนวนข้อสอบ = 120

จำนวนผู้เข้าสอบ = 244

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

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

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

==========

SCORE STATISTICS

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

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

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

DIFFICULTY INDEX (p value)

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

DISCRIMINATION INDEX (D or r value)

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

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

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

Item Analysis and Option Analysis

Faculty of Medicine Siriraj Hospital
Mahidol University

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Item Analysis and Option Analysis

Faculty of Medicine Siriraj Hospital
Mahidol University

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Item Analysis and Option Analysis

Faculty of Medicine Siriraj Hospital
Mahidol University

No. : 37	p Value : 0.29								r_{pbi} : -0.02	
A		B		C		D		* E		
r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	
-0.05	2.05	0.22	52.05	-0.14	7.38	-0.20	9.84	-0.02	28.69	

No. : 38	p Value : 0.75								r_{pbi} : 0.11	
* A		B		C		D		E		
r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	
0.11	74.59	-0.11	22.95	-0.14	0.82	0.08	0.82	0.08	0.82	

No. : 39	p Value : 0.51								r_{pbi} : 0.23	
A		B		* C		D		E		
r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	
-0.02	10.25	-0.21	27.46	0.23	51.23	-0.07	9.02	0.09	1.64	

No. : 40	p Value : 0.21								r_{pbi} : 0.13	
A		* B		C		D		E		
r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	
0.00	40.57	0.13	20.90	0.00	4.51	0.07	17.62	-0.21	16.39	

No. : 41	p Value : 0.42								r_{pbi} : -0.03	
A		B		C		* D		E		
r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	
0.02	7.38	0.07	43.03	-0.02	0.41	-0.03	41.80	-0.10	7.38	

No. : 42	p Value : 0.79								r_{pbi} : 0.33	
A		* B		C		D		E		
r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	
-0.22	5.33	0.33	79.10	-0.20	4.92	-0.02	2.87	-0.15	7.79	

No. : 43	p Value : 0.81								r_{pbi} : 0.37	
* A		B		C		D		E		
r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	
0.37	80.74	-0.33	14.75	0.01	0.82	-0.14	2.05	-0.07	1.64	

No. : 44	p Value : 0.56								r_{pbi} : 0.34	
A		B		* C		D		E		
r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	
-0.14	1.64	-0.18	6.56	0.34	55.74	-0.22	20.08	-0.05	15.98	

No. : 45	p Value : 0.86								r_{pbi} : 0.39	
A		B		C		D		* E		
r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	
-0.16	2.05	-0.11	0.82	-0.04	1.23	-0.33	9.84	0.39	86.07	

No. : 46	p Value : 0.81								r_{pbi} : 0.31	
A		* B		C		D		E		
r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	
-0.19	10.66	0.31	80.74	-0.09	2.87	-0.15	1.64	-0.15	4.10	

No. : 47	p Value : 0.93								r_{pbi} : 0.26	
A		* B		C		D		E		
r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	
-0.14	2.46	0.26	93.44	-0.01	0.82	-0.17	1.64	-0.15	1.64	

No. : 48	p Value : 0.07								r_{pbi} : -0.20	
A		B		C		* D		E		
r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	
-0.20	12.70	-0.08	4.51	-0.18	2.87	-0.20	6.56	0.37	73.36	

No. : 49	p Value : 0.95								r_{pbi} : 0.21	
A		B		C		* D		E		
r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	
0.00	0.00	0.00	0.00	-0.21	4.92	0.21	95.08	0.00	0.00	

No. : 50	p Value : 0.83								r_{pbi} : 0.24	
A		B		* C		D		E		
r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	
0.00	0.00	0.00	0.00	0.24	83.20	-0.23	15.98	-0.09	0.82	

No. : 51	p Value : 0.76								r_{pbi} : 0.26	
* A		B		C		D		E		
r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	
0.26	76.23	-0.14	2.87	-0.04	2.46	0.07	0.41	-0.23	18.03	

No. : 52	p Value : 0.70								r_{pbi} : 0.24	
A		B		C		* D		E		
r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	
-0.15	0.82	-0.21	11.89	0.01	12.70	0.25	70.08	-0.16	4.51	

No. : 53	p Value : 0.51								r_{pbi} : 0.31	
A		* B		C		D		E		
r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	
0.02	4.51	0.31	50.82	-0.07	2.05	-0.07	2.87	-0.28	39.75	

No. : 54	p Value : 0.37								r_{pbi} : 0.28	
A		* B		C		D		E		
r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	
-0.07	9.43	0.28	36.89	-0.19	13.52	-0.09	16.80	-0.04	23.36	

Item Analysis and Option Analysis

Faculty of Medicine Siriraj Hospital
Mahidol University

No. : 55	p Value : 0.71					r_{pbi} : 0.25			
A		B		C		* D		E	
r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%
-0.18	2.87	-0.20	14.75	-0.08	5.74	0.25	70.90	0.01	5.74

No. : 56	p Value : 0.81					r_{pbi} : 0.29			
A		* B		C		D		E	
r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%
-0.02	1.23	0.29	81.15	-0.15	7.38	-0.10	4.92	-0.22	5.33

No. : 57	p Value : 0.26					r_{pbi} : 0.19			
A		B		C		* D		E	
r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%
-0.08	6.15	-0.17	29.51	-0.01	15.57	0.19	26.23	0.03	22.54

No. : 58	p Value : 0.66					r_{pbi} : 0.29			
A		B		C		* D		E	
r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%
-0.16	25.00	-0.14	2.46	-0.22	0.41	0.29	65.98	-0.14	6.15

No. : 59	p Value : 0.73					r_{pbi} : 0.36			
A		B		C		* D		E	
r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%
-0.13	0.82	-0.25	19.67	-0.26	5.33	0.36	73.36	0.10	0.82

No. : 60	p Value : 0.93					r_{pbi} : 0.28			
A		B		C		D		* E	
r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%
0.00	0.00	-0.13	4.10	-0.27	2.87	-0.03	0.41	0.28	92.62

No. : 61	p Value : 0.89					r_{pbi} : 0.26			
A		B		C		D		* E	
r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%
0.05	0.41	-0.30	2.46	-0.13	5.74	-0.06	2.46	0.26	88.93

No. : 62	p Value : 0.89					r_{pbi} : 0.38			
A		B		C		* D		E	
r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%
-0.32	7.38	-0.09	0.82	-0.17	3.28	0.38	88.52	0.00	0.00

No. : 63	p Value : 0.69					r_{pbi} : 0.05			
A		B		C		* D		E	
r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%
0.00	0.00	-0.12	1.64	-0.02	29.51	0.05	68.85	0.00	0.00

No. : 64	p Value : 0.81					r_{pbi} : 0.20			
A		B		* C		D		E	
r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%
-0.09	0.82	0.05	2.46	0.20	80.74	-0.16	11.89	-0.10	3.69

No. : 65	p Value : 0.68					r_{pbi} : 0.10			
A		B		* C		D		E	
r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%
-0.06	9.43	-0.15	1.64	0.10	68.44	-0.04	1.23	-0.01	19.26

No. : 66	p Value : 0.55					r_{pbi} : 0.32			
A		B		* C		D		E	
r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%
-0.22	23.36	-0.08	11.48	0.32	54.92	-0.11	6.15	-0.07	4.10

No. : 67	p Value : 0.45					r_{pbi} : 0.29			
A		B		C		* D		E	
r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%
-0.20	26.64	-0.07	17.62	-0.05	1.23	0.29	45.49	-0.06	8.61

No. : 68	p Value : 0.28					r_{pbi} : -0.03			
A		B		* C		D		E	
r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%
0.02	14.34	0.07	1.64	-0.03	27.87	0.06	10.25	-0.04	45.90

No. : 69	p Value : 0.39					r_{pbi} : 0.37			
A		B		C		* D		E	
r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%
-0.05	23.77	-0.07	13.93	-0.22	0.41	0.37	38.93	-0.28	22.95

No. : 70	p Value : 0.25					r_{pbi} : 0.13			
A		* B		C		D		E	
r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%
-0.02	7.79	0.13	24.59	-0.10	1.64	0.06	10.66	-0.10	54.92

No. : 71	p Value : 0.80					r_{pbi} : 0.09			
* A		B		C		D		E	
r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%
0.09	80.33	-0.03	1.64	-0.13	3.28	0.00	5.74	-0.03	9.02

No. : 72	p Value : 0.65					r_{pbi} : 0.37			
A		B		C		D		* E	
r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%	r_{pbi}	%
-0.25	6.97	-0.05	6.56	-0.23	20.08	-0.05	1.23	0.37	65.16

16 March 2017

Constructed Response Items

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

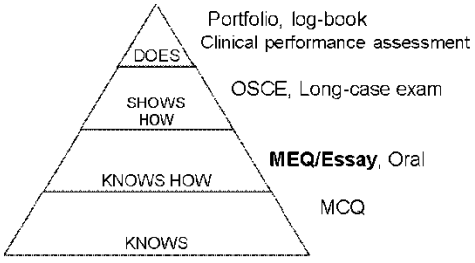
1

Written Tests

Two major types of written test forms
1. Selected Response items
2. Constructed response items

2

Assessment approach



Miller's pyramid

3

Written examination

- Level I: Recall, Recognition
  - ทดสอบความจำ
- Level II: Comprehension, Interpretation
  - ทดสอบความเข้าใจ สรุปข้อมูล การแปลผลต่างๆ
- Level III: Application, Problem solving
  - วิเคราะห์ปัญหา เพื่อการวินิจฉัยโรค/ภาวะ
  - การตัดสินใจในการแก้ปัญหา (การรักษา)

Use of an educational taxonomy for evaluation of cognitive performance. J Med Educ 1981 Feb;56(2):115-21

4

Comparison

	Selected Response	Constructed Response
Measured construct	Concrete knowledge, basic interpretation, some applications	Complex cognitive ability: problem solving, interpretation, decision making
Item construction	Simple	Complex
Cost of scoring	Low	Expensive
Type of scoring	Objective	Subjective
Rater effects	No effect	Significant factor
Reliability	High	Low

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

5

Limitations of Selected Response Items

- Cueing and guessing correct answers
- Difficulty in developing good items
- Testing of trivial content
- Limited ability to assess higher level of cognitive learning

Assessment of knowledge with written test forms. In International handbook of research in medical Education, part 2. Dordrecht: Kluwer, 2002, p. 647-672.

6

suprapath.son@mahidol.ac.th

1

Constructed Response Items

A variety of written formats in which examinees is required to *create answers* spontaneously in response to questions

- **Traditional essay questions**
  - Long essay
  - Short essay
- **Modified essay questions**
  - Standard modified essay questions (MEQ)
  - Patient management problem (PMP)
  - Key features problem (KFP)
  - Short Answer question (SAQ)

7

Objectives

เมื่อสิ้นสุดการบรรยายและการร่วมกิจกรรม อาจารย์ผู้เข้าร่วมอบรมสามารถ

- อธิบายข้อดีและข้อจำกัดของข้อสอบชนิด constructed response items
- บอกขั้นตอนที่สำคัญในการสร้างข้อสอบ modified essay questions ได้
- ร่วมในกระบวนการพัฒนาข้อสอบ modified essay questions สำหรับนักศึกษาระดับคลินิก

8

Constructed response items: Strengths

- Examinees 'responses are non-cued: more authentic
- Able to measure higher-order cognitive tasks: application, analysis, synthesis, and evaluation
- Motivation for clinical learning

9

Constructed response items: Limitations

- Difficult to develop and score
- Inefficient exam format
- Expensive
- Subjectivity
- Low reliability
- Construct underrepresentation

10

Traditional essay questions

- Long essay examinations
  - An exam is consist of a few open-ended essay questions, each requires lengthy written responses from examinees
- Short essay examinations
  - An exam is consist of many open-ended essay questions, each requires short written answer consisting of a sentence or two

11

Comparison

	Long Essay	Short Essay
Content coverage	Narrow	Broad
Item development	Easy	Difficult
Scoring guideline development	Very difficult	Easier
Students' answers	Infinite possibilities	More focused scope
Reliability	Very low	Low
Time used	More	Less
Good use	Assessment of complex cognitive abilities: analysis, synthesis, evaluation, and presentation of ideas	Assessment of simplified, structured problems with limited answers

12

suprapath.son@mahidol.ac.th

Modified Essay Question

- การประยุกต์ให้เสมือนการแก้ปัญหาผู้ป่วยในชีวิตจริง
- การแก้ปัญหาของผู้ป่วยรายหนึ่งๆประกอบด้วยหลายขั้นตอน
  - จะไม่มีข้อมูลทั้งหมดตั้งแต่เริ่มเห็นผู้ป่วย
  - ต้องคอยๆสืบค้นหาข้อมูลเพิ่มเติมและวิเคราะห์ ตัดสินใจแก้ปัญหาไปทีละขั้นตอน
  - เมื่อทำแต่ละขั้นตอนแล้ว ไม่สามารถย้อนกลับไปแก้ไขสิ่งที่ได้ทำไปก่อนหน้านี้ได้

13

Clinical Problem Solving Methods

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

14

Forward Reasoning
(Data driven process)

Complete data gathering
history taking, physical exam, lab findings, radiological studies, etc.

⇩

Generate diagnosis and differential diagnoses

⇩

Plan the treatment

15

Backward reasoning
(Hypothesis Driven Process)

Chief complaint / Condition

⇩

Identification of problem (s)

⇩

Formulation of hypothesis (es)

⇩

Information gathering

⇩

Testing the hypothesis (es)
(collect more information)

⇩

Re-rank hypothesis (Tentative diagnosis)

⇩

Intervention & Management

16

MEQ Process

A brief clinical situation (scenario, vignette)

⇩

Hypothesis generation / Differential diagnosis

⇩

(Focused) Data gathering (Hx, PE)

⇩

Hypothesis refinement

⇩

Management/Intervention

⇩

Incidental question related to clinical/basic sciences
(จะมีหรือไม่มีก็ได้)

17

Standard Modified Essay Questions

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

รัชดาภรณ์ลิขิตกุล การประเมินความรู้ในการแก้ปัญหาผู้ป่วยทางคลินิก ศิริราชสาร 2564; 47(2) : 125 – 134

18

Standard MEQ

- **Chief complaint (A brief scenario)**
  - ผู้ป่วย: เพศ อายุ และภูมิหลังที่จำเป็น
  - ปัญหา: สั้นๆ แต่รัดกุม เพียงพอที่จะนำมาวิเคราะห์ และตั้งสมมุติฐานกว้างๆ ได้

    ควรเกี่ยวข้องกับหลายๆ สาขาวิชา
- **A question on differential diagnosis (Hypothesis generation)**
  - ควรตั้งคำถามให้รัดเจน และจำเพาะ

19

Standard MEQ

- **Questions to collect additional information (Data gathering)**
  - คำถามเกี่ยวกับข้อมูลทางคลินิก **(Hx & PE)** เพื่อมาสนับสนุน / คัดค้าน สมมุติฐาน ที่ตั้งไว้ใน

    (focused data gathering)
- **Additional clinical information**
  - อาจให้ข้อมูลทั้งหมด หรือ ให้ข้อมูลบางส่วน แล้วใช้คำถามต่อ ว่ายังต้องการข้อมูลอะไรอีกบ้าง

20

Standard MEQ

- **Differential diagnosis (Hypothesis refinement)**
  - คำถามเกี่ยวกับการวินิจฉัยโรคที่น่าจะเป็น โดยอาศัยข้อมูลทั้งหมดที่ให้
- **Interpretation of laboratory findings**
  - คำถามการแปลข้อมูลผลการตรวจทางห้องปฏิบัติการ ภาพรังสี คลื่นไฟฟ้าหัวใจ เป็นต้น

21

Standard MEQ

- **Management**
  - คำถามการรักษาจำเพาะ การรักษาตามอาการ / คำสั่งการรักษา
  - การป้องกัน ส่งเสริมสุขภาพ
- **Exploring knowledge (optional)**
  - คำถามทดสอบความรู้เกี่ยวกับวิทยาศาสตร์การแพทย์พื้นฐาน

22

Modified Essay Question

Advantages

- Construct responses
- Mimic actual clinical problem solving
- Focus on higher order cognitive abilities

23

Modified Essay Question

Limitations

- Construct underrepresentation
- Difficult to develop
- Unexpected responses
- Subjective scoring
- Cannot assess affective or psychomotor abilities

24

Key Features Problem

- A constructed response question focusing on clinical decision making skills
- Elicit examinees' responses concerning only the critical steps in the resolution of each problem (the problem's key features)
- Allow for more cases, items for testing a broader content domain
- Responses can be selected or constructed

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

25

Key Features Problem

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

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

26

Short Answer Question (SAQ)

ข้อสอบชนิดบรรยายที่มีลักษณะดังนี้

- มีโจทย์ผู้ป่วยสมมติ
- ถามคำถาม (2-3 ข้อ) ที่เกี่ยวข้องกับโจทย์ผู้ป่วย
- คำถามที่ถามต้องการคำตอบเป็นคำหรือวลีสั้น ๆ ที่ตรงประเด็นเท่านั้น

27

Developing an MEQ

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

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

28

Assembling Problem-Writing Groups

Item writers: background and clinical expertise are pertinent to the context of the examination

- Ensure that the problems used are well grounded in practice and represent a wide range of real-life practice.
- A group of writers help review the content.

29

Select A Problem

- Refer to test specification table
- Select an appropriate clinical problem
  1. ปัญหาที่พบบ่อย และจำลองมาจากผู้ป่วยจริง
  2. ปัญหาหรืออาการสำคัญที่ยังไม่สามารถจำแนกสาเหตุได้แน่นอน
  3. ปัญหาที่นักศึกษาหรือแพทย์ประจำบ้านมักพลาดบ่อย
  4. ปัญหาที่เกี่ยวข้องกับหลายระบบ เช่น ผู้ป่วยมีปัญหาระบบ G ควบกับ nutrition และ/หรือ electrolyte imbalances เป็นต้น

30

Select A Problem (cont.)

- Select an appropriate clinical problem (cont.)
  5. ปัญหาที่สามารถประเมินทักษะการแก้ปัญหาและการตัดสินใจ
     - **Life threatening หรือ emergency situation**
     - **Diagnosis: relevant history, physical examination or investigations**
     - **Subsequent or definite management**
     - **Preventive care, health promotion, rehabilitation**

31

Defining A Key Feature

- Ask a problem writer
  - ปัญหาที่สำคัญที่สุดในการจัดการกับผู้ป่วยที่นำเสนอ
  - ขั้นตอนสำคัญที่ขาดไม่ได้ในการรักษาผู้ป่วย
- remark
  - key features ไม่จำเป็นต้องจำกัด.เฉพาะ biomedical บาง สถานการณ์อาจเป็นเรื่อง ethical, medicolegal, prevention
  - ปรึกษาหารือในกลุ่มผู้เขียนโจทย์ จนได้ consensus ว่าขั้นตอนใดจัดว่า essential และ critical

32

Defining A Key Feature (cont.)

- Typical decisions or actions tested in KFP
  - ประวัติเพิ่มเติมที่สำคัญ
  - การตรวจร่างกายที่สำคัญที่ต้องมองหา หรือตรวจเพิ่มเติม
  - การวินิจฉัยโรค หรือ วินิจฉัยแยกโรค
  - การสืบค้นเพิ่มเติมเพื่อ confirm หรือ exclude การวินิจฉัย
  - การรักษาที่เฉพาะเจาะจงกับโรค

33

Defining A Key Feature (cont.)

- Qualifiers: คำคุณศัพท์ที่บ่งบอกความสำคัญของการตัดสินใจ
  - Immediate (สิ่งที่ต้องทำ)ทันที
  - Initial (สิ่งที่ต้องทำ) เบื้องต้น
  - Longterm (สิ่งที่ต้องทำ) ในระยะยาว
  - Definitive (การรักษา การดูแล ...) ที่จำเพาะ
  - Urgent ฉุกเฉิน เร่งด่วน
  - Most important สำคัญที่สุด
  - Most likely น่าจะเป็นไปได้มากที่สุด
  - Must not miss (สิ่งที่) พลาดไม่ได้ ห้ามพลาด ฯลฯ

34

From A Problem to A Case

Following a decision of key features, the problem writers select one case scenario:

- Age, gender
- Setting of the encounter
- KFP on diagnosis: brief case
- KFP on management: longer case and includes laboratory information

35

Writing the Questions

- Write the questions that test the defined key features
- Most case scenario are followed by two or three questions, each question test one key feature
- The number of answers may vary from one to ten, typically 3-5 answers

36

Selecting Question Formats

- Two alternatives
  - **(1) Write-in (WI) format**: write a very short note or single words
  - **(2) Short menu (SM) format**: select from a list up to 25 items

Medical Council of Canada and Royal Australian College of General Practitioners suggested WI format as a more effective one

37

Specify the number of required answers

ระบุให้ชัดเจนในโจทย์ว่าจะให้ทำอะไร อย่างไร ให้บอกชื่อโรคกี่ชื่อ เช่น

- จงบอกชื่อโรคที่ผู้ป่วยรายนี้น่าจะเป็นมากที่สุด 1 โรค
- จงบอกสิ่งตรวจพบจากการตรวจร่างกายที่สำคัญที่จะช่วยในการยืนยันการวินิจฉัยโรค มา 3 ประการ
- จงเขียนคำสั่งการรักษาสำหรับผู้ป่วยรายนี้ในใบคำสั่งการรักษาที่จัดให้

38

Preparing Scoring Keys

- Only one acceptable answer
  - Correct diagnosis
- Multiple acceptable answers
  - Differential diagnosis
- Partial credit system
  - Complete answer
  - Incomplete answer

39

Preparing Scoring Keys (cont.)

- Penalty
  - Absence of "must have" answers
    - Give a score of "0" despite the presence of other less important answers
  - Presence of "unnecessary" investigations or treatment
    - Two options:
      - negative score (but not cross items)
      - no score (C)
  - Harmful treatment
    - negative score (but not cross items)

40

Time

- อาจารย์ผู้ออกข้อสอบ ควรทดลองตอบคำถามด้วยตนเองและจับเวลา หรือ ให้เพื่ออาจารย์ทดลองทำข้อสอบ
- เวลาที่นักศึกษาใช้ในการตอบ จะมากกว่าเวลาที่อาจารย์ใช้ในการตอบคำถามนั้น ๆ ประมาณ 30-50%
- หากข้อมูลที่ให้เพิ่มเติมในแต่ละหน้ามีความยาวมาก ต้องกำหนดเวลาให้เพียงพอสำหรับอ่านและแปลข้อมูล

41

Validation and References

- Validation
  - Pilot the problem with colleagues new to the problem => discussion, revision
- References
  - Useful, especially in the field of rapidly developing intervention and discovery

42

ตารางสรุปข้อสอบ MEQ ปีการศึกษา.....

สถานีที่.....จำนวนข้อสอบทั้งหมด.....ข้อ เวลาสอบรวม ..... นาที

ข้อที่	เรื่องที่ออกข้อสอบ	จำนวนข้อย่อย	Physician tasks / Competencies										
			Problem Identifi -cation	Hypothesis generation	Data Gathering	Data Interpretation	Clinical Reasoning	Patient Management	Patient Education	Ethical analysis	Evidence -based	BasicKnow -ledge	อื่นๆ
1.													
2.													
3.													
4.													
5.													
6.													
7.													

ข้อสอบ Modified Essay Questions (MEQ)

นักศึกษาแพทย์ชั้นปี..... ปีการศึกษา.....

สถาบัน คณะแพทยศาสตร์ศิริราชพยาบาล

รายวิชา

อาจารย์ผู้ออกข้อสอบ

Problem / Topic

Objectives 1.

2.

3.

4.

วันที่ออกข้อสอบ

จำนวนคำถาม คำถาม

เวลาประมาณ นาที

คะแนนเต็ม 100 คะแนน

Physician Tasks ☑	**คะแนนเต็ม**
☐ Health promotion and maintenance	
☐ Mechanism of diseases	
☐ Data Gathering (Hx & PE)	
☐ Data Gathering (Investigation)	
☐ Hypothesis Generation (Differential diagnosis)	
☐ Hypothesis Refinement (Diagnosis)	
☐ Emergency management	
☐ Acute management	
☐ Long term management	
☐ Counseling education	
☐ Basic knowledge	
คะแนนเต็ม	**100**

เกณฑ์ผ่าน คะแนน

โจทย์ข้อสอบ

.....

.....

.....

.....

คำถามที่ 1. ..... (คะแนน)

(นาที)

ข้อมูลเพิ่มเติม.....

คำถามที่ 2. ..... (คะแนน)

(นาที)

ข้อมูลเพิ่มเติม.....

คำถามที่ 3. ..... (คะแนน)

(นาที)

ข้อมูลเพิ่มเติม.....

คำถามที่ 4. ..... (คะแนน)

(นาที)

เฉลยข้อสอบ

คำถามที่ 1. (คะแนน)

1..... .....คะแนน

2..... .....คะแนน

3..... .....คะแนน

4..... .....คะแนน

เกณฑ์ผ่าน.....คะแนน

คำถามที่ 2. (คะแนน)

1..... .....คะแนน

2..... .....คะแนน

3..... .....คะแนน

4..... .....คะแนน

เกณฑ์ผ่าน.....คะแนน

คำถามที่ 3. (คะแนน)

1..... .....คะแนน

2..... .....คะแนน

3..... .....คะแนน

4..... .....คะแนน

เกณฑ์ผ่าน.....คะแนน

คำถามที่ 4. (คะแนน)

1..... .....คะแนน

2..... .....คะแนน

3..... .....คะแนน

4..... .....คะแนน

เกณฑ์ผ่าน.....คะแนน

the metric of medical education

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

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

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

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

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

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

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

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

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

INTRODUCTION

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

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

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

Overview

What is already known on this subject

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

What this study adds

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

Suggestions for further research

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

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

A NEW APPROACH

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

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

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

SAMPLE KEY FEATURES PROBLEM: —DIARRHOEA

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

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

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

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

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

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

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

Figure 1 A sample key features problem.

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

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

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

DEVELOPING KEY FEATURES PROBLEMS

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

Assembling problem-writing groups

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

Selecting a problem, defining its key features

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

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

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

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

Figure 2 Critical clinical decisions or actions tested in KFPs.

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

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

Figure 3 Common qualifiers in key features.

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

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

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

Figure 4 Typical elements in KFP clinical scenarios.

Writing the questions

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

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

Selecting question formats

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

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

Specifying the number of required answers

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

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

PREPARING SCORING KEYS

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

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

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

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

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

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

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

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

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

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

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

VALIDATION AND REFERENCES

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

COMPUTERISED PRESENTATION OF KFP FORMATS

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

QUALITY ASSURANCE ISSUES IN ITEM DEVELOPMENT

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

STANDARD SETTING OF KFP FORMATS

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

CONCLUSION

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

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

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

REFERENCES

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

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

BMC Medical Education

BioMed Central

Research article

Open Access

Assessment of higher order cognitive skills in undergraduate education: modified essay or multiple choice questions? Research paper

Edward J Palmer*[1,2] and Peter G Devitt[2]

Address: [1]Centre for Learning and Professional Development, University of Adelaide, Adelaide, Australia and [2]Dept of Surgery, University of Adelaide, Adelaide, Australia

Email: Edward J Palmer* - edward.palmer@adelaide.edu.au; Peter G Devitt - peter.devitt@adelaide.edu.au

* Corresponding author

Published: 28 November 2007

BMC Medical Education 2007, **7**:49 doi:10.1186/1472-6920-7-49

Received: 11 April 2007
Accepted: 28 November 2007

This article is available from: http://www.biomedcentral.com/1472-6920/7/49

Abstract

Background: Reliable and valid written tests of higher cognitive function are difficult to produce, particularly for the assessment of clinical problem solving. Modified Essay Questions (MEQs) are often used to assess these higher order abilities in preference to other forms of assessment, including multiple-choice questions (MCQs). MEQs often form a vital component of end-of-course assessments in higher education. It is not clear how effectively these questions assess higher order cognitive skills. This study was designed to assess the effectiveness of the MEQ to measure higher-order cognitive skills in an undergraduate institution.

Methods: An analysis of multiple-choice questions and modified essay questions (MEQs) used for summative assessment in a clinical undergraduate curriculum was undertaken. A total of 50 MCQs and 139 stages of MEQs were examined, which came from three exams run over two years. The effectiveness of the questions was determined by two assessors and was defined by the questions ability to measure higher cognitive skills, as determined by a modification of Bloom's taxonomy, and its quality as determined by the presence of item writing flaws.

Results: Over 50% of all of the MEQs tested factual recall. This was similar to the percentage of MCQs testing factual recall. The modified essay question failed in its role of consistently assessing higher cognitive skills whereas the MCQ frequently tested more than mere recall of knowledge.

Conclusion: Construction of MEQs, which will assess higher order cognitive skills cannot be assumed to be a simple task. Well-constructed MCQs should be considered a satisfactory replacement for MEQs if the MEQs cannot be designed to adequately test higher order skills. Such MCQs are capable of withstanding the intellectual and statistical scrutiny imposed by a high stakes exit examination.

Background

Problem-solving skills are an essential component of the medical practitioner's clinical ability and as such must be taught, learned and assessed during training. Entire curricula have been re-designed with this concept in mind. Problem-based learning is used in many teaching institutions and has its supporters and detractors. Despite criticism, it is undeniable that what problem-based learning

BMC Medical Education 2007, **7**:49 http://www.biomedcentral.com/1472-6920/7/49

sets out to achieve in terms of encouraging and developing the skills of synthesis, evaluation and problem-solving are valued components of a good medical education. In conjunction with the promotion of these skills, an effective assessment process is required. It has long been recognised that in the assessment of clinical competence problem-solving ability has been one of the most difficult areas to measure and quantify [1]. The modified essay question (MEQ) is one of several tools developed to try and assess this skill [2].

The MEQ is a compromise between the multiple-choice question (MCQ) and the essay. A well constructed MCQ will be unambiguous, clearly set to a defined standard and easy to mark (usually automatically), but more often than not tests little more than recall of fact [3]. An essay might test higher powers of reasoning and judgement but will be time-consuming to mark and risk considerable variation in standards of marking [4]. The MEQ is designed to sit in between these two test instruments in terms of the ability to test higher cognitive skills and the ease of marking to a consistent standard. The aim of the modified essay question is to broadly measure both the absolute amount of knowledge retained by the candidate and the ability of the candidate to use that knowledge to reason through and evaluate clinical problems. It accomplishes this by providing a clinical scenario with a number of steps. Progression through these stages should test the candidate's ability to understand, reason, evaluate and critique.

Construction of appropriate MEQs can be difficult [5] and a major criticism of this form of assessment is that MEQs often do little more than test the candidate's ability to recall a list of facts and frustrate the examiner with a large pile of papers to be hand-marked [6].

Although there is evidence to suggest that well constructed MEQs will test higher order cognitive skills [5], and that they can test different facets of understanding than MCQs [7], it is reasonable to ask if MEQ assessments in higher education are well constructed and if they are capable of assessing higher order cognitive skills. This paper describes such a study and is designed to gauge the effectiveness of the MEQ as a summative test tool in a clinical course. We have defined the effectiveness of the questions by their ability to measure higher cognitive skills, as determined by a modification of Bloom's taxonomy, and its quality as determined by the presence of item writing flaws.

Methods

Fourth Year clinical students at the University of Adelaide underwent a written test as part of their overall assessment of performance for a nine-week surgical attachment. The same test instrument was used at the start of the attachment and on completion. The test material consisted of 50 MCQs and three MEQs (a total of 8 stages) and the questions were designed so that both types would cover similar test material. The content, focusing on core material, was matched in both the MCQ and the MEQ components of the examination. The MCQs had one correct answer and four distractors and were constructed to standard guidelines for MCQ construction [8,9].

In addition, the MEQ components of the Final MB BS examination papers for two consecutive years at the University of Adelaide were analysed. The first paper had 15 MEQs with a total of 68 stages, the other had 15 MEQs with a total of. 70 stages. The papers for each examination were assembled by one member of Faculty, who gathered contributions from individual clinicians. There was no formal instruction for the contributors on how to construct an MEQ, which would assess higher order cognitive skills, and the examination organiser undertook the final review of the submitted material.

In total, 33 MEQs made up of 146 stages were collected for analysis. The MEQs were written by at least 12 separate authors using the standard methodology for developing assessments within the faculty.

Each multiple-choice question was quantified independently as to its level of cognitive skill tested [10] and its structural validity [11] by two assessors. Each modified essay question and their individual components was also categorised independently by the two assessors according to the cognitive level measured by each question and its component parts. The assessors discussed their individual assessment and then produced a final grading for each MCQ and MEQ. The inter-rater agreement was calculated using Kappa statistics.

The data was classified using a modification of Bloom's hierarchy of cognitive learning [12,13]. Three levels were defined and classified as shown in Table 1. Level I, covered knowledge and recall of information, Level II covered comprehension and application, understanding and the ability to interpret data, and Level III tested problem-solving, the use of knowledge and understanding in new circumstances.

Table 1: Modified Bloom's taxonomy

Level I:	Knowledge -recall of information
Level II:	comprehension and application -understanding and being able to interpret data
Level III:	problem-solving -use of knowledge and understanding in new circumstances.

BMC Medical Education 2007, **7**:49 http://www.biomedcentral.com/1472-6920/7/49

The rating scale shown in Table 2 was used to judge the rigor of the multiple-choice questions according to the presence of any item-writing flaws.

The item-writing flaws were defined as:

- Repetition of part of the stem in an option
- Use of qualifiers within an option
- Complicated or ambiguous stem
- Negative questions not clearly stated
- Use of double negatives
- Absolute options (e.g., never, always, all-of-the-above)

The cover test has been defined as the ability to surmise the answer from the stem of an item alone, with the correct answer and the distractors covered up [9].

Results

Table 3 illustrates an example of the coding of 2 MCQs. Neither of the MCQs in this table displayed item-writing flaws. Item 1 in the table was judged to be testing lower order cognitive skills than item 2.

Table 4 illustrates stages of an MEQ requiring different levels of cognitive skill to answer. The first two items in the table come from the same MEQ. The last item was obtained from a different question.

The assessors showed a close correlation in their assessment of the questions according to the modified Bloom's taxonomy categorisation. The reliability between the two assessors and the final mark was good with values of Kappa equal to 0.7 and 0.8 for the MCQs and 0.7 and 0.8 for the MEQs.

The overall performances of the MCQs and the MEQs were compared for their ability to test higher cognitive skills (Figure 1). Just over 50% of the MCQs in the Fourth Year examination paper focussed only on recall of knowledge and the largest proportion of MEQs also focussed on this low level cognitive skill. A similar proportion of MCQs and MEQs tested middle order cognitive skills and, rather surprisingly, MCQs were better at addressing the highest order cognitive skills compared with MEQs.

Table 2: Rating scale used to judge the rigor of the multiple-choice questions according to the presence of any item-writing flaws.

Rating	Conditions required to achieve rating
1.	Pass the cover test and no item-writing flaws
2.	Pass the cover test and 1 to 2 item-writing flaws
3.	Cover test dubious and no item-writing flaws
4.	Fail the cover test and 1 to 2 item-writing flaws
5.	Fail the cover test and more than 2 item-writing flaws

Each of the Final Examination papers for 2005 and 2006 contained 15 MEQs and there were a total of 68 and 70 sections respectively (average 4.5 and 4.7 sections per question). In the 2005 paper 51% of the questions tested factual recall (Bloom level I), 47% tested data interpretation (Bloom level II) and only 2% tested critical evaluation. The pattern was similar for the 2006 paper with 54% testing Bloom level I cognitive skills and the remainder (46%) testing Bloom level II.

The 33 MEQs had an average Bloom categorisation of 1.35 with a standard deviation of 0.4. The distribution is shown in Figure 2.

The assessors showed a close correlation in their assessment of the multiple-choice questions according to the item writing flaws categorisation. The reliability between the two assessors and the final mark was moderate, with Kappa equal to 0.5 and 0.6.

An analysis of the structural validity of the MCQs showed that 80% passed the cover test and contained no item-writing flaws. Twenty percent of questions were flawed, but most of these flaws were only of a minor nature and only one question out of the fifty was sufficiently flawed to call into question its structural validity.

Discussion

For an assessment to be effective, there are a number of issues to be considered. Resource considerations are important, and this may have some impact on the style of exam chosen. True-false, multiple-choice and extended matching questions can be marked automatically and may have a relatively low impact on academic time, compared to the marking of MEQ and essay questions. Based on resource considerations alone, MEQs may be considered an inferior form of assessment, but there are other issues, which must be considered.

The reliability and validity of an assessment is vitally important. A reliable assessment will provide consistent results if applied to equivalent cohorts of students. MCQs benefit from a high reliability when the set of questions is valid and there are sufficient numbers of questions, as do True-False questions [14]. MEQs and standard essay questions can have good reliability provided multiple markers are used. Validity of content should always be carried out regardless of the type of assessment tool used. At a mini-

BMC Medical Education 2007, **7**:49 http://www.biomedcentral.com/1472-6920/7/49

Table 3: Sample coding of MCQs

Question	Modified Bloom's taxonomy categorisation	Explanation
A 16 year old obese schoolgirl is admitted with acute pancreatitis. The most likely underlying cause would be A. familial. B. hyperparathyroidism. C. alcohol. D. gallstones. E. trauma.	1	This question is a test of knowledge recall only.
8. A 68-year-old man is hospitalised with his third attack of acute cholecystitis in two years. He is started on a course of antibiotics. He suffered a myocardial infarction one month ago. An isotope scan performed six weeks prior to his present illness showed a non-functioning gallbladder. Which one of the following is the most appropriate treatment? A. immediate percutaneous cholecystolithotomy. B. start on chenodeoxycholic acid. C. allow patient to settle and then perform cholecystectomy within 48 hours. D. allow patient to recover and delay surgery for 5 months. E. proceed to immediate cholecystectomy.	3	There is assumed knowledge in this question. The student needs to make a judgement and evaluation to choose the most appropriate management option.

mum this should include content validity and construct validity. Other measures of validity such as concurrent and predictive validity are also relevant but can be far more challenging to determine. The ability of assessments to discriminate effectively between good and poor candidates, as well as the fidelity of the assessment are also important considerations in evaluating an assessment tool.

We have shown that in a standard mid-course multiple-choice examination paper a substantial component of that examination will focus on testing higher cognitive skills. Yet conversely and perversely, in an examination specifically designed as part of the exit assessment process a disproportionately high percentage of modified essay questions did little more than measure the candidates' ability to recall and write lists of facts. This may be inappropriate when it is considered that the next step for most of the examinees is a world where problem-solving skills are of paramount importance. The analysis has shown that it is possible to produce an MCQ paper that tests a broad spectrum of a curriculum, measures a range of cognitive skills and does so, on the basis of structurally sound questions. It is important to recognise that these results are from one institution only, and the processes used to design assessments may not be typical of other institutions. The generalizability of the results is also worth considering. In this study there were many authors involved in writing the questions. Although it was not possible to isolate individual authors, at least a dozen individuals

Table 4: Sample coding of MEQs

Question	Modified Bloom's taxonomy categorisation	Explanation
A 46 year old woman presents to the emergency department with a three month history of early satiety and anorexia. Over the last two weeks she has been vomiting most days and has been unable to eat or drink much over the last few days. Describe what other information you would seek from the history that would help you establish a diagnosis and justify your answers.	3	Knowledge recall is required, but there is significant interpretation of data required. This makes this a Bloom level 2 at minimum. However, there is a need to evaluate other data, not provided explicitly in this problem in order to arrive a t a diagnosis (problem solving skills). This makes this question a Bloom level 3.
From the history you think that the patient has gastric outlet obstruction. Describe the physical findings you would look for on examination and explain why they might occur.	2	Knowledge recall is required but the student requires understanding of a number of different processes to answer the question correctly. There is no problem solving required, thus making this a Bloom level 2 question.
<from a different problem> Assuming that a mammogram was to be performed as part of the work-up, what are the features suggesting malignancy that would be sought?	1	Knowledge recall of features of malignancy. Requires no understanding of the overall problem.

BMC Medical Education 2007, **7**:49 http://www.biomedcentral.com/1472-6920/7/49

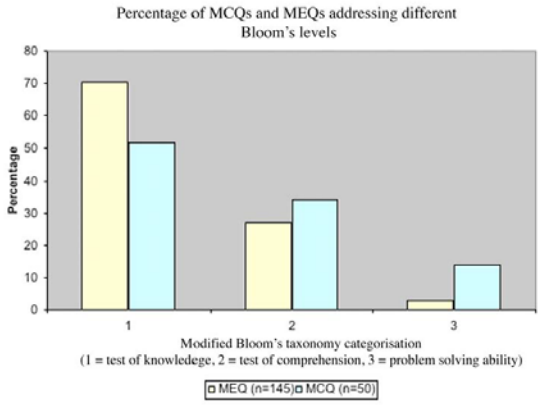


Figure 1
Percentage of MCQs and MEQs addressing different Bloom's levels of cognitive skills.

were involved, and there was little variation in the overall Bloom categorization of the MEQs. This suggests that the findings of this study may be transferable to other schools.

The apparent structural failure of the MEQ papers was not likely the result of a conscious design decision on the part of those who wrote the questions, but may have been a lack of appreciation of what an MEQ is designed to test. This resulted in a substantial proportion of the questions measuring nothing more than the candidates' ability to recall and list facts.

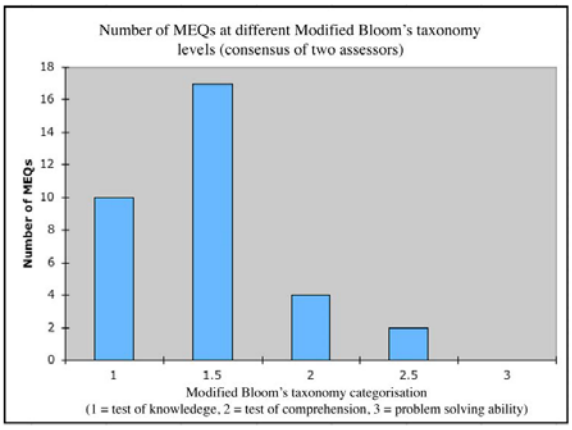


Figure 2
Number of MEQs at different Modified Bloom's taxonomy levels (consensus of two assessors).

This relatively poor performance of MEQs has been observed by others. Feletti [15] reported using the MEQ as a test instrument in a problem-based curricula. In their study the percentage of the examination that tested factual recall varied between 11% and 20%. The components testing problem-solving skills ranged from 32% to 45%. That the proportion of factual recall questions in the current study was higher than that observed by Feletti might well reflect a lack of peer-review when the examination was set. The Feletti data showed that as the number of items increased in the examination, the ability to test cognitive skills, other than factual recall, fell. In other words, the shorter the time available to answer an item, the more likely the material would focus on recall of fact. The University of Adelaide papers allowed 12 minutes a question or less than 3 minutes per stage. This is considerably less than the 2 – 20 minutes per item in the Feletti study.

The open-ended question has low reliability [15] and an examination based on this format is unable to sample broadly. The essay has only moderate inter-rater reliability for the total scores in free-text marking and low reliability for a single problem [16]. Such an examination is also expensive to produce and score, particularly when measured against a clinician's time. It makes little sense to use this type of assessment to test factual knowledge, which can be done much more effectively and efficiently with the MCQ.

Our study has confirmed the impressions reported by others that MEQs tend to test knowledge as much as they measure higher cognitive skills [5]. If an MEQ is to be used to its full value it should present a clinical problem and examine how the students sets about dealing with the situation with the step-wise inclusion of more data to be analysed and evaluated. Superficially, this is what the MEQs in this study set out to do, but when the questions were examined closely, most failed and did no more than ask the candidates to produce a list of facts.

The present study has shown that it is possible to construct a multiple-choice examination paper, which tests those cognitive skills for which the MEQ is supposedly the instrument of choice. These observations raises the question of why it is necessary to have MEQs at all, but the potential dangers of replacing MEQs with MCQs must be considered.

It is generally thought that MCQs focus on knowledge recall and MEQs test the higher cognitive skills. When the content of both assessments is matched the MCQ will correlate well with the MEQ and the former can accurately predict clinical performance [2]. This undoubtedly relies upon a well-written MCQ designed to measure more than knowledge recall.

A good MCQ is difficult to write. Many will contain item writing flaws and most will do no more than test factual recall. Our study has shown that this does not necessarily have to be the case, but it cannot be assumed that anyone can write a quality MCQ unaided and without peer review.

If MCQs are to be used to replace MEQs or similar open-ended format, the issue of cueing must be considered. The effect of cueing is usually positive and can lead to a higher mean score [17]. Conventional MCQs have a cueing effect which has been reported as giving an 11-point advantage compared with open-ended questions. It has been shown that if open-ended questions do not add to the information gained from an MCQ, this difference in the mean score may not matter, particularly if it can lead to the use of a well structured MCQ testing a broad spectrum of material with an appropriate range of cognitive testing [18]. Grading could be adjusted to take into account the benefits of cueing.

Other options to improve the testing abilities of the MCQ type of format is to use extended matching questions and uncued questions [19]. These have been put forward as advances on the MCQ, but these test formats can be easily misused with the result that they may end up focusing only on knowledge recall [4,19,20].

The criticisms levelled at MCQs are more a judgement of poor construction [11,21] and the present study suggests that a similar criticism should be levelled at MEQs. We would go further, and suggest that assessment with well-written MCQs has more value (in terms of broad sampling of a curriculum and statistical validity of the test instrument) than a casually produced MEQ assessment. This is not suggest that MEQs should never be used, as they do have the capability to measure higher cognitive skills effectively [5], and there is evidence to suggest that MEQs do measure some facets of problem solving that an MCQ might not [7].

The measurement of problem-solving skills is important in medicine. MEQs seem ideally suited for this process, but it is possible to use a combination of MEQs and MCQs in a sequential problem solving process, where the ability to solve problems can be separated to some extent from the ability to retain facts [22]. The computer may be the ideal format for this, and there are examples of problem solving exercises using the electronic format readily available [23].

When designing an assessment, which may consist of MCQs or MEQs, it is important to recognise the potential strengths of both formats. This study has shown that if an MEQ is going to be used to assess higher order cognitive skills, there needs to be a process in place where adequate instruction is given to the MEQ authors. If this instruction is not available, and the authors can construct high quality MCQs, the assessment may be better served by containing more MCQs than MEQs. The reduced effort in marking such an assessment would be of benefit to faculties struggling with limited resources.

Conclusion

Apart from its ability to assess appropriate cognitive skills, any assessment instrument should be able to withstand the scrutiny of content and construct validity, reliability, fidelity and at the same time discriminate the performance levels of the cohort being tested. We suggest that a well-constructed peer-reviewed multiple-choice question meets many of the educational requirements and advocate that this format be considered seriously when assessing students. Benefits of automated marking, and potentially high reliability at low cost make MCQs a viable option when writing high stakes assessments in clinical medicine.

Competing interests

The author(s) declare that they have no competing interests.

Authors' contributions

PGD conceived of the study. EP and PGD designed, coordinated and carried out the study. EP carried out the statistical analysis. Both authors participated in the manuscript and read and approved the final version.

References

1. Marshall J: **Assessment of problem-solving ability.** *Medical Education* 1977, **11**:329-34.
2. Rabinowitz HK: **The modified essay question: an evaluation of its use in a family medicine clerkship.** *Medical Education* 1987, **21**:114-18.
3. Epstein RM: **Assessment in Medical Education.** *N Engl J Med* 2007, **356**:387-96.
4. Wood EJ: **What are extended Matching Sets Questions?** *Bioscience Education ejournal* 2003, **1**: [http://www.bioscience.heacademy.ac.uk/journal/vol1/beej-1-2.pdf].
5. Irwin WG, Bamber JH: **The cognitive structure of the modified essay question.** *Medical Education* 1982, **16**:326-31.
6. Ferguson KJ: **Beyond multiple-choice questions: using case-based learning patient questions to assess clinical reasoning.** *Medical Educ* 2006, **40(11)**:1143-.
7. Rabinowitz HK, Hojat MD: **A comparison of the modified essay question and multiple choice question formats: Their relationships to clinical performance.** *Fam Med* 1989, **21**:364-367.
8. Haladyna TM, Downing SM, Rodriguez MC: **A review of multiple-choice item-writing guidelines for classroom assessment.** *App Meas Educ* 2002, **13**:309-334.
9. Case S, Swanson D: **Constructing Written Test Questions For the Basic and Clinical Sciences.** *National Board of Examiners* 2003.
10. Bloom B, Englehart M, Furst E, Hill W, Krathwohl D: **Taxonomy of educational objectives: The classification of educational goals. Handbook I: Cognitive domain.** New York, Toronto: Longmans, Green; 1956.
11. Palmer E, Devitt P: **Constructing multiple choice questions as a method for learning.** *Ann Acad Med Singap* 2006, **35**:604-08.
12. Crooks TJ: **The Impact of Classroom Evaluation Practices on Students.** *Rev Educ Res* 1988, **58**:438-81.

13. Buckwalter JA, Schumacher R, Albright JP, Cooper RR: **Use of an educational taxonomy for evaluation of cognitive performance.** *J Med Educ* 1981, **56**:115-21.
14. Downing SM: **True-false, alternate-choice, and multiple-choice items.** *Educ meas, issues pract* 1992, **11**:27-30.
15. Feletti GI, Smith EKM: **Modified Essay Questions: are they worth the effort?** *Medical Education* 1986, **20**:126-32.
16. Schuwirth LWT, van der Vleuten C: **ABC of learning and teaching in medicine: Written assessment.** *BMJ* 2003:643-45.
17. Schuwirth LWT, van der Vleuten CPM, Donkers HHLM: **A closer look at cueing effects in multiple-choice questions.** *Med Educ* 1996, **30**:44-49.
18. Wilkinson TJ, Frampton CM: **Comprehensive undergraduate medical assessments improve prediction of clinical performance.** *Med Educ* 2004, **38**:1111-16.
19. Veloski JJ, Rabinowitz HK, Robeson MR: **A solution to the cueing effects of multiple choice questions: the Un-Q format.** *Med Educ* 1993, **27**:371-75.
20. Wood TJ, Cunnington JPW, Norman GR: **Assessing the Measurement Properties of a Clinical Reasoning Exercise.** *Teach Learn Med* 2000, **12**:196-200.
21. Collins J: **Education techniques for lifelong learning: writing multiple-choice questions for continuing medical education activities and self-assessment modules.** *Radiographics* 2006, **26**:543-51.
22. Berner ES, Bligh TJ, Guerin RO: **An indication for a process dimension in medicl problem-solving.** *Med Educ* 1977, **11**:324-328.
23. **eMedici** [http://www.emedici.com]. Web page accessed 2007

Pre-publication history

The pre-publication history for this paper can be accessed here:

http://www.biomedcentral.com/1472-6920/7/49/prepub

การสร้างข้อสอบอัตนัยประยุกต์

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

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

ลักษณะพื้นฐานของข้อสอบอัตนัยประยุกต์

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

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

ข้อสอบอัตนัยแบบดั้งเดิมที่ดีนั้นผู้ออกข้อสอบสามารถประเมินทักษะการคิดวิเคราะห์ขั้นสูงได้ แต่อุปสรรคสำคัญที่ทำให้ไม่สามารถบรรลุวัตถุประสงค์ดังกล่าวได้คือการสร้างข้อสอบที่ผู้สอบตั้งเป้าหมายให้ตรวจให้คะแนนได้ง่ายเป็นสำคัญ ทำให้ข้อสอบอัตนัยแบบดั้งเดิมส่วนใหญ่ทำการประเมินเพียงความรู้ระดับความจำหรือความเข้าใจพื้นฐานเท่านั้น

สมมติฐานพื้นฐานในการตอบข้อสอบอัตนัยแบบดั้งเดิมคือการวิเคราะห์และหาแนวทางแก้ปัญหาเป็นกระบวนการที่ทำในขั้นตอนเดียว ดังนั้นข้อสอบจึง

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

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

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

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

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

พัฒนาการของข้อสอบอัตนัยประยุกต์

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

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

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

ตัวอย่างข้อสอบอัตนัยประยุกต์แบบแก้ปัญหาสำคัญ

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

คำถามที่ 1.1 ให้ผู้สอบเขียนปัญหาสำคัญที่สุดของผู้ป่วยรายนี้ 1 อย่าง

ตอนที่ 2 ผู้ป่วยได้รับการประเมินว่ามีภาวะขาดสารน้ำปานกลางถึงรุนแรง ท่านต้องการให้สารน้ำทางหลอดเลือดดำแก่ผู้ป่วย

คำถามที่ 2.1 จงเขียนคำสั่งการรักษาเพื่อให้สารน้ำที่เหมาะสมแก่ผู้ป่วย

คำถามที่ 2.2 จงส่งตรวจเพิ่มเติมทางห้องปฏิบัติการเพื่อช่วยวินิจฉัยผู้ป่วยรายนี้ 2 การตรวจ

จากตัวอย่างข้างต้นจะเห็นว่าผู้ออกข้อสอบไม่ได้เริ่มจากการถามว่าจะซักประวัติ หรือตรวจร่างกายอะไรในผู้ป่วยที่มีภาวะท้องร่วงรุนแรง เนื่องจากผู้ออกข้อสอบเห็นว่าปัญหาสำคัญในการดูแลผู้ป่วยในภาวะนี้เป็นเรื่องการประเมินความรุนแรงของการขาดสารน้ำและการให้น้ำเกลือทดแทนในปริมาณที่เหมาะสมร่วมกับการสืบค้นหาสาเหตุของท้องร่วง ดังนั้นโจทย์ข้อนี้จึงมีเพียงสองตอนและใช้เวลาสอบไม่เกินสิบนาที

ขั้นตอนการสร้างข้อสอบอัตนัยประยุกต์

การสร้างข้อสอบอัตนัยประยุกต์ที่มีคุณภาพดีควรมีการดำเนินการเป็นขั้นตอน ดังนี้[4,20]

1. ตั้งกลุ่มพัฒนาข้อสอบ

ข้อสอบอัตนัยประยุกต์ที่ดีควรเป็นการแก้ปัญหาที่อาศัยความรู้จากหลากหลายวิชา การมีทีมคณาจารย์ที่มีประสบการณ์และความชำนาญแตกต่างกันมาช่วยกันสร้างข้อสอบจะได้สถานการณ์ผู้ป่วยที่เหมือนจริงในเวชปฏิบัติและสามารถประเมินความรู้ของผู้เข้าสอบได้ครอบคลุมสหสาขาวิชา และมั่นใจได้ว่าการเฉลยคำตอบทำได้อย่างรอบคอบ

2. เลือกปัญหาทางคลินิกที่จะทำการประเมินผู้สอบ

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

ลักษณะปัญหาทางคลินิกที่ควรเลือกมาสอบด้วยข้อสอบอัตนัยประยุกต์ ได้แก่

- ปัญหาที่พบได้บ่อยในเวชปฏิบัติ
- ปัญหาที่แพทย์เกิดความผิดพลาดในการดูแลผู้ป่วยค่อนข้างบ่อย
- ปัญหาที่ยังไม่สามารถวินิจฉัยสาเหตุได้ชัดเจน
- ปัญหาที่มีความเกี่ยวพันกับหลายระบบ

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

3. กำหนดปัญหาสำคัญ

เมื่อทีมคณาจารย์เลือกปัญหาทางคลินิกที่จะทำการสอบแล้ว คณาจารย์ต้องตั้งคำถามว่าขั้นตอนใดในการดูแลผู้ป่วยที่มีปัญหาดังกล่าวจัดเป็นขั้นตอนสำคัญที่สุดในการจัดการปัญหานั้น ซึ่งขั้นตอนดังกล่าวจะได้รับการกำหนดให้เป็น ปัญหาสำคัญของสถานการณ์ผู้ป่วยที่จะใช้สอบ ในบางกรณีที่ทีมคณาจารย์ไม่สามารถเลือกขั้นตอนสำคัญในปัญหาทางคลินิกนั้น ๆ จากวิธีดังกล่าวได้

อาจใช้คำถามว่าขั้นตอนใดในการดูแลผู้ป่วยที่มีปัญหาดังกล่าวเป็นขั้นตอนที่นักศึกษาแพทย์หรือแพทย์ประจำบ้านทำผิดพลาดมากที่สุด[4]

มีข้อแนะนำสองประการสำหรับการกำหนดปัญหาสำคัญในแต่ละสถานการณ์ ได้แก่

- สิ่งที่ต้องตัดสินใจในผู้ป่วยแม้เป็นสิ่งที่ถูกต้องและควรปฏิบัติอาจไม่ได้เป็นขั้นตอนสำคัญที่จะต้องนำมาสอบเสมอไป การปฏิบัติต่อผู้ป่วยหลายอย่างที่ทำกันเป็นปกติ โดยไม่ต้องคิดวิเคราะห์ เป็นขั้นตอนที่ไม่ค่อยทำผิดพลาด มักไม่ใช่ปัญหาสำคัญในสถานการณ์นั้น

- ปัญหาสำคัญไม่จำกัดอยู่เฉพาะประเด็นปัญหาทาง ชีววิทยาการแพทย์ (biomedical) เท่านั้น ในบางสถานการณ์ปัญหาสำคัญอาจเป็นประเด็นทางจริยธรรม กฎหมาย หรือ การส่งเสริมสุขภาพและป้องกันโรคก็ได้

4. เขียนโจทย์คำถาม

เมื่อมีสถานการณ์ผู้ป่วยและขั้นตอนที่เป็นปัญหาสำคัญในสถานการณ์นั้นแล้ว ทีมคณาจารย์ต้องเขียนโจทย์คำถามที่มีความชัดเจน เพื่อประเมินว่าผู้เข้าสอบมีความสามารถในการตัดสินใจในการแก้ปัญหาสำคัญในสถานการณ์ดังกล่าวหรือไม่ โดยทั่วไปแล้วลักษณะโจทย์คำถามที่ใช้บ่อยในข้อสอบอัตนัยประยุกต์ได้แก่

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

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

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

การกำหนดจำนวนคำตอบนี้จะทำให้ผู้เข้าสอบต้องเลือกสิ่งที่ถูกต้องเหมาะสมที่สุดเท่านั้นมาเขียนตอบ หากผู้เข้าสอบเขียนคำตอบเกินจำนวนที่กำหนด อาจารย์ผู้ตรวจข้อสอบจะไม่อ่านคำตอบที่เกินมา การปฏิบัติเช่นนี้จะช่วยกำจัดปัญหาการตรวจกระดาษคำตอบที่ผู้เข้าสอบเขียนคำตอบแบบหว่านแห ให้ครอบคลุมทุกอย่างโดยที่ผู้เข้าสอบเองไม่มีความรู้ ความเข้าใจว่าสิ่งใดเป็นประเด็นสำคัญในการดูแลผู้ป่วยในขั้นตอนนั้น ๆ

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

5. กำหนดเกณฑ์การให้คะแนน

ขั้นตอนสุดท้ายในการสร้างข้อสอบอัตนัยประยุกต์คือการกำหนดเกณฑ์การให้คะแนน ซึ่งเป็นขั้นตอนที่มีความท้าทาย และสร้างความลำบากใจให้แก่อาจารย์ผู้ออกข้อสอบหลายท่าน เนื่องด้วยเกรงว่าจะเฉลยคำตอบไม่ครอบคลุมสิ่งที่ผู้เข้าสอบจะเขียนตอบมา หรือเกิดความไม่เป็นธรรมขึ้น ในที่นี้ผู้นิพนธ์ขอเสนอแนะแนวทางในการกำหนดเกณฑ์ให้คะแนนดังนี้

- แนะนำให้กำหนดคะแนนเต็มในการแก้ปัญหา

สถานการณ์หนึ่ง ๆ เป็น 100 คะแนน เท่ากันในทุกสถานการณ์ เพื่อให้ไม่ต้องทำการปรับคะแนนสอบหลังการตรวจข้อสอบ

- กรณีที่มีคำตอบที่ถูกต้องยอมรับได้เพียงคำตอบเดียว เช่นข้อมูลจากโจทย์มีความชัดเจนว่าผู้ป่วยเป็นโรคอะไร แล้วโจทย์ให้ผู้เข้าสอบตอบชื่อโรค หากผู้เข้าสอบตอบตรงตามเฉลยที่ตั้งไว้ให้ได้คะแนนเต็ม หากตอบคำตอบอื่นนอกจากนั้นไม่ได้คะแนน

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

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

- คำตอบที่ไม่ถูกต้อง ไม่สมควรปฏิบัติแก่ผู้ป่วย โดยทั่วไปแล้วพิจารณาไม่ให้คะแนน ซึ่งก็จัดเป็นการทำโทษในระดับหนึ่งแล้ว เพราะผู้สอบมีสิทธิเขียนคำตอบได้จำนวนจำกัด การที่ไม่ให้คะแนนในคำตอบที่ไม่เหมาะสม ก็จะทำให้คะแนนสูงสุดที่ผู้สอบจะทำได้ลดลงไปแล้ว การปฏิบัติที่ไม่ถูกต้องที่มีผลเสียรุนแรงต่อผู้ป่วยเท่านั้นที่ควรจะพิจารณาให้คะแนนติดลบ และแม้มีการให้คะแนนติดลบก็ไม่ควรมีการติดลบข้ามไปถึงข้อสอบข้ออื่นในชุดข้อสอบนั้น

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

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

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

(2) ไล่ดูโจทย์คำถามทีละข้อพร้อมเฉลย แล้วทำสัญลักษณ์ * ไว้ในคำตอบที่คาดว่าผู้ที่มีความรู้ ความสามารถคาบเส้นจะตอบในข้อสอบแต่ละตอน

(3) ทำการรวมค่าคะแนนที่ได้รับการทำสัญลักษณ์ * ไว้ตั้งแต่ข้อแรกจนถึงข้อสุดท้าย จะได้คะแนนเกณฑ์ผ่านในการแก้ปัญหาสถานการณ์นั้น ๆ[22]

แนวทางการพัฒนาข้อสอบอัตนัยประยุกต์ในคณะแพทยศาสตร์ศิริราชพยาบาล

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

(1) **เนื้อหาข้อสอบ**

ข้อสอบอัตนัยประยุกต์ที่ใช้ในการสอบประมวลความรอบรู้ทางการแพทย์ของคณะแพทยศาสตร์ศิริราชพยาบาลที่ผ่านมาหลายข้อเป็นเนื้อหาวิชาที่ยากและเป็นความรู้ลึกในระดับผู้เชี่ยวชาญเฉพาะทาง แนวทางในการพัฒนาการสอบอัตนัยประยุกต์อันดับแรกคือการพัฒนาเนื้อหาให้เหมาะสมกับการประเมินความรู้ของแพทย์เวชปฏิบัติทั่วไป

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

ข้อแนะนำในเรื่องเนื้อหาที่สำคัญคืออาจารย์ผู้ออกข้อสอบต้องตระหนักว่าการสอบนี้เป็นการประเมินความรู้เวชปฏิบัติทั่วไป มิใช่การประเมินความรู้เชิงลึกในศาสตร์ของแต่ละสาขาวิชา โรคหรือภาวะที่นำมาออก

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

(2) **รูปแบบคำถาม**

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

(3) **จำนวนสถานการณ์ผู้ป่วยที่ใช้สอบ**

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

หากพิจารณาจากข้อแนะนำของผู้เชี่ยวชาญในการประเมินผลที่ได้อภิปรายไปก่อนหน้านี้ที่แนะนำให้ใช้ข้อสอบอัตนัยประยุกต์แบบแก้ปัญหาสำคัญ แนวทางการพัฒนาข้อสอบอัตนัยประยุกต์ของคณะฯ ให้มีความครอบคลุมสถานการณ์ผู้ป่วยที่มากขึ้น และมีความเที่ยงของคะแนนสอบมากขึ้นคือการใช้ข้อสอบแบบแก้ปัญหาสำคัญมาแทนการจัดการปัญหาของผู้ป่วยตั้งแต่ต้นจนจบ กล่าวคือในแต่ละสถานการณ์ผู้ป่วย ข้อสอบควรมุ่งถามคำถามสำคัญเพียงสองหรือสามข้อ และเพิ่มจำนวนสถานการณ์ผู้ป่วยให้มากขึ้นนั่นเอง

(4) **การนำเสนอข้อสอบ**

การทำข้อสอบอัตนัยประยุกต์ ผู้สอบต้องทำงานภายใต้ข้อจำกัดด้านเวลา เวลาที่ใช้ในการตอบข้อสอบอัตนัยประยุกต์เป็นผลรวมของเวลาที่ใช้อ่านโจทย์ คิดวิเคราะห์ และเขียนคำตอบ ปัญหาสำคัญประการหนึ่งที่สร้างความลำบากให้กับผู้สอบคือปริมาณข้อมูลที่นำเสนอให้ผู้สอบอ่านในสถานการณ์ผู้ป่วยแต่ละรายนั้นมีมาก ทำให้ผู้สอบต้องใช้เวลาในการอ่านมากและเหลือเวลาสำหรับเขียนคำตอบน้อย ถึงแม้ว่าในการนำเสนอข้อมูลของข้อสอบอัตนัยประยุกต์จะได้มีการแยกข้อมูลเดิมที่เคยนำเสนอไปก่อนหน้านี้ ออกจากข้อมูลใหม่ที่เพิ่มเติมขึ้นมาในการนำเสนอข้อสอบแต่ละตอนแล้วก็ตาม ด้วย

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

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

ระบบจัดสอบอัตนัยประยุกต์ด้วยคอมพิวเตอร์อย่างเต็มรูปแบบที่ไม่ต้องมีการเขียนตอบในกระดาษเลยนั้นมีการจัดทำในต่างประเทศ[12,23] แต่ต้องยอมรับว่าการ

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

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

บทสรุป

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

เอกสารอ้างอิง

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

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

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

ขอแก้ไขเป็น

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

16 March 2017

Long-case Examination

- One of assessment instruments
- Clinical/Practical Assessments
- Long- and short-case examination
  - Short-case examination: individual component
  - Long-case examination: assessment on the patient as a whole

Long Case Examination

- The examinees
  - spend a long period of time
  - explore and work up a single patient case
- An examiner assesses
  - history taking
  - physical examination
  - communication skills
  - diagnostic skills
  - plan of investigations and management
  - professionalism of the examinees

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
- Powerful tool of feedback

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

Long-case Examination

- **Problems**
  - Objectivity
  - Validity
  - Reliability

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

Stokes, 1974

1

Long-case Examination

- Use of a non-standardised real patient
- May provide a unique opportunity to test
  - the physician's tasks and interaction with a real patient
- Has poor content validity
  - Less reliable and lacks consistency
  - Reproducibility of the score is 0.39
- In high stake summative assessment long case should be avoided

Norcine,2002
Int J Health Sci (Qassim). 2008; 2(2):3–7

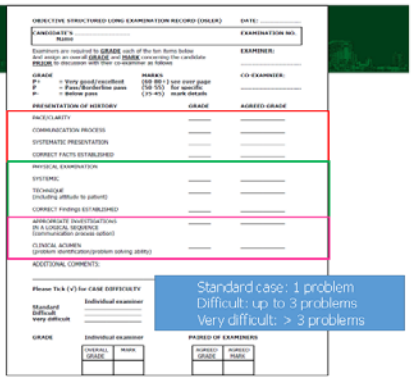
OSLER

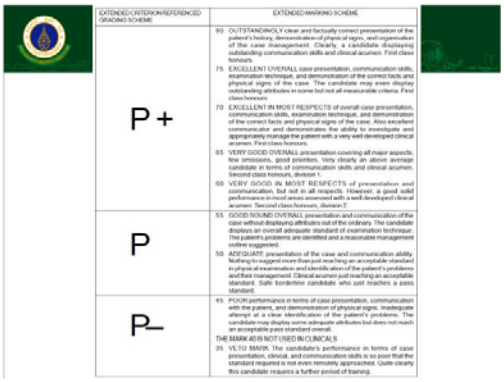
The Objective Structured Long Examination Record (OSLER)

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

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





Long-case Examination

- **Three variables**
  - Candidates***
  - Examiners
  - Patients

2

Long-case Examination

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

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 ศรว. ในการสอบ long case ข้อกำหนดของ ศรว. ในการสอบ long case examination
  1. จำนวนผู้ป่วยอย่างน้อย 2 ราย
  2. โรค หรือ ปัญหาสอดคล้องกับเกณฑ์มาตรฐานผู้ประกอบวิชาชีพเวชกรรมของแพทยสภา
  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)

3

Medical Teacher

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

AMEE Medical Education Guide No. 9. Assessment of clinical competence using the Objective Structured Long Examination Record (OSLER)

Fergus Gleeson

To cite this article: Fergus Gleeson (1997) AMEE Medical Education Guide No. 9. Assessment of clinical competence using the Objective Structured Long Examination Record (OSLER), Medical Teacher, 19:1, 7-14, DOI: 10.3109/01421599709019339

To link to this article: http://dx.doi.org/10.3109/01421599709019339

Published online: 03 Jul 2009.

Medical Teacher, Vol. 19, No. 1, 1997

AMEE Medical Education Guide No. 9. Assessment of clinical competence using the Objective Structured Long Examination Record (OSLER)

FERGUS GLEESON
James Connolly Memorial Hospital & The Royal College of Surgeons in Ireland, Dublin, Ireland

SUMMARY *Much criticism has been directed at the assessment of clinical competence and at the long case in particular in recent years. In the traditional long case candidates spend one hour with a patient from whom they take a history and whom they examine. An examiner is not present. The student is then examined by a pair of examiners over a 20–30 minute period. This has been to the extent that the problems associated with the long case in terms of objectivity, validity and reliability are such that some critics have suggested that it should be abandoned altogether. Others would take the view that before we dispense with this method we should attempt to remodel and improve it. Furthermore, tradition and practically would suggest that the long case will be with us for some time to come. The justifiable criticism of the long case is directed on a number of fronts, a major one being that the history-taking process is not observed by the examiners. Bearing these criticisms in mind, the Objective Structured Long Examination Record (OSLER) has been developed. The OSLER is a 10-item analytical record of the traditional long case which attempts as far as is possible within the limits of practicality to improve the objectivity, validity and reliability of existing practices. All candidates are assessed over 20–30 minutes by the examiners on the same 10 items, thus improving reliability and items are included that are representative of what would be regarded as having an acceptable degree of construct or face validity with regard to the long case. Attention is paid to communication skills and the history-taking process in particular. In attempting to standardize the long case and minimize the 'luck of the draw' aspect, examiners are requested to formally document the difficulty of the case. The figure of 10 with regard to the number of items assessed is not coincidental and is a deliberate act to include a minimum of the essential in terms of what should be assessed. This allows examiners to concentrate on the candidate's performance with a structured guide that is not so intrusive as to interrupt the examiner's concentration. The four items on history include pace and clarity of presentation, communication skills process, systematic approach and establishment of the case facts. Three items on physical examination include systematic approach, examination technique and establishment of the correct physical findings. During these activities the candidate's affective behaviour is also assessed. The remaining three items include construction of appropriate investigations in a logical sequence, appropriate management and final clinical acumen. The latter item draws on the previous nine to assess candidates' ability to identify and solve problems. The initial assessment is essentially criterion referenced through a P+, P, P− system which is followed by the selection of an appropriate mark, each of which has its own written descriptive profile. The perfect method for long case clinical assessment has yet to be established. Indeed perfection may be no more than a pious hope bearing in mind that any method will always be a compromise between objectivity, validity and reliability on one hand and practicality on the other. While the search for the perfect long case method continues, the OSLER is suggested as a practical approach to what is universally recognized as an ongoing assessment challenge.*

Introduction

Assessment is treated with great reverence in the vast majority of medical schools. Lowry (1993), however, has recently posed the question: Is assessment as powerful as we think, and if it is, are most medical educators using it

Correspondence: Fergus Gleeson, FRCPI, Associate Professor of Medicine, James Connolly Memorial Hospital, Dublin, Ireland. Fax: 8204708.

0142-159X/97/010000-00 © 1997 Journals Oxford Ltd

effectively? Clinical assessment in many medical schools, in spite of frequent criticism, has continued to be a combination of what are commonly termed long- and short-case examinations. This combination is likely to persist if one accepts that clinical assessment, to be truly valid, must be patient centred. Over the past 20 years, the short case examination has received much attention with the introduction of the Objective Structured Clinical Examination (OSCE) (Harden & Gleeson, 1979). With most attention being paid to these improvements, the long case has largely been ignored. While improvements such as the OSCE have focused attention on the individual components of clinical competence, it is widely agreed that there is still need for a method to assess students on the patient as a whole. The traditional long-case examination has been our method of fulfilling this role. There has been much justified criticism of the long case in which different examiners examine different candidates on different patients. This has very rightly been referred to as the 'luck of the draw' (Stokes, 1974). What, therefore, educationally constitutes a good assessment method? It should be objective, valid and reliable. Why does the existing long case fail to meet such criteria to an acceptable degree? Such assessments are frequently heavily subjective in that there is little prior agreement between pairs of examiners or indeed by institutions as a whole as to what constitutes a valid assessment. In other words, there is a lack of or no agreement as to what has to be measured during the course of the examination. What constitutes a valid assessment? Such an assessment measures what it is supposed to measure, i.e. clinical skills. These skills include the ability to obtain information by way of history, physical examination and investigations, to use this information to solve patients' problems and finally to utilize the solution to problems by way of management. In most existing long-case assessments history taking as such is not validly assessed. While the product may be assessed, the far more important history-taking process is not observed and therefore not validly assessed. This is a highly significant omission when one considers the relative value of the history in terms of overall diagnostic problem solving (Hampton *et al.*, 1975; Miall, 1992).

It is for this reason that the OSCE has achieved much of its deserved success as observed history taking plays a significant part in such assessments. However, while the OSCE displays a relatively high content validity, it has a relatively low face or what is termed construct validity. In other words, while it assesses the parts very well it does not assess the whole candidate/patient interaction on one and the same patient which, after all, is what occurs in the practice of medicine. In view of what has been stated in terms of objectivity and validity, there is thus a high probability that there will be inconsistencies or a lack of reliability in the marking of the long case if there is not a clear agenda to be followed. While examiners may at times be following a similar agenda, the items may receive significantly different emphasis by the individuals of pairs of examiners so that marking inconsistencies are a recognized problem (Fleming *et al.*, 1976).

While the importance of patients' histories is universally recognized, the emphasis placed on physical examination in the long case needs to be critically reappraised for a number of reasons. One of these is the lack of gross physical signs in the majority of patients examined in practice. Furthermore, many very difficult clinical problems do not have any physical signs and are therefore frequently not used in such assessments (Weatherall, 1991). Social and psychological factors play a significant part in the day-to-day problems encountered in both hospital and community practice.

Increasingly therefore, recognition of the value of communication skills is being highlighted (Irwin *et al.*, 1989; Doherty *et al.*, 1990; McManus *et al.*, 1993). The communication skill necessary to acquire information on difficult clinical problems is very real and consequently places a responsibility on institutions and their examiners to establish that such skills have been developed by way of assessment. This need therefore emphasizes again the relative importance of the history. Another key factor is the degree of case difficulty over a wide range of long cases, which is very variable and must be given due recognition by examiners. A further danger in the course of long-case examinations is that during the assessment, unless there is a structure to be followed, the emphasis may shift from the clinical to that of a viva or oral assessment. This is not an infrequent occurrence and in such circumstances the validity of the clinical examination is obviously seriously compromised. There are therefore genuine concerns about the existing traditional long case which frequently result in the making of global pass/fail decisions in a non-structured fashion. Such decisions at times result in questionable outcomes in terms of justice to the candidate and to the public at large, which is the ultimate reason for all assessments. Such a scenario could be likened to referees adjudicating in different games, using different rules and in the end miraculously producing an overall winner. Can one imagine a similar scene in any other field of human endeavour?

In the traditional long case students spend one hour with a patient from whom they take a history and whom they examine. An examiner is not present. The student is then examined by the examiner over a 20–30 minute period.

Long-case assessments at both undergraduate and postgraduate level are in most instances carried out over a short period of time, e.g. one week. During this time frame, large numbers of candidates have to be assessed. There is therefore a need for an improved long-case format to assess such large numbers that is practical to implement but at the same time recognizes the essential criteria of objectivity, validity and reliability in so far as this is possible under the time constraints already referred to. In spite of the obvious problems that exist with regard to the long case, there is a natural reluctance to change established practices unless very real benefits are possible. For any educational innovation to succeed, there are certain criteria to be fulfilled (Collingwood, 1979). The innovation must have a relative advantage over existing practice. The complexity of the innovation must not be such as to evoke an immediate and negative attitude. It must have trialability in that it can be introduced and removed in the event of failure without producing a major convulsion in the system. Finally, it must be seen to have observability in that the more visible the effect of the innovation, the more likely will be its acceptance.

The current unstructured global marking of the long case has major potential for unreliable assessment. An essential requirement therefore is the need to structure a

number of items for examiners to deliberate on. This results in turn in the introduction of the concept of the checklist. For a comprehensive long-case examination, the potential length of such a list would be so great as to be impractical to implement. Such an instrument would end up being both an invalid and unreliable instrument in that the examiner would spend more time concentrating on the checklist rather than on the actual measurement of the candidate's performance. More realistic approaches such as the observed long case proposed by Newble (1991) and a similar approach by Price & Byrne (1994), for assessment skills in psychiatry are both very expensive in terms of examiner time. Both approaches require examiners to be present for the whole history-taking process carried out by the candidates. This extra time element would make such assessments impractical for the vast majority of institutions, particularly in those situations in which large numbers of candidates have to be assessed in a relatively short time frame. The method adopted, therefore, must be comprehensive enough to allow for valid judgements by examiners, be practical to use and at the same time be perceived as fair by the candidates. Such perception demands that, as with all assessment instruments, it must be seen to be objective. It must also be structured so that all candidates are assessed using the same criteria leading to greater consistency or reliability.

A valid method for the assessment of long-case clinical competence must include essential principles. These are the recognition of and observation of the history-taking process. While such is being observed the examiner has the opportunity of assessing the communication skill of the candidate. Physical examination skill is essential, as is the ability to construct a series of investigations. All of the foregoing allow the examiner to deliberate on the candidate's ability to identify and solve problems. Finally there is a requirement to assess the candidate's ability to manage the problem, which again involves skills of communication as well as overall management. During all these activities, the examiners will also have the opportunity of assessing the affective behaviour or attitude of the candidate towards the patient. The assessment instrument must be practical in terms of its length and usage by the examiner whose primary function is to concentrate on the candidate's performance. The Objective Structured Long Examination Record (OSLER) has been developed in an attempt to fulfil the stated foregoing criteria and principles. Examiners spend 20–30 minutes with the student who has already examined and taken a history from the patient.

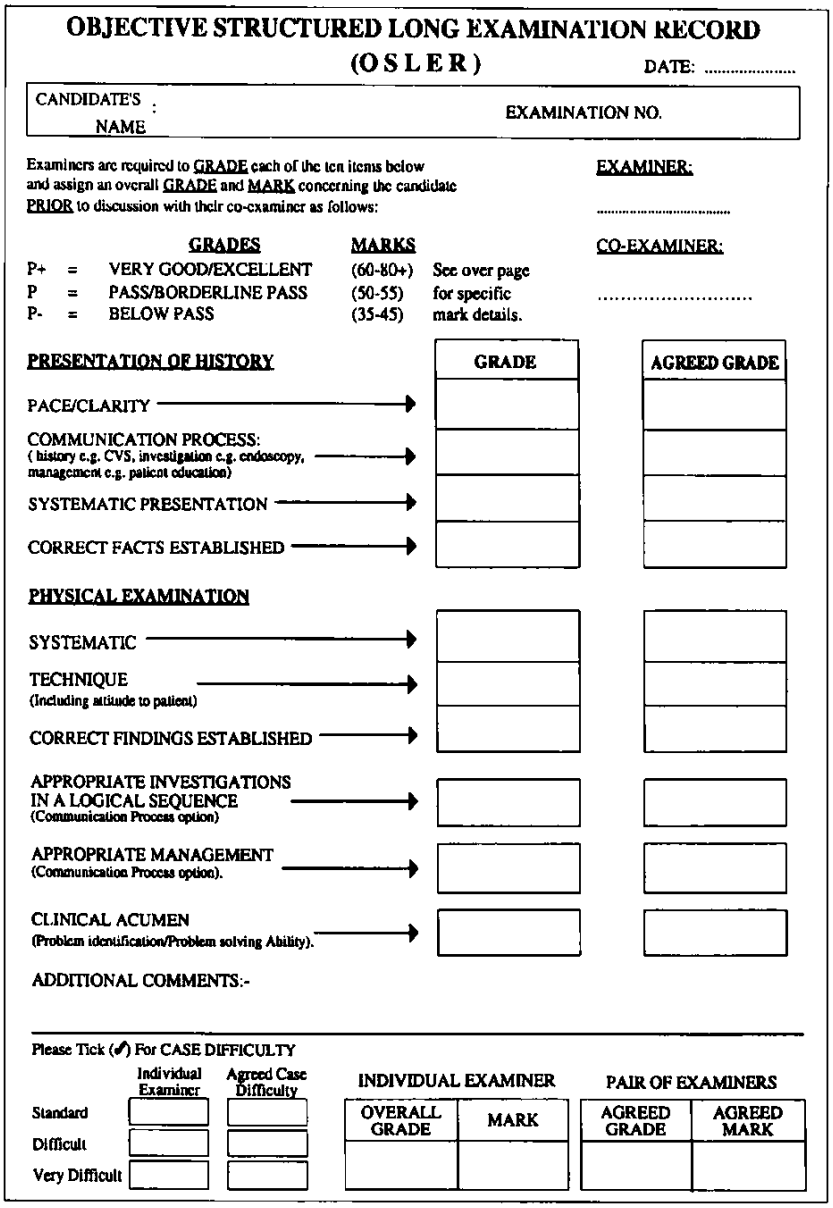
Method

Presentation of history

The OSLER consists of 10 items (Figure 1) which include four on history, three on physical examination and the remaining three cover investigation, management and clinical acumen. The figure of 10 is not coincidental and is a deliberate act to include as much as is essential but as little as possible. This is to allow the examiner to concentrate on the candidate's performance with a guide that is not so intrusive as to interrupt the examiner's concentration. The four items assessed on the history are pace and clarity of presentation, communication skills process, systematic approach and establishment of the case facts. Pace/Clarity essentially assesses communication between the candidate and the examiner. Pace of presentation measures rate of speech with appropriate pauses. Too rapid and it is unintelligible, too slow and it is inefficient in terms of time economics. Clarity is obviously allied to pace but at the same time recognizes the need and ease with which the examiner observes the unfolding story that is the history. Greater emphasis is now being placed on communication skills in medical schools (GMC, 1993) and the inclusion of the first item recognizes this fact. Graduates of medical schools are employed worldwide and if the candidate cannot effectively communicate with the examiner, he/she cannot be validly assessed. More importantly, if the candidate cannot make him/herself understood by the examiner, what chance has the patient? It is essential therefore that the examiner has an opportunity of observing the communication skills of the candidate with the patient through the second item, communication process. This is achieved by requesting the candidate to take a history for three minutes concentrating on one system, e.g. cardiovascular, or segment of the history, e.g. social history. By observing this process and listening to the remainder of the history, the examiner can form an opinion as to the candidate's ability to communicate with the patient. Alternatively the communication skill of the candidate can be assessed during the assessment of the investigation or management sections. This could be achieved by the candidate describing to the patient a particular investigation, e.g. colonoscopy. Alternatively a candidate could be asked to explain to the patient, as part of the management, the usage and dangers of anticoagulants. By listening to the remainder of the history, the candidate's ability to systematically go through the story in a logical sequence can be assessed. Finally it is essential that the candidate demonstrates his/her ability to accurately establish the correct facts of the case.

Physical examination

Three items are a minimum of the essential for inclusion in relation to physical examination. Here again the process as well as the product is being observed and assessed. A systematic approach will reveal something of the candidate's ability to logically approach the subject to obtain the necessary information to problem solve. However, the key to successful physical examination lies in a well-developed technique. This item deals with the candidate's psychomotor skills and, like all such skills, frequent practice is the essential requisite. An experienced astute examiner will be in a position to decide on the merits of a candidate in this section. Not alone are the pure psychomotor skills being observed but also the candidate's confidence and attitude towards the patient. Influences other than technique can affect the performance of psychomotor skills on any particular occasion; however, the candidate with a truly professional approach which includes attention to detail can overcome such influences. The most obvious of these is the relative difficulty of the case the candidate is assigned. The 'luck of the draw' is a well-accepted factor and the experienced examiner will recognize this. All examiners therefore need to be consistently conscious of this factor and an



OBJECTIVE STRUCTURED LONG EXAMINATION RECORD
(O S L E R)

DATE: .....

CANDIDATE'S NAME : EXAMINATION NO.

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

EXAMINER: .....

CO-EXAMINER: .....

GRADES			MARKS	
P+	=	VERY GOOD/EXCELLENT	(60-80+)	See over page
P	=	PASS/BORDERLINE PASS	(50-55)	for specific
P-	=	BELOW PASS	(35-45)	mark details.

	GRADE	AGREED GRADE
PRESENTATION OF HISTORY		
PACE/CLARITY →		
COMMUNICATION PROCESS: (history e.g. CVS, investigation e.g. endoscopy, management e.g. patient education) →		
SYSTEMATIC PRESENTATION →		
CORRECT FACTS ESTABLISHED →		
PHYSICAL EXAMINATION		
SYSTEMATIC →		
TECHNIQUE (Including attitude to patient) →		
CORRECT FINDINGS ESTABLISHED →		
APPROPRIATE INVESTIGATIONS IN A LOGICAL SEQUENCE (Communication Process option) →		
APPROPRIATE MANAGEMENT (Communication Process option). →		
CLINICAL ACUMEN (Problem identification/Problem solving Ability). →		

ADDITIONAL COMMENTS:-

Please Tick (✓) For CASE DIFFICULTY

	Individual Examiner	Agreed Case Difficulty
Standard		
Difficult		
Very Difficult		

INDIVIDUAL EXAMINER		PAIR OF EXAMINERS	
OVERALL GRADE	MARK	AGREED GRADE	AGREED MARK

Figure 1. The OSLER

assessment of the case difficulty is included in the OSLER to aid this process. It should of course also be borne in mind that, in later practice, the 'luck of the draw' will apply on a daily basis and the candidate should have the flexibility to demonstrate that he/she can handle any given situation under the prevailing circumstances. Whatever difficulties are encountered the candidate has to correctly identify the clinical signs to proceed satisfactorily to manage the patient's problem.

Investigation, management and clinical acumen

For the item on investigation, the examiner is requested to assess the candidate's ability to construct appropriate investigations for the case in question in a logical sequence. Frequently, appropriate investigations might be suggested but the sequence would be inappropriate either in terms of invasiveness of the patient or in terms of costs. In addition the examiner also has an opportunity to assess the candidate's ability to logically sequence his/her thought processes in a limited time. This is an additional skill which is essential for later efficient practice. Management is the next skill to be assessed. Here the candidate can range from either killing to curing the patient. The examiner has a duty not to release a candidate on an unsuspecting public who is not properly prepared. This concept can be rela-

tively blurred in a situation where the candidate performs well in the earlier items but in this critical area can be found wanting with potentially disastrous consequences.

Clinical acumen is the overall ability of the candidate to identify the patient's problems and to put the diverse parts of the case together to produce a whole product in terms of problem identification and the ability to solve such problems in overall management terms. Increasingly, the importance of identifying problem-solving ability is being recognized (Barrows & Feltonich, 1987; *Lancet*, 1989; Cassirer, 1992). The inclusion of this item therefore is an essential criterion of clinical competence as the examiners have to attempt to extrapolate from this situation the candidate's ability to perform consistently over a range of such situations or cases. This crucial decision by the examiner has suspect potential if it is made in a global fashion as frequently occurs without the support of the clearly identified previously described nine items. There is evidence to demonstrate that the ability to solve problems will vary from case to case (Elstein *et al.*, 1978). This in turn makes it all the more important to recognize and include this item for valid judgements by examiners. To further assist the examiner in this respect and also to minimize the 'luck of the draw' element for the candidate as far as this is feasible, the difficulty of the case is noted.

Case difficulty

As long cases vary in their degree of difficulty, it is necessary for examiners to establish the relative difficulty of the case under consideration. Not to do so would seriously compromise the validity and reliability of the overall assessment. The case difficulty has been arbitrarily divided into 'standard cases', which would represent a single problem, 'Difficult' cases, which would include up to three problems and 'very difficult' cases, with greater than three problems. However, it will be appreciated that a single problem could amount to a very difficult case. Examiners therefore have to grade difficulty in the context of the case in question and it will be obvious therefore that this decision has to be made prior to commencing the assessment itself.

Grading and marking

It has long been recognized that awarding marks in the long case is unreliable (Wilson *et al.*, 1969) and short training periods for examiners have yielded little improvement (Ludbrook & Marshall, 1971). This is not too surprising as there has been little examiner training on methods that in turn frequently lack objectivity and validity. Prior to the awarding of a mark in the OSLER, a grading system has been adopted. Performance therefore is graded as P + (very good/excellent), P (pass/bare pass) and P − (below pass) for each of the 10 items followed by an overall grade for the complete performance. This is how the vast majority of examiners instinctively make initial assessment decisions. This could be described as an extended criterion-referenced method in that candidates are measured against the criterion for the standard of the clinical assessment in question, i.e. undergraduate or postgraduate. Having decided on an appropriate grade for each individual item and then an overall grade, examiners using the OSLER are then in a position to select an appropriate mark from a designated list of possible marks, each of which is backed up with a stated written mark Profile (Figure 2). Individual examiners, having decided on their overall grade and mark for the candidate, are then in a position to confer with their co-examiner during which time they agree a grade for each of the 10 individual items, an overall grade and finally an agreed mark. This combination of grading and marking amounts to 138 formal decisions being made for any one individual candidate when both examiners are taken into account.

Discussion

The OSLER has now been used for 10 years, during which time important data has emerged. The detailed information that is available following such OSLER assessments has highlighted serious defects in basic clinical skills. This has been noted in both undergraduate and more particularly in postgraduate studies (Gleeson, 1992). The identification of such defects was not too surprising as such findings have been noted in other studies (Maguire & Rutler, 1976; Wiener & Nathanson, 1976; Wray & Friedland, 1983; Sox *et al.*, 1985; Chan Yan *et al.*, 1988). Of even more significance has been the documented immediate marked improvement between two OSLER assessments on 230 postgraduate students within 48 hours (Gleeson, 1995). The time interval was such that only the feedback knowledge of such defects could have influenced the improvement. This finding is all the more important as feedback is regarded as a key step in the development of such skills (Ende, 1983). In a recent *Lancet* commentary the following was stated: "OSLER seems to be a powerful tool for providing feedback and therefore has great potential to increase clinical competence" (Van Der Vleuten, 1996).

What are the advantages of the OSLER in the context of educational assessment criteria? Objectivity is enhanced by prior agreement on what is to be assessed. In any long case there are three variables which are the candidate, the examiners and the patient. Ideally the only variable should be the candidate. Strenuous efforts are being made to standardize patients, particularly through simulation, in North America. For the foreseeable future, however, such standardization will not be practicable or, indeed, for many desirable. In the meantime we must strive to standardize our examiners by assisting them to be as reliable or consistent as possible in their assessments. Recognition of this is already obvious by having two examiners assessing each candidate. The end result, however, is not as perfect as one would anticipate on many occasions. It is not acceptable or good practice for a pair of examiners to confer on the merits of a candidate prior to awarding an individual grade or mark. Many examining authorities increasingly recognize this problem but in some instances have been slow to insist on its implementation. The OSLER, with its increased number of items and fixed structure, will assist individuals of a pair of examiners in their decision making and thus make it easier for examining authorities to insist on the implementation of individual marking prior to examiners conferring. Examiners also require to be conscious that they are assessing broad clinical skills in addition to detailed case-specific skills. There is evidence that such an

The pass mark is 50. Marks should be given in 5s (e.g. 80, 75, 70, 65, 60 etc) in accordance with the following guidelines. Intermediate marks, e.g. 53, 67 should not be used.

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

Examiners should not be hesitant in awarding high or low marks when justified.

Figure 2. The OSLER marking profile.

approach is more reliable (Van Thiel *et al.*, 1991). Is the traditional long case valid, i.e. does it measure what it is supposed to measure in assessing how the student handles the patient as a whole? Clearly there are problems when it comes to measuring history taking and this has already been referred to. By highlighting the construct and content validity, i.e. increasing the number and construct of the items on history taking to be measured, these problems could be expected to be improved by the OSLER. In addition to content validity, overall construct or face validity will be improved by ensuring that all 10 items are formally assessed in a structured manner. It can and does happen that, in existing assessment methods, some item(s) receive undue attention to the exclusion of others. By having a fixed number of items to be measured, examiners will not have to generalize from what they have assessed to what they should have assessed, as is frequently the situation.

Most assessment innovations run into problems of practicality in terms of organizational logistics. The OSLER is singularly unaffected in this respect. Indeed it could be described as organizational friendly as the organization is identical to existing practices. The OSLER could also be described as examiner friendly in that it assists the examiner as an 'aide-memoire' in reminding him/her to consistently cover the same general areas for all candidates to be assessed. The provision of a checklist of items for the

long case was suggested over 20 years ago (Fleming *et al.*, 1974) as a reasonable approach. This in turn makes it candidate friendly in that the assessment will be regarded as more fair by the candidates. There are also a number of other advantages associated with usage of the OSLER that should make it potentially more acceptable. Since it is essentially in line with the traditional long case, it fulfils the innovational criteria already referred to, thus making it more acceptable to more conservative forces. The same number of examiners are required and the examiner time is identical. Structured examinations are frequently criticized by examiners, who feel that their 'independence' is in some way interfered with. In strict educational and institutional objective terms such a stand would be unacceptable: however, the fact remains that such a view is strongly held by a significant number. The OSLER has attractions for such situations in that it allows the examiner to continue to operate as before. The examiner continues to exercise his/her independence, particularly through the item on clinical acumen. However, it will be obvious that the grade in this area will have to correlate with the grades recorded in the other nine items. The feedback potential already referred to is obvious in terms of identification of clinical skills defects.

If one accepts that for clinical assessment to be truly valid it must be patient centred then it would seem reasonable to conclude that the long case is going to be maintained to a greater or lesser extent in the short to medium term at least. Instead of bemoaning this fact, an effort should be made to maximize its potential while at the same time minimizing its faults until such time as a method emerges which will allow full observation of the candidate during the long case. The OSLER as described is both examiner (user) friendly and candidate friendly and could be implemented with relatively little effort. The small extra effort required to implement it would be offset by the more detailed data obtained on candidates' performances rather than the more frequent global data currently available. The perfect method for the assessment of clinical competence has to date not been developed and for reasons of practicality will not be available in the foreseeable future. Until such time, some improvement in the long case is necessary. The OSLER is suggested as that improvement.

Note on contributor

PROFESSOR GLEESON is Associate Professor of Medicine at the Royal College of Surgeons in Ireland and Consultant Physician/Gastroenterologist at James Connolly Memorial Hospital, Dublin. He has had a long standing interest in the assessment of clinical competence and was closely involved in the development of the Objective Structured Clinical Examination (OSCE).

References

BARROWS, H.S. & FELTOUICH, D.J. (1987) The clinical reasoning process, *Medical Education*, 21, pp. 86–91.

CASSIRER, J.P. (1992) Clinical problem-solving—a new feature in the journal, *New England Journal of Medicine*, 326, pp. 60–61.

CHAN YAN, C., GILLIES, J., REUDY,. J.H., MONTANER, J.S.G. & MARSHALL, S.A. (1988) Clinical skills of medical residents: a review of physical examination, *Canadian Medical Association Journal*, 139, pp. 629–632.

COLLINGWOOD, V. (1979) Planning of innovation in higher education, *Programmed Learning and Educational Technology* 16, pp. 8–15.

DOHERTY, E., O'BOYLE, C.A., SHANNON, W., MCGEE, H. & BURY, G.(1990) Communication skills training in undergraduate medicine, *Irish Medical Journal*, 83, pp. 54–56.

ELSTEIN, A.S., SHULMAN, L.S. & SPRAFKA, S.A. (1978) *Medical Problem Solving—An Analysis of Clinical Reasoning* (Cambridge, MA and London, UK, Harvard University Press).

ENDE, J. (1983) Feedback in clinical medical education, *Journal of the American Medical Association*, 250, pp. 777–781.

FLEMING, P.R., MANDERSON, W.G., MATTHEWS, M.B., SANDERSON, P.H. & STOKES, J.F. (1974) Evolution of an examination: MRCP(UK), *British Medical Journal*, 2, pp. 99–106.

FLEMING, P.R., SANDERSON P.H., STOKES, J.F. & WALTON, H.J. (1976) *Examinations in Medicine*. (Edinburgh, Churchill Livingstone).

GENERAL MEDICAL COUNCIL (1993) *Tomorrow's Doctors—Recommendations on Undergraduate Medical Education* (London, General Medical Council).

GLEESON, F. (1992) Defects in postgraduate clinical skills as revealed by the Objective Structured Long Examination Record (OSLER), *Irish Medical Journal*, 85, pp. 11–14.

GLEESON, F. (1995) The effect of immediate feedback on clinical skills using the OSLER, in: *Proceedings of the Sixth Ottawa Conference* (Toronto, University of Toronto Bookstore Custom Publishing).

HAMPTON, J.R., HARRISON, M.J.G., MITCHELL, J.R.A., PRICHARD J.S. & SEYMOUR, C. (1975) Relative contributions of history-taking, physical examination and laboratory investigations to diagnosis and management of medical out-patients, *British Medical Journal*, 2, pp. 486–489.

HARDEN, R.M. & GLEESON, F.A. (1979) ASME Medical Education Booklet No. 8: Assessment of Medical Competence Using an Objective Structured Clinical Examination (OSCE), *Medical Education*, 13, pp. 39–54.

IRWIN, W.G., MCCELLAND, R. & LOVE A.H.G. (1989) Communication skills training for medical students: an integrated approach, *Medical Education*, 23, pp. 387–394.

Lancet (1989) Editorial: Solving problems taking decisions, *Lancet* ii, p. 1306.

LOWRY, S. (1993) *Medical Education* (London, BMJ).

LUDBROOK, J. & MARSHALL, V.R. (1971) Examiner training for clinical examinations, *British Journal of Medical Education*, 5, pp. 152–155.

MAGUIRE, G.P, & RUTTER, D.R. (1976) History taking for medical students: deficiencies in performance, *Lancet* 3, pp. 556–558.

MCMANUS, I.C., VINCENT, C.A., THOM, S. & KIDD, J. (1993) Teaching communication skills to clinical students, *British Medical Journal*, 306, pp. 1322–1327.

MIALL, L.S. & DAVIES, H. (1992) An analysis of paediatric decision-making: how should students be taught?, *Medical Education*, 26, pp. 317–320.

NEWBLE, D.I. (1991) The observed long case in clinical assessment, *Medical Education*, 25, pp. 369–373.

PRICE, J. & BYRNE, G.J.A. (1994) The direct clinical examination: an alternative method for the assessment of clinical psychiatry skills in undergraduate medical students, *Medical Education*, 28, pp. 120–125.

SOX, H.C., MORGAN, W.L., NEUFELD, U.R., SHELDON, G.F. & TONESK, X. (1985) Subgroup, report on clinical skills, *Journal of Medical Education*, 59, pp. 139–147.

STOKES, J. (1972) *The Clinical Examination—Assessment of Clinical Skills*, Medical Education Booklet No. 2. (Dundee, Association for the Study of Medical Education).

VAN DER VLEUTEN (1996) Making the best of the 'long case'; *Lancet*, 347, pp. 704–705.

VAN THIEL, J., KRAAN, H.F. & VAN DER VLEUTEN, C.P.M. (1991) Reliability and feasibility of measuring medical interviewing skills: the revised Maastricht history-taking and advice checklist, *Medical Education*, 25, pp. 224–229.

WEATHERALL, D.J. (1991) Examining undergraduate examinations, *Lancet*, 338, pp. 37–39.

WIENER, S. & NATHANSON, M. (1976) Physical examination: frequently observed errors, *Journal of the American Medical Association*, 236, pp. 852–855.

WILSON, G.M., LEVER, R., HARDEN, R.McG., ROBERTSON, J.L.S. & MACRITCHIE, J. (1969) Examination of clinical examiners, *Lancet* i, pp. 37–40.

WRAY, N.P. & FRIEDLAND J.A. (1983) Detection and correction of house staff error in physical diagnosis, *Journal of the American Medical Association* 249, pp. 1035–1037.

16 March 2017

Portfolio

The 21st century assessment tool

LERTBUNNAPHONG T, M.D.

รู้จัก Portfolio มั้ย

My daughter story

The story of Portfolio

เก่ง และ ดี

The story of Portfolio

21st century

What is PORTFOLIO?

Collection of evidences

What is PORTFOLIO?

Evaluation of outcomes

What is PORTFOLIO?

Make evolution!!!

4 ขั้นตอนในการพัฒนา **Portfolio**

1. กำหนดเป้าหมาย
2. กำหนดหลักฐาน
3. สะท้อนความรู้สึก ปรับเปลี่ยนทัศนคติ
4. ประเมินผล

4 ขั้นตอนในการพัฒนา **Portfolio**

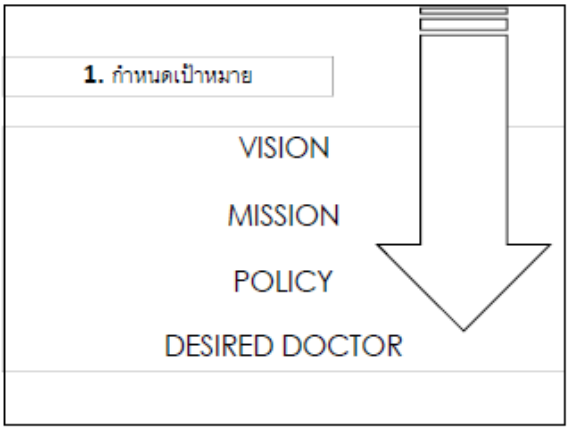
1. กำหนดเป้าหมาย
2. กำหนดหลักฐาน

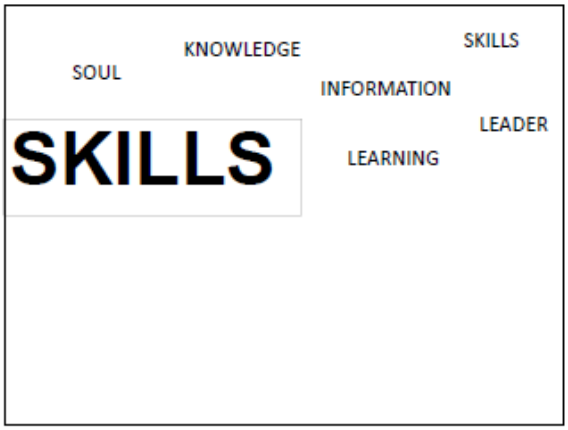
แต่ละกลุ่มกำหนดเป้าหมายและหลักฐานใน **Portfolio**
(เวลา **10** นาที)

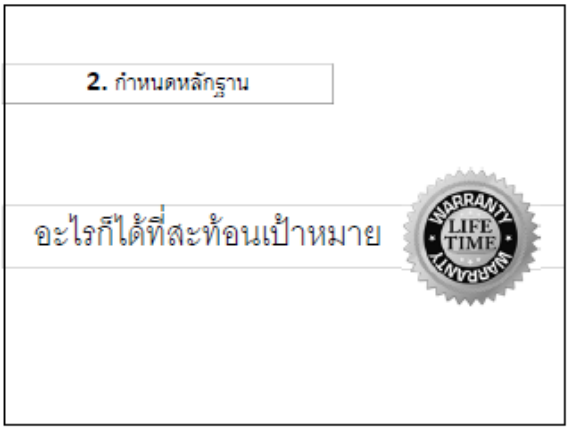
นักศึกษาแพทย์	แพทย์ใช้ทุน	แพทย์ประจำบ้าน

Who is your choice; Dr.A or Dr.B?

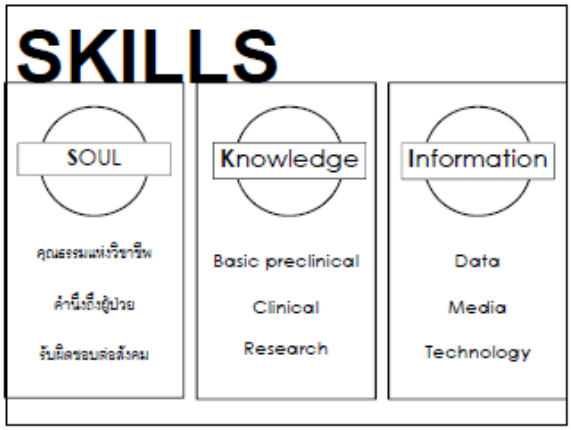
WHY?

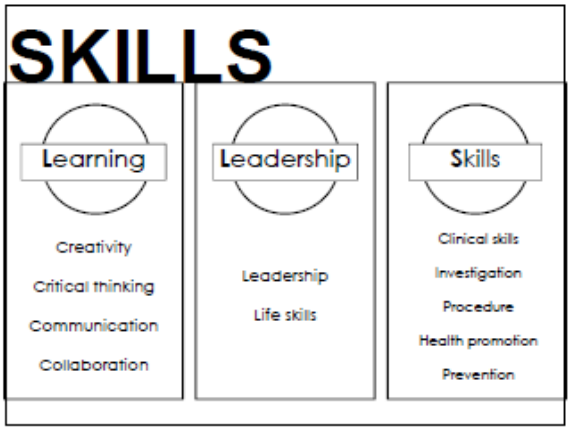

1. กำหนดเป้าหมาย
VISION
MISSION
POLICY
DESIRED DOCTOR


SOUL
KNOWLEDGE
SKILLS
INFORMATION
LEADER
SKILLS
LEARNING


2. กำหนดหลักฐาน
อะไรก็ได้ที่สะท้อนเป้าหมาย
WARRANTY
LIFE TIME
WARRANTY

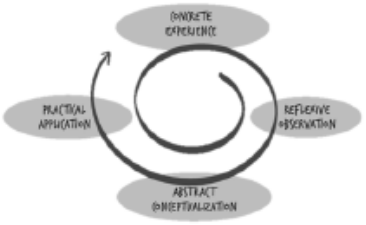

Platinum Level


SKILLS
SOUL
คุณธรรมแห่งวิชาชีพ
คำนึงถึงผู้ป่วย
รับผิดชอบต่อสังคม
Knowledge
Basic preclinical
Clinical
Research
Information
Data
Media
Technology


SKILLS
Learning
Creativity
Critical thinking
Communication
Collaboration
Leadership
Leadership
Life skills
Skills
Clinical skills
Investigation
Procedure
Health promotion
Prevention

3. สะท้อนความรู้สึก ปรับเปลี่ยนทัศนคติ

CRITICAL REFLECTION



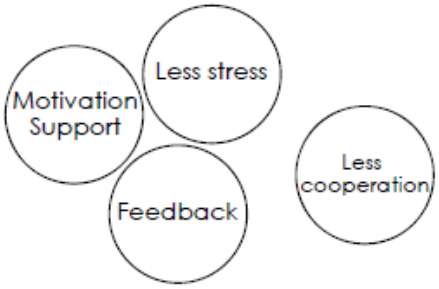
Learning without reflection is **waste**

Confucius

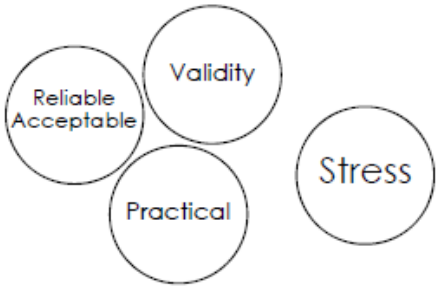
4.ประเมินผล

Summative VS Formative

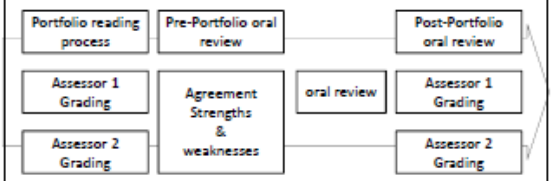
Formative evaluation

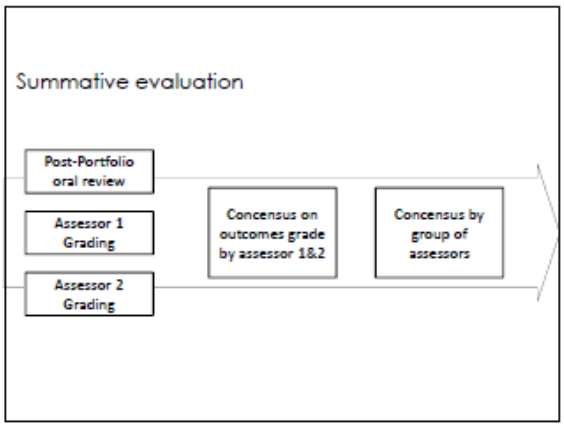


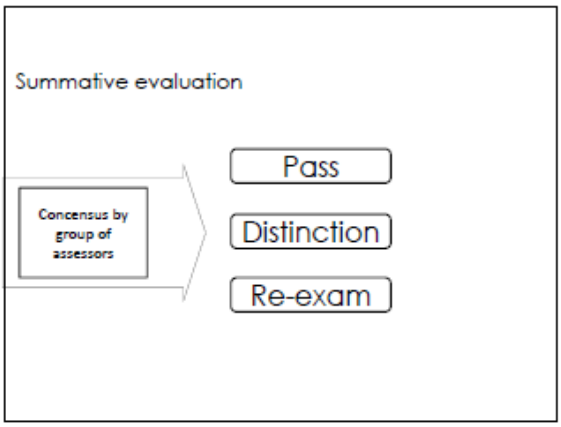
Summative evaluation

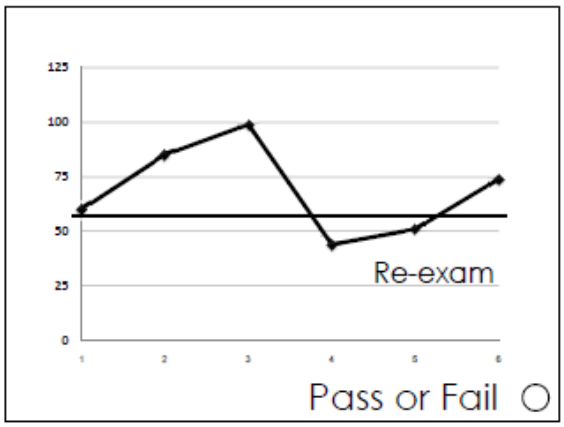


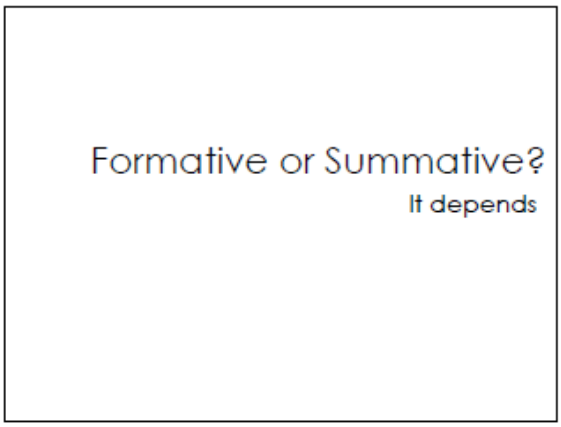
Summative evaluation

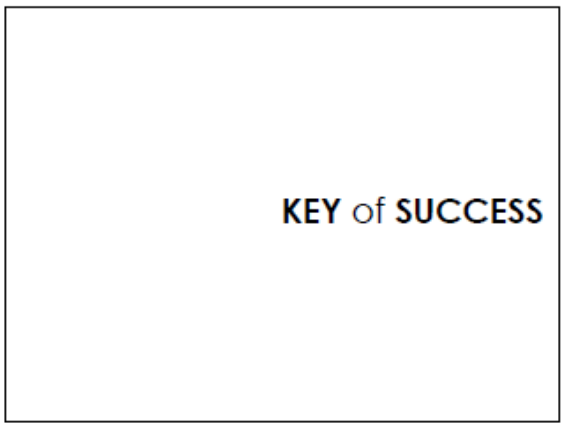



Summative evaluation
Post-Portfolio oral review
Assessor 1 Grading
Assessor 2 Grading
Concensus on outcomes grade by assessor 1&2
Concensus by group of assessors


Summative evaluation
Concensus by group of assessors
Pass
Distinction
Re-exam


125
100
75
50
25
0
Re-exam
Pass or Fail


Formative or Summative?
It depends


KEY of SUCCESS


organization supports
×
medical teachers
×
medical students

Outcomes
Evidences
Reflection
Assessment

There are great strengths in each assessment

once correct one is selected for each outcome

John Dent
ESME online course 2012

ผลลัพธ์การปฏิบัติงานของ

นายแพทย์ X

อาจารย์ที่ปรึกษา อาจารย์ A

ตามการประเมินด้วยแฟ้มสะสมพัฒนาการ (Portfolio)

ปีการศึกษา 2554-2556

Competency based portfolio assessment

Academic year 2011-2013

สาส์นจากหัวหน้าภาควิชา

ภาควิชาสูติศาสตร์-นรีเวชวิทยา คณะแพทยศาสตร์ศิริราชพยาบาล มหาวิทยาลัยมหิดล ขอแสดงความยินดีกับ**นายแพทย์** A ที่สำเร็จการฝึกอบรมแพทย์ประจำบ้าน สาขาสูติศาสตร์-นรีเวชวิทยา ระหว่างปีการศึกษา 2553-2555

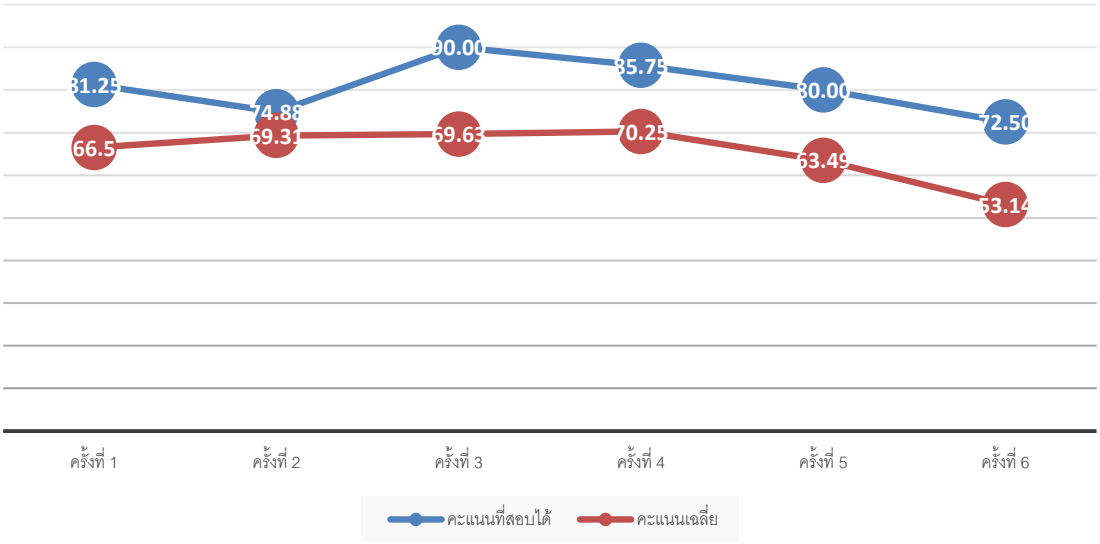
ตลอดระยะเวลาสามปีที่ผ่านมา ภาควิชาฯ ได้ดำเนินการประเมินคุณสมบัติด้านต่างๆ ของท่าน ได้แก่ ความรู้ ทักษะหัตถการ การวิจัย และพฤติกรรมการปฏิบัติงาน ในรูปแบบ Portfolio ดังผลสรุปในเอกสารฉบับนี้

ภาควิชาฯ ขออำนวยพรให้ท่านประสบความสำเร็จในการดำเนินชีวิตครอบครัว และหน้าที่การงานตลอดไป

ศาสตราจารย์คลินิก นายแพทย์ชาญชัย วันทนาศิริ
หัวหน้าภาควิชาสูติศาสตร์-นรีเวชวิทยา
คณะแพทยศาสตร์ศิริราชพยาบาล
มหาวิทยาลัยมหิดล

การประเมินความรู้ทางสูติศาสตร์-นรีเวชวิทยา

(Knowledge assessment)



ชั้นปีที่ 1

การสอบครั้งที่	คะแนนรวม	คะแนนเฉลี่ย	คะแนนที่สอบได้	ลำดับที่
1	100	66.50	81.25	1
2	100	69.31	74.88	3
3	100	69.63	90.00	1
4	100	70.25	85.75	1
5	100	63.49	80.00	1
6	100	53.14	72.50	2

ผลการสอบตามหลักสูตรประกาศนียบัตรบัณฑิตชั้นสูงสาขาวิทยาศาสตร์การแพทย์คลินิก:

The Higher Graduate Diploma (Clinical Medical Sciences) **คณะแพทยศาสตร์ศิริราชพยาบาล**

☒ผ่าน ได้รับประกาศนียบัตรเมื่อ 25 พฤษภาคม 2555

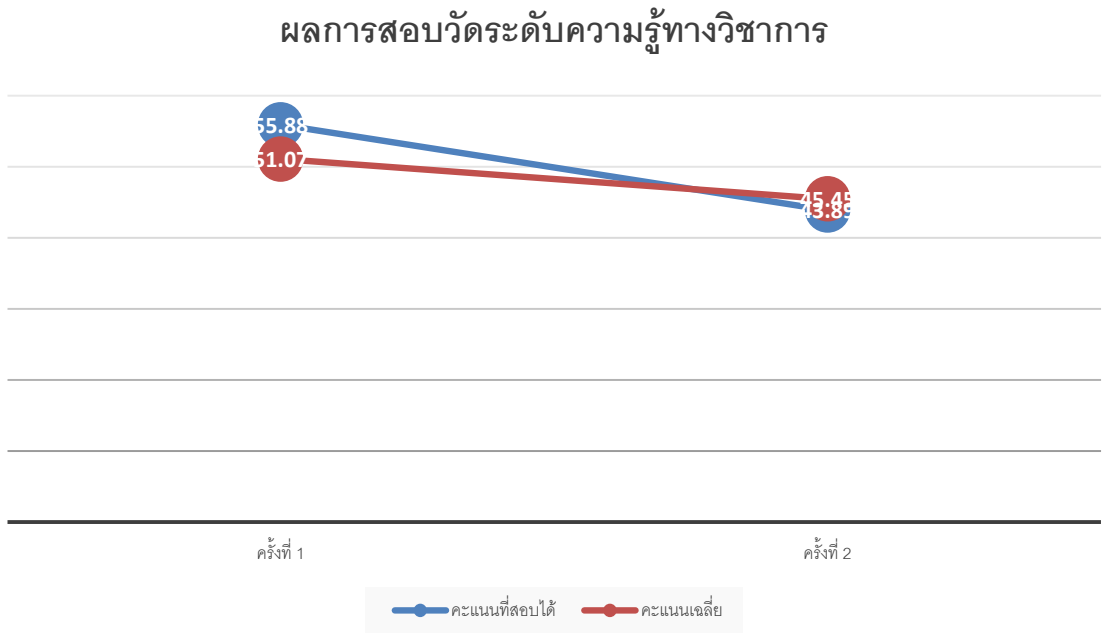
☐ไม่ผ่าน

การสอบวิชาภาษาอังกฤษและกฎหมายทางการแพทย์ครั้งที่ 1

☒**ผ่าน** ☐**ไม่ผ่าน**

การสอบวิชาภาษาอังกฤษและกฎหมายทางการแพทย์ครั้งที่ 2 **(กรณีสอบไม่ผ่านครั้งแรก)**

☐**ผ่าน** ☐**ไม่ผ่าน**



ชั้นปีที่ 2

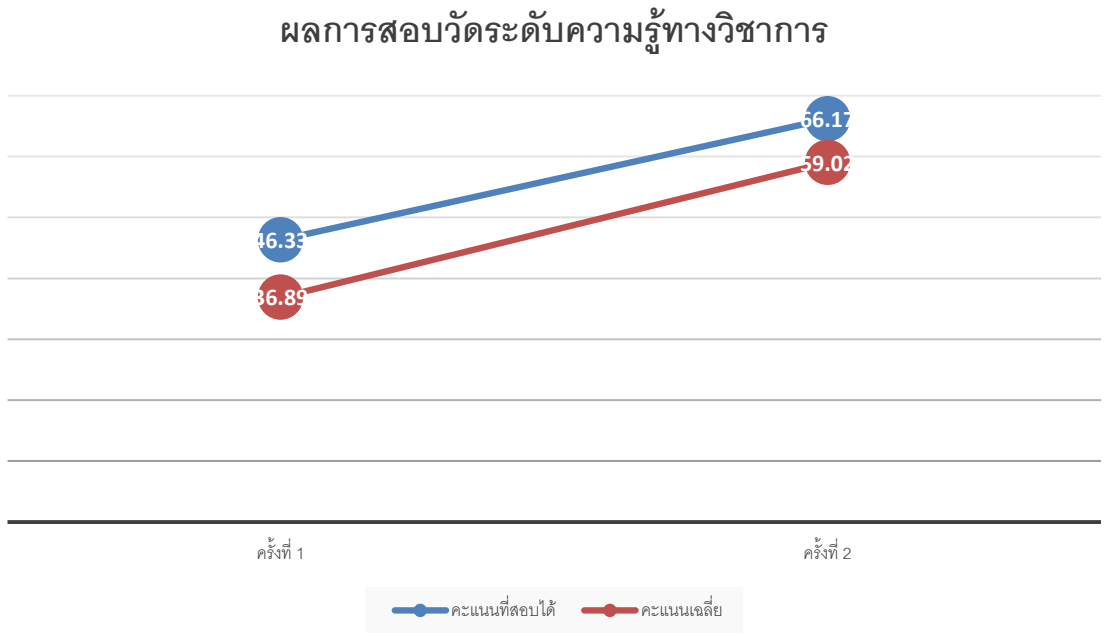
การสอบครั้งที่	คะแนนรวม	คะแนนเฉลี่ย	คะแนนที่สอบได้	ลำดับที่
1	100	51.07	55.88	5
2	100	45.45	43.89	10

การสอบ OSLER ในสถาบัน ครั้งที่ 1

☒**ผ่าน** ☐**ไม่ผ่าน**

การสอบ Basic science ราชวิทยาลัยสูตินรีแพทย์แห่งประเทศไทย

☒**ผ่าน** ☐**ไม่ผ่าน**



ชั้นปีที่ 3

การสอบครั้งที่	คะแนนรวม	คะแนนเฉลี่ย	คะแนนที่สอบได้	ลำดับที่
1	100	36.89	46.33	2
2	100	59.02	66.17	1

การสอบ OSLER ในสถาบัน ครั้งที่ 2

☒ผ่าน ☐ไม่ผ่าน

การสอบ OSLER ราชวิทยาลัยสูตินรีแพทย์แห่งประเทศไทยครั้งที่ 1

☒ผ่าน ☐ไม่ผ่าน

การสอบ OSLER ราชวิทยาลัยสูตินรีแพทย์แห่งประเทศไทยครั้งที่ 2 (กรณีสอบครั้งแรกไม่ผ่าน)

☐ผ่าน ☐ไม่ผ่าน

การสอบงานวิจัย ราชวิทยาลัยสูตินรีแพทย์แห่งประเทศไทย

☒ผ่าน ☐ไม่ผ่าน

หัตถการสำคัญทางสูติศาสตร์-นรีเวชวิทยาที่ปฏิบัติ ขณะเป็นแพทย์ประจำบ้านชั้นปีที่ 3

(Clinical skills assessment when being the 3rd year resident)

การผ่าตัดทางนรีเวช

การผ่าตัด	จำนวน
Total abdominal hysterectomy +/- bilateral salpingooophorectomy	19
Vaginal hysterectomy +/- AP repair	4
Adnexal surgery: Salpingectomy/Salpingotomy/Salpingostomy	21
Cervical conization	11

การผ่าตัดทางสูติศาสตร์

การผ่าตัด	จำนวน
Cesarean delivery	55
Tubal sterilization	3
Dilatation and curettage	16
Vacuum extraction/Forceps extraction	4
Breech assisting	
Manual removal of placenta	2

หมายเหตุ

จำนวนหัตถการเป็นจำนวนโดยประมาณ เนื่องจากอยู่ระหว่างกระบวนการพัฒนาและปรับปรุงระบบเก็บข้อมูลหัตถการแพทย์ประจำบ้าน ภาควิชาสูติศาสตร์-นรีเวชวิทยา

การทำงานวิจัยระดับแพทย์ประจำบ้าน
(Research competency)

เรื่อง Prevalence and Associating Factors of Sexual Dysfunction in Women Who Use Intrapartum Device (IUD)

อาจารย์ผู้ควบคุมผู้ช่วยศาสตราจารย์นายแพทย์ธันยารัตน์ วงศ์วนานุรักษ์

ข้อมูลสำคัญสำหรับงานวิจัย

1. ผ่าน SIRB เมื่อ 21 กุมภาพันธ์ 2555
   เลขที่ 813/2554 (EC3)

2. ประกวดการนำเสนองานวิจัยในการประชุมราชวิทยาลัยสูตินรีแพทย์แห่งประเทศไทย
   วันที่ 26 พฤศจิกายน 2556
   ☐ เข้าร่วมนำเสนอ ไม่ได้รับรางวัล
   ☒ เข้าร่วมนำเสนอ ได้รับรางวัล ชมเชย

3. การตีพิมพ์ในวารสารวิชาการ
   ☐ ไม่ได้ตีพิมพ์
   ☒ ได้รับการตีพิมพ์ (ระบุรายละเอียดวารสาร) J Med Assoc Thai 2014
   Full text. E-Journal: http://Jmatonline.com

ผลการประเมินเจตคติและพฤติกรรมการปฏิบัติงานของแพทย์ประจำบ้าน

(Multisources feedback)

แพทย์ประจำบ้านจะได้รับการประเมินในประเด็นต่อไปนี้

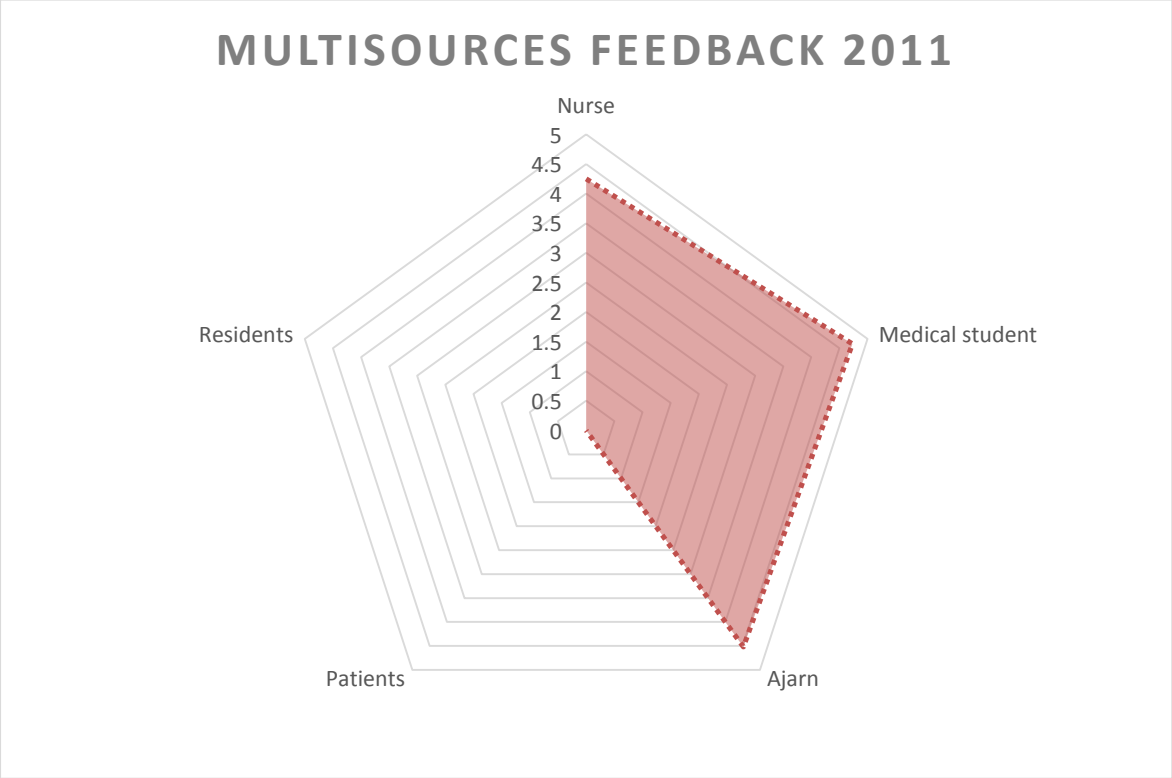
1.ความรู้ความสามารถด้านวิชาการ

2.ทักษะพื้นฐานในการปฏิบัติงาน

ได้แก่ ทักษะการสื่อสารกับเพื่อนร่วมงานและผู้ป่วย/ญาติ การบันทึกรายงานผู้ป่วย การทำงานร่วมกับผู้อื่น และบุคลิกภาพขณะปฏิบัติงาน

3.คุณธรรมและจริยธรรม

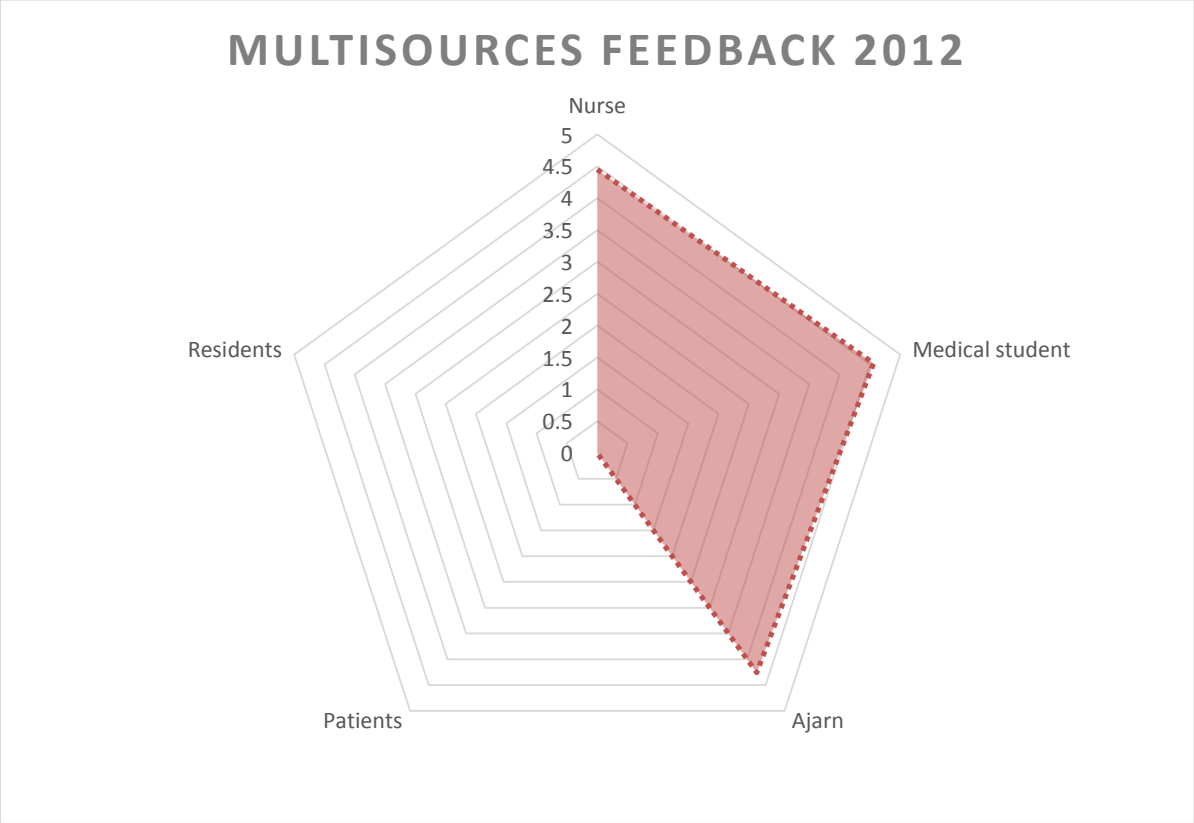
ได้แก่ ความรับผิดชอบ ความเสียสละ ความตรงต่อเวลา ความซื่อสัตย์ การปฏิบัติตามระเบียบข้อบังคับ และอัธยาศัย/น้ำใจ/ความเอื้อเฟื้อต่อผู้อื่น



ชั้นปีที่ 1ปีการศึกษา 2554

Rotation	อาจารย์ (5 คะแนน)	นักศึกษาแพทย์ (5 คะแนน)	พยาบาล (5 คะแนน)	แพทย์ประจำบ้าน	ผู้รับบริการ
พระศรีฯ 9/2			4.61		
LR เช้า		5.00	3.76		
LR พิเศษเช้า			4.61		
นรีเวช 1	4.90	5.00	4.00		
นรีเวช 1 (2)	4.50	4.90	4.00		
พระศรีฯ 10/2			4.46		
พระศรีฯ 9/1+ANC			5.00		
LR ดึก			4.00		
LR พิเศษบ่าย			4.30		
นรีเวช 2	4.20	4.50	4.56		
Onco	4.50	4.30	3.84		
พระศรีฯ 10/3		5.00	4.30		
พระศรีฯ 10/1		4.46	3.91		
คะแนนเฉลี่ย	**4.52**	**4.73**	**4.25**		

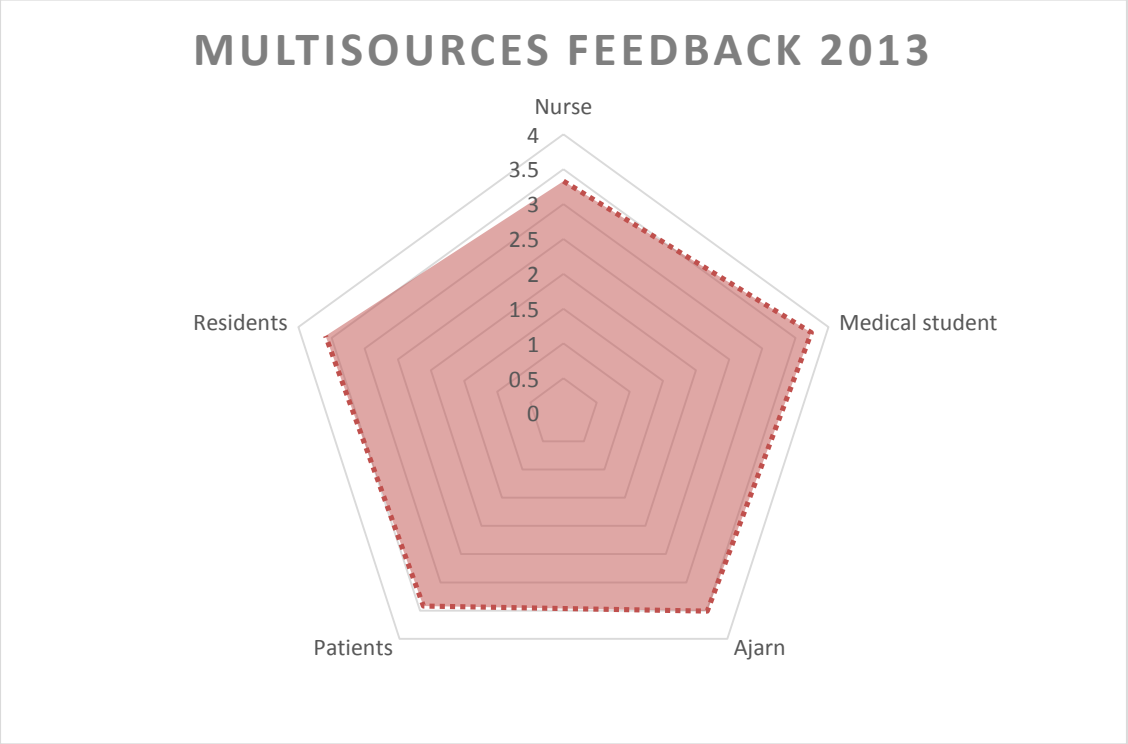
*ยังไม่มีการประเมินจากแพทย์ประจำบ้านและผู้รับบริการในปีการศึกษา 2554



ชั้นปีที่ 2 ปีการศึกษา 2555

Rotation	อาจารย์ (5 คะแนน)	นักศึกษาแพทย์ (5 คะแนน)	พยาบาล (5 คะแนน)	แพทย์ประจำบ้าน	ผู้รับบริการ
นรีเวช 1	3.93	4.00	4.53		
เลิดสิน	4.67				
พระศรีฯ 13/1	4.35		4.61		
LR ดึก			4.00		
Onco	4.17	4.20	3.23		
พระศรีฯ 14/2			5.00		
นรีเวช 2	4.11	4.50	5.00		
สระบุรี	4.47				
พระศรีฯ 13/2	4.40		4.70		
พระศรีฯ 10/1		4.70	4.50		
พระศรีฯ 14/1	4.00		4.23		
LR เช้า		5.00	4.69		
พระศรีฯ 10/3		5.00	4.56		
คะแนนเฉลี่ย	**4.26**	**4.56**	**4.45**		

*ยังไม่มีการประเมินจากแพทย์ประจำบ้านและผู้รับบริการในปีการศึกษา 2555



ชั้นปีที่ 3 ปีการศึกษา 2556

Rotation	อาจารย์ (4 คะแนน)	แพทย์ประจำบ้าน (4 คะแนน)	พยาบาล (4 คะแนน)	นักศึกษาแพทย์ (4 คะแนน)	ผู้รับบริการ (4 คะแนน)
นรีเวช 1	3.50	3.80	3.40	3.90	3.03
STD	3.70		3.20		
พระศรีฯ 10/1		2.62	4.00	3.70	3.36
LR พิเศษ		3.90	3.08		
OPD GYN			2.90		3.40
Septic		3.75	3.10	4.00	3.26
วิสัญญี	3.75				
นรีเวช 2	3.90	4.00	3.85	3.87	3.74
Infertile	3.20				
นครปฐม	3.00				
OPD ANC			3.75		3.73
ONCO	3.60	3.81	3.02		
LR เช้า		3.25	3.08	3.50	
Surgery	3.47				
คะแนนเฉลี่ย	**3.51**	**3.59**	**3.33**	**3.74**	**3.42**

***เริ่มการประเมินจากนักศึกษาแพทย์และผู้รับบริการ ในปีการศึกษา 2556**

ผลลัพธ์การปฏิบัติงานของ

แพทย์หญิง Y

อาจารย์ที่ปรึกษา อาจารย์ B

ตามการประเมินด้วยแฟ้มสะสมพัฒนาการ (Portfolio)

ปีการศึกษา 2554-2556

Competency based portfolio assessment

Academic year 2011-2013

สาส์นจากหัวหน้าภาควิชา

ภาควิชาสูติศาสตร์-นรีเวชวิทยา คณะแพทยศาสตร์ศิริราชพยาบาล มหาวิทยาลัยมหิดล ขอแสดงความยินดีกับ**แพทย์หญิง** B ที่สำเร็จการฝึกอบรมแพทย์ประจำบ้าน สาขาสูติศาสตร์-นรีเวชวิทยา ระหว่างปีการศึกษา 2553-2555

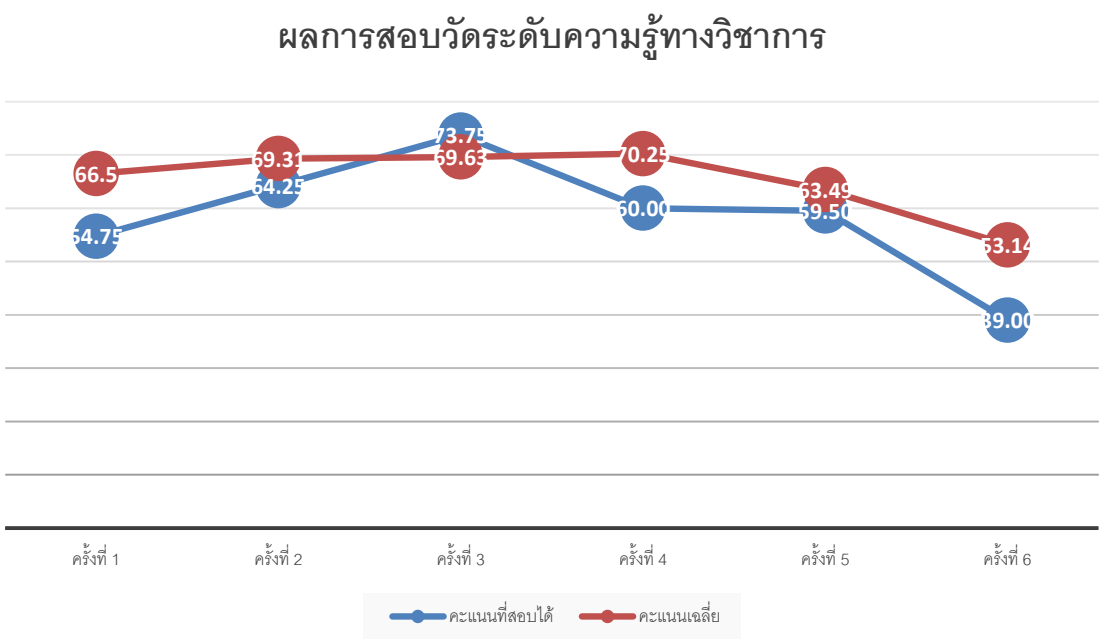
ตลอดระยะเวลาสามปีที่ผ่านมา ภาควิชาฯ ได้ดำเนินการประเมินคุณสมบัติด้านต่างๆ ของท่าน ได้แก่ ความรู้ ทักษะหัตถการ การวิจัย และพฤติกรรมการปฏิบัติงาน ในรูปแบบ Portfolio ดังผลสรุปในเอกสารฉบับนี้

ภาควิชาฯ ขออำนวยพรให้ท่านประสบความสำเร็จในการดำเนินชีวิตครอบครัว และหน้าที่การงานตลอดไป

ศาสตราจารย์คลินิก นายแพทย์ชาญชัย วันทนาศิริ
หัวหน้าภาควิชาสูติศาสตร์-นรีเวชวิทยา
คณะแพทยศาสตร์ศิริราชพยาบาล
มหาวิทยาลัยมหิดล

การประเมินความรู้ทางสูติศาสตร์-นรีเวชวิทยา

(Knowledge assessment)



ชั้นปีที่ 1

การสอบครั้งที่	คะแนนรวม	คะแนนเฉลี่ย	คะแนนที่สอบได้	ลำดับที่
1	100	66.50	54.75	13
2	100	69.31	64.25	11
3	100	69.63	73.75	4
4	100	70.25	60.00	13
5	100	63.49	59.50	11
6	100	53.14	39.00	13

ผลการสอบตามหลักสูตรประกาศนียบัตรบัณฑิตชั้นสูงสาขาวิทยาศาสตร์การแพทย์คลินิก:
The Higher Graduate Diploma (Clinical Medical Sciences) **คณะแพทยศาสตร์ศิริราชพยาบาล**

☒ผ่าน ได้รับประกาศนียบัตรเมื่อ 25 พฤษภาคม 2555

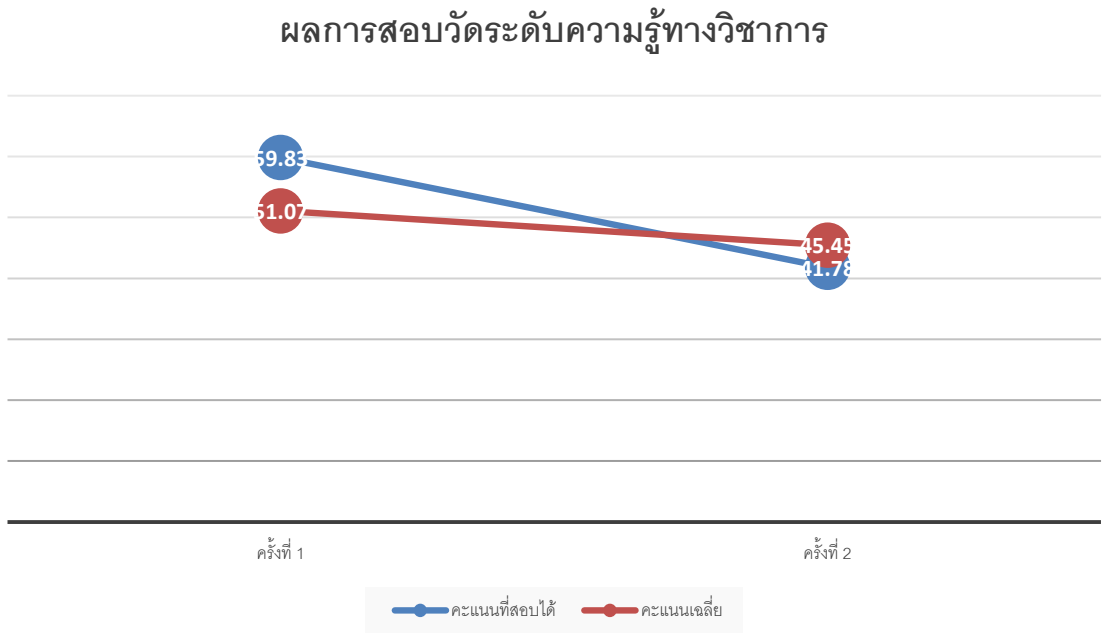
☐ไม่ผ่าน

การสอบวิชาภาษาอังกฤษและกฎหมายทางการแพทย์ครั้งที่ 1

☒ผ่าน ☐ไม่ผ่าน

การสอบวิชาภาษาอังกฤษและกฎหมายทางการแพทย์ครั้งที่ 2 **(กรณีการสอบครั้งที่** 1 **ไม่ผ่าน)**

☐ผ่าน ☐ไม่ผ่าน



ชั้นปีที่ 2

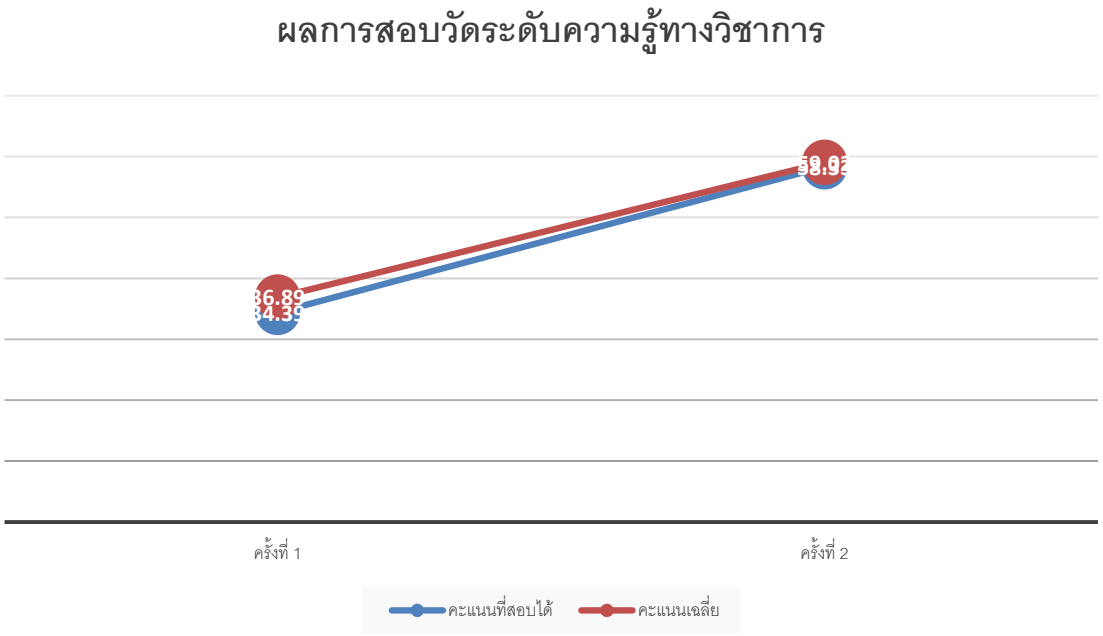
การสอบครั้งที่	คะแนนรวม	คะแนนเฉลี่ย	คะแนนที่สอบได้	ลำดับที่
1	100	51.07	59.83	4
2	100	45.45	41.78	12

การสอบ OSLER ในสถาบัน ครั้งที่ 1

☒**ผ่าน** ☐**ไม่ผ่าน**

การสอบ Basic science ราชวิทยาลัยสูตินรีแพทย์แห่งประเทศไทย

☒**ผ่าน** ☐**ไม่ผ่าน**



ชั้นปีที่ 3

การสอบครั้งที่	คะแนนรวม	คะแนนเฉลี่ย	คะแนนที่สอบได้	ลำดับที่
1	100	36.89	34.39	10
2	100	59.02	58.33	10

การสอบ OSLER ในสถาบัน ครั้งที่ 2

☒ผ่าน ☐ไม่ผ่าน

การสอบ OSLER ราชวิทยาลัยสูตินรีแพทย์แห่งประเทศไทยครั้งที่ 1

☐ผ่าน ☒ไม่ผ่าน

การสอบ OSLER ราชวิทยาลัยสูตินรีแพทย์แห่งประเทศไทยครั้งที่ 2 (กรณีสอบครั้งแรกไม่ผ่าน)

☒ผ่าน ☐ไม่ผ่าน

การสอบงานวิจัย ราชวิทยาลัยสูตินรีแพทย์แห่งประเทศไทย

☒ผ่าน ☐ไม่ผ่าน

หัตถการสำคัญทางสูติศาสตร์-นรีเวชวิทยาที่ปฏิบัติ ขณะเป็นแพทย์ประจำบ้านชั้นปีที่ 3

(Clinical skills assessment when being the 3rd year resident)

การผ่าตัดทางนรีเวช

การผ่าตัด	จำนวน
Total abdominal hysterectomy +/- bilateral salpingooophorectomy	14
Vaginal hysterectomy +/- AP repair	7
Adnexal surgery: Salpingectomy/Salpingotomy/Salpingostomy	4
Cervical conization	2

การผ่าตัดทางสูติศาสตร์

การผ่าตัด	จำนวน
Cesarean delivery	43
Tubal sterilization	1
Dilatation and curettage	5
Vacuum extraction/Forceps extraction	5
Breech assisting	
Manual removal of placenta	6

หมายเหตุ

จำนวนหัตถการเป็นจำนวนโดยประมาณ เนื่องจากอยู่ระหว่างกระบวนการพัฒนาและปรับปรุงระบบเก็บข้อมูลหัตถการแพทย์ประจำบ้าน ภาควิชาสูติศาสตร์-นรีเวชวิทยา

การทำงานวิจัยระดับแพทย์ประจำบ้าน

(Research competency)

เรื่อง Prevalence of Abnormal Menstrual Patterns among Copper Intrauterine Devices (IUDs)Users in Women Attending Family Planning Clinic, Siriraj Hospital

อาจารย์ผู้ควบคุม ผู้ช่วยศาสตราจารย์นายแพทย์สุรศักดิ์ อังสุวัฒนา

ข้อมูลสำคัญสำหรับงานวิจัย

1. **ผ่าน** SIRB **เมื่อ**28 สิงหาคม 2555
   เลขที่ 415/2555(EC3)

2. **ประกวดการนำเสนองานวิจัยในการประชุมราชวิทยาลัยสูตินรีแพทย์แห่งประเทศไทย**
   วันที่ 26 พฤศจิกายน 2556
   ☒**เข้าร่วมนำเสนอ ไม่ได้รับรางวัล**
   ☐**เข้าร่วมนำเสนอ ได้รับรางวัล**

3. **การตีพิมพ์ในวารสารวิชาการ**
   ☒**ไม่ได้ตีพิมพ์**
   ☐**ได้รับการตีพิมพ์ (ระบุรายละเอียดวารสาร)**

ผลการประเมินเจตคติและพฤติกรรมการปฏิบัติงานของแพทย์ประจำบ้าน

(Multisources feedback)

แพทย์ประจำบ้านจะได้รับการประเมินในประเด็นต่อไปนี้

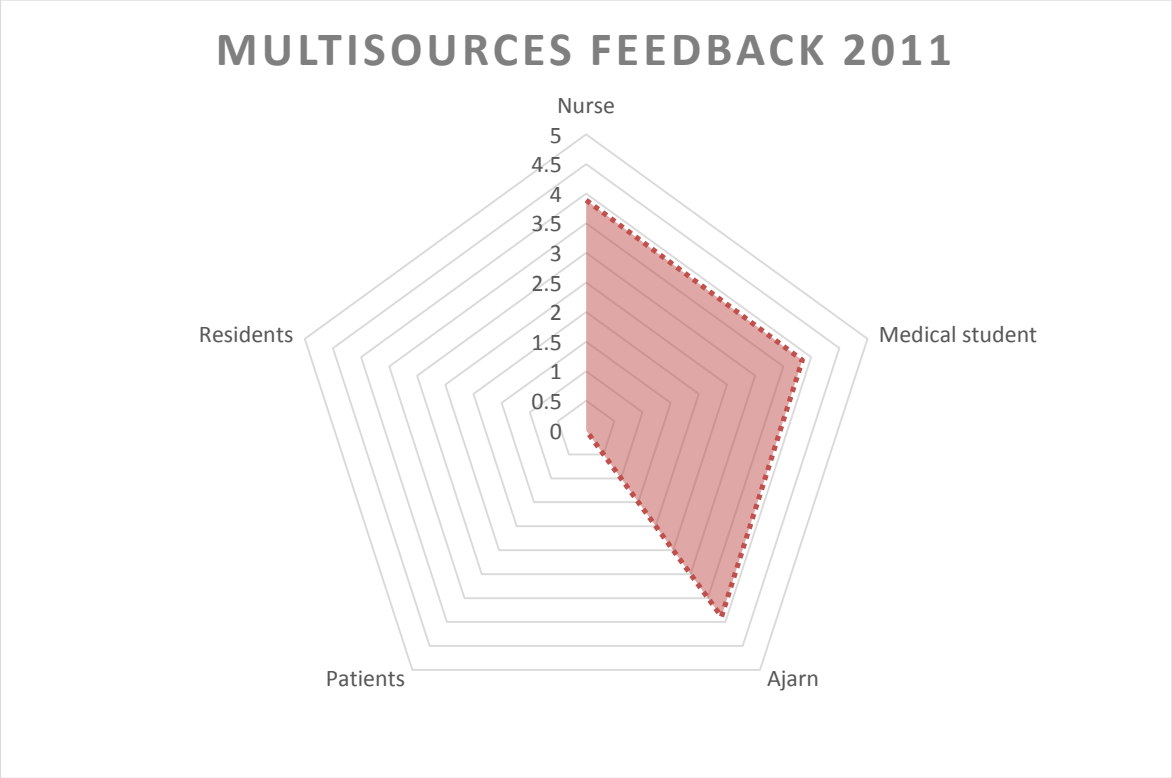
1.ความรู้ความสามารถด้านวิชาการ

2.ทักษะพื้นฐานในการปฏิบัติงาน

ได้แก่ ทักษะการสื่อสารกับเพื่อนร่วมงานและผู้ป่วย/ญาติ การบันทึกรายงานผู้ป่วย การทำงานร่วมกับผู้อื่น และบุคลิกภาพขณะปฏิบัติงาน

3.คุณธรรมและจริยธรรม

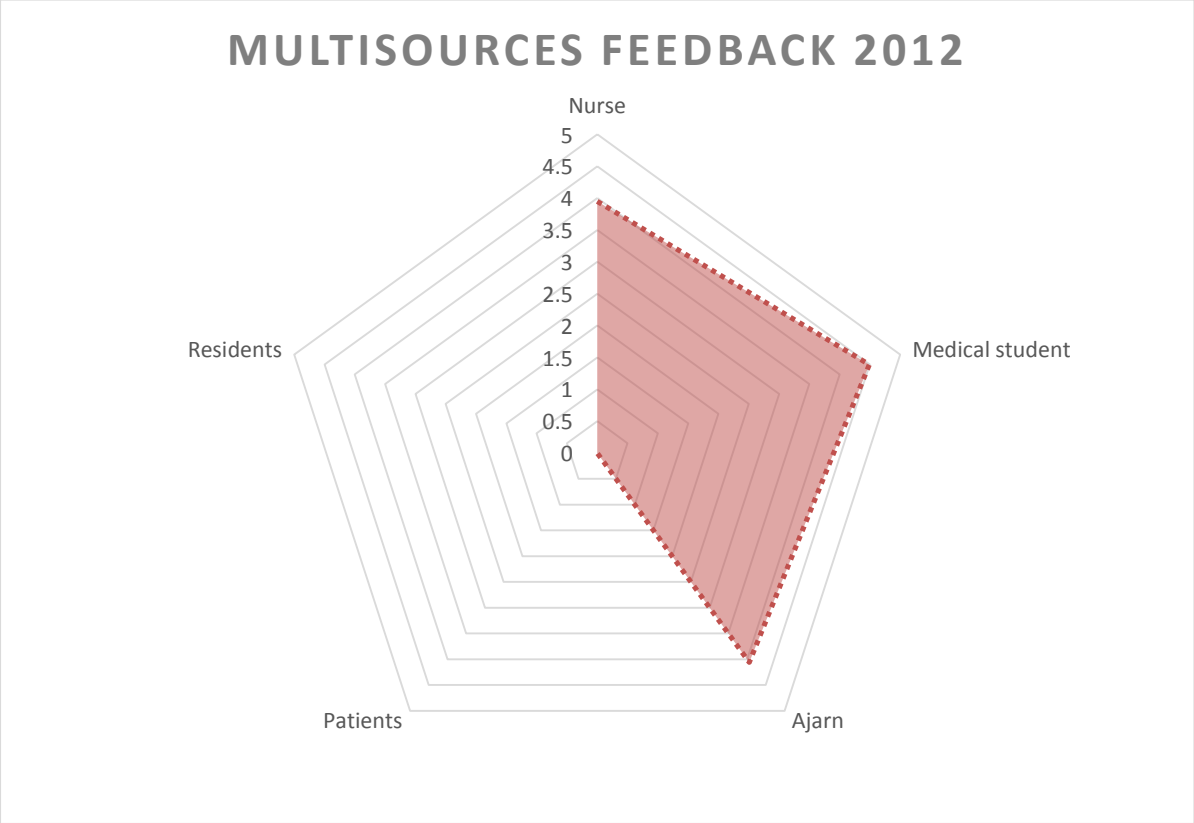
ได้แก่ ความรับผิดชอบ ความเสียสละ ความตรงต่อเวลา ความซื่อสัตย์ การปฏิบัติตามระเบียบข้อบังคับ และอัธยาศัย/น้ำใจ/ความเอื้อเฟื้อต่อผู้อื่น



ชั้นปีที่ 1 ปีการศึกษา 2554

Rotation	อาจารย์ (5 คะแนน)	นักศึกษาแพทย์ (5 คะแนน)	พยาบาล (5 คะแนน)	แพทย์ประจำบ้าน	ผู้รับบริการ
พระศรีฯ 9/2			4.00		
LR เช้า		3.58	4.00		
LR พิเศษเช้า			3.90		
นรีเวช 1	3.50	3.40	4.10		
นรีเวช 1 (2)	4.00	3.41	3.92		
พระศรีฯ 10/2			3.92		
พระศรีฯ 9/1+ANC			4.03		
LR ดึก			3.76		
LR พิเศษบ่าย			3.23		
นรีเวช 2	4.20	3.17	5.00		
Onco	3.88	5.00	4.07		
พระศรีฯ 10/3		3.83	2.92		
พระศรีฯ 10/1		4.50	3.84		
คะแนนเฉลี่ย	3.89	3.84	3.89		

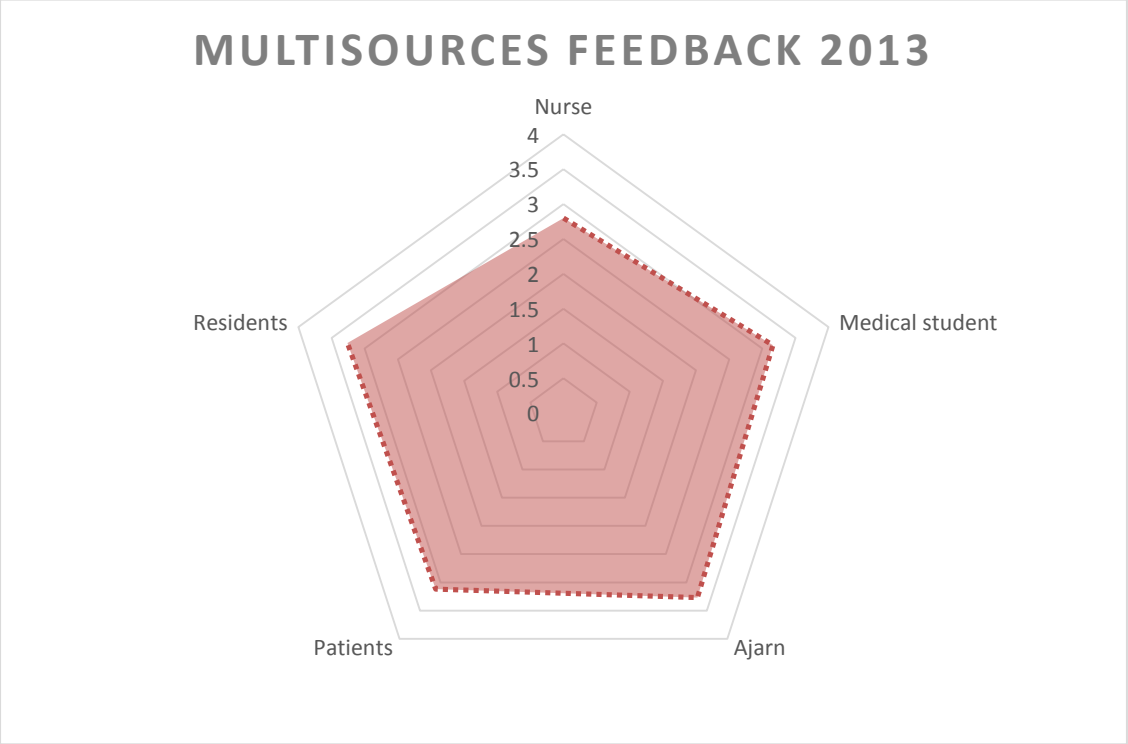
*ยังไม่มีการประเมินจากแพทย์ประจำบ้านและผู้รับบริการในปีการศึกษา 2554



ชั้นปีที่ 2 ปีการศึกษา 2555

Rotation	อาจารย์ (5 คะแนน)	นักศึกษาแพทย์ (5 คะแนน)	พยาบาล (5 คะแนน)	แพทย์ประจำบ้าน	ผู้รับบริการ
นรีเวช 1	3.95	4.67	4.84		
เลิดสิน	4.00				
พระศรีฯ 13/1	4.00		3.69		
LR ดึก			4.07		
Onco	4.05	3.77	3.76		
พระศรีฯ 14/2			4.42		
นรีเวช 2	3.82	4.50	4.23		
พระศรีฯ 13/2	4.35		3.08		
พระศรีฯ 10/1		5.00	3.25		
พระศรีฯ 14/1	4.30		4.00		
LR เช้า		4.58	4.20		
พระศรีฯ 10/3		4.42	4.00		
คะแนนเฉลี่ย	**4.06**	**4.49**	**3.95**		

*ยังไม่มีการประเมินจากแพทย์ประจำบ้านและผู้รับบริการในปีการศึกษา 2555



ชั้นปีที่ 3 ปีการศึกษา 2556

Rotation	อาจารย์ (4 คะแนน)	แพทย์ประจำบ้าน (4 คะแนน)	พยาบาล (4 คะแนน)	นักศึกษาแพทย์ (4 คะแนน)	ผู้รับบริการ (4 คะแนน)
นรีเวช 1	3.30	3.40	2.29	3.12	3.28
STD	3.50		3.00		
พระศรีฯ 10/1		3.30	2.20	2.66	2.64
LR พิเศษ			3.06		
OPD GYN			3.40		3.07
Septic		3.50	3.00	3.40	3.33
วิสัญญี	2.70				
นรีเวช 2	3.40	3.48	3.13	3.44	3.40
Infertile	3.30				
นครปฐม	3.30				
OPD ANC			2.85		3.04
ONCO	3.05	2.80	2.21	3.75	
LR เช้า		3.10	2.95	2.63	
Surgery	3.62				
คะแนนเฉลี่ย	**3.27**	**3.26**	**2.80**	**3.16**	**3.12**

***เริ่มการประเมินจากนักศึกษาแพทย์และผู้รับบริการ ในปีการศึกษา 2556**

AN INTERNATIONAL ASSOCIATION FOR MEDICAL EDUCATION

Portfolios for Assessment and Learning

Jan van Tartwijk
Erik W Driessen

AMEE GUIDE 45
Assessment

AMEE Guides in Medical Education

www.amee.org

Welcome to AMEE Guides Series 2

The AMEE Guides cover important topics in medical and healthcare professions education and provide information, practical advice and support. We hope that they will also stimulate your thinking and reflection on the topic. The Guides have been logically structured for ease of reading and contain useful take-home messages. Text boxes highlight key points and examples in practice. Each page in the guide provides a column for your own personal annotations, stimulated either by the text itself or the quotations. Sources of further information on the topic are provided in the reference list and bibliography.

Guides are divided into series according to subject:

- Teaching and Learning
- Research Methods
- Education Management
- Curriculum Planning
- Assessment

The Guides are designed for use by individual teachers to inform their practice and can be used to support staff development programmes.

'Living Guides'

An important feature of this new Guide series is the concept of supplements, which will provide a continuing source of information on the topic. Published supplements will be available to all who have purchased the Guide.

If you would like to contribute a supplement based on your own experience, please contact the Guides Series Editor, Professor Trevor Gibbs (tjg.gibbs@gmail.com).

Supplements may comprise either a 'Viewpoint', when you communicate your views and comments on the Guide or the topic more generally, or a 'Practical Application', where you report on implementation of some aspect of the subject of the Guide in your own situation. Submissions for consideration for inclusion as a Guide supplement should be maximum 1,000 words.

Other Guides in the new series

A list of topics in this exciting new series is listed on the back inside cover.

Institution/Corresponding address:

Dr Jan van Tartwijk, ICLON – Leiden University Graduate School of Teaching, Leiden University,
PO Box 905, 2300 AX Leiden, The Netherlands

Tel: +31 71 527 3845

Fax: +31 71 527 5342

Email: jtartwijk@iclon.leidenuniv.nl

The authors:

Dr Jan van Tartwijk works at the ICLON – Leiden University Graduate School of Teaching. In his research and teaching he focuses on teacher-student communication processes in the classroom and the use of portfolios in medical education and teacher education.

Dr Erik Driessen works at the Department of Educational Development and Research at Faculty of Medicine of the University of Maastricht. He specializes in assessment and the use of portfolios in medical education.

Both have a long history with working with portfolios. Jan van Tartwijk started experimenting with portfolios in teacher education and faculty development in 1994. In 1999, he joined Erik Driessen and Cees van der Vleuten at Maastricht University, where they implemented portfolios in the undergraduate program of the Faculty of Medicine of the University of Maastricht. Since then, they have published a series of articles and books about using portfolios in higher education and have advised numerous faculties and originations in medical education and elsewhere about the use of portfolio for learning and assessment. Their corporation is not limited to the topic of portfolios; they also work together on research on how to stimulate and assess self-critical thinking and reflection.

Part of this AMEE Guide was first published in Medical Teacher:
Van Tartwijk J & Driessen EW (2009). Portfolios for assessment and learning. AMEE Guide No.45.
Medical Teacher, 31(9): 790-801.

Guide Series Editor: Trevor Gibbs (tjg.gibbs@gmail.com)
Published by: Association for Medical Education in Europe (AMEE), Dundee, UK
Designed by: Lynn Thomson

ISBN: 978-1-903934-57-9

Contents

Abstract

In 1990, Miller wrote that no tools were available for assessment of what a learner does when functioning independently at the clinical workplace (Miller 1990). Since then portfolios have filled this gap and found their way into medical education, not only as tools for assessment of performance in the workplace, but also as tools to stimulate learning from experience.

We give an overview of the content and structure of various types of portfolios, describe the potential of electronic portfolios, present techniques and strategies for using portfolios as tools for stimulating learning and for assessment, and discuss factors that influence the success of the introduction. We conclude that portfolios have a lot of potential but that their introduction also often leads to disappointment, because they require a new perspective on education from mentors and learners and a significant investment of time and energy.

TAKE HOME MESSAGES

- The goals of working with a portfolio need to be clear.
- It is not problematic to use portfolios concurrently to formatively promote learning as well as for summative assessment. Summative assessment is important to ensure that portfolio learning maintains its status alongside other assessed subjects.
- The effectiveness of learning is enhanced when a mentor supports the portfolio process. Mentorship requires a substantial time investment but is crucial for the successful use of portfolios. The effectiveness of assessment can be enhanced by combining the portfolio with an interview.
- Use a flexible learner-centred portfolio format. A rigid structure in which every detail of portfolio content is prescribed will elicit negative reactions from portfolio users.
- Too much structure is a greater risk than too little structure, but learners do need clear directions and guidance to support the development and assessment of broad competencies.
- Working with a portfolio is time consuming both for learners and mentors. This is more of a problem in postgraduate training and continuous medical education than in undergraduate education.

Introduction

Today's doctors find themselves confronted not only with patients who are increasingly knowledgeable and assertive, but also with pressure to apply new findings and evidence in day-to-day practice, and with the necessity to collaborate with other health professionals in ever larger teams and communities. To deal with these complexities, doctors need generic competencies to enhance effective communication, organization, teamwork and professionalism. These generic competencies are sometimes labelled as doctors' "soft skills" in contrast to "hard clinical skills". In recent years, learning, teaching and assessment of these generic competencies has gained unexpected urgency among politicians and the general public. Headlines decrying incidents involving dysfunctional doctors and hospital departments with dramatic impact on morbidity and mortality figures catapulted generic competencies to the forefront of attention as indispensable qualities for doctors. As a result, professional associations and governments began to voice increasingly urgent demands to include these generic competencies in education and assessment (General Medical Council, 2000). At the same time, consistent with the general trend towards outcome-based education, the focus in medical education shifted from the educational process itself towards the competencies of doctors at the end of training and at important junctures during the training process (Norcini et al., 2008). The competencies described by professional organizations such as the Royal College of Physicians and Surgeons of Canada (1996) became the framework for assessment and, as a consequence, for the content and organization of programmes for medical education in many countries.

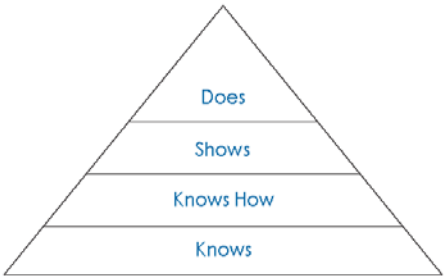
...doctors need generic competencies to enhance effective communication, organization, teamwork and professionalism.

However, stimulating the development of competencies (Box 1) and the assessment of its result is complicated. Already in 1990, Miller described the challenges involved in assessing clinical competence. He presented a framework for clinical assessment, shaped like a pyramid (Figure 1), whose layers from bottom to top represent increasingly complex levels of mastery, with the lower levels providing the foundation for the higher levels (Miller, 1990).

BOX 1
Competence

The concept of competence is much used and much debated (Stoof et al., 2002; Dreyfus, 2004). Here, we define it as an integrated body of knowledge, skills, and (professional) attitudes enabling proficient performance in certain real life settings, i.e. the "Does" level in Miller's framework.

FIGURE 1
Framework for clinical assessment: Miller's Pyramid (cf. Miller, 1990)



The bottom level is concerned with *knowledge*. This is the knowledge relating to the skills that learners must master for their future professional practice. This knowledge is best assessed by written tests. The next level represents application of the knowledge from level 1. Learners should know *how* to apply their knowledge when performing skills. For instance, at this level, learners are expected to know how to diagnose a patient and which aspects of a patient's presentation to attend to. The *knows how* level can also be assessed by written tests. One level up, at level 3, the issue of interest is that learners demonstrate their ability to use their knowledge to *take appropriate action in a simulated environment*. This level combines knowledge and action (cognition and behaviour). Not only should learners know how to diagnose a patient, they should also be able to actually perform the appropriate actions, for example a physical examination in a simulated patient (*shows how*). The top of the pyramid is concerned with *independent performance within the complex environment of day-to-day practice*. This requires integration of knowledge, skills, attitudes, and personal characteristics. Performance at the top of the pyramid is manifested when learners are working independently in professional practice. Typically, adequate performance at this level requires integrated performance of different roles; not only the role of medical expert but also that of counsellor, participant in the doctor- patient relationship, a leadership role in relation to nursing staff, etc. Good performance at the Does level (of Miller's Pyramid) implies competence.

Good performance at the Does level (of Miller's Pyramid) implies competence.

In 1990, Miller observed that there were no instruments to evaluate performance consistent with the top of the pyramid (Miller, 1990). At the same time, scholars in the field of teacher education and teacher assessment were struggling with the same problem (Bird, 1990). Here too, the key challenge was how to assess performance in real life settings. Shulman (1998) describes the Teacher Assessment Project that was set up with the purpose of exploring and developing new approaches to the evaluation of teaching in primary and secondary education. He recounts that it was considered undesirable to assess teacher competence solely on the basis of ratings in assessment centres, because experiments showed that the information provided by assessment centres alone was not enough to identify competent and excellent teachers. Information about whether teachers succeeded in making the most of their pupils' learning opportunities *within* their own complex working environment was needed as well. It was also

recognised that there can be striking variations among teaching settings. For instance, it makes quite a difference whether one teaches at an urban school in a deprived area with its myriad of social problems or at a high school in a middle class suburban environment. As part of efforts to achieve fair judgement of teacher performance in a broad array of settings and situations, the *portfolio* concept was borrowed from the arts and architecture (Box 2).

BOX 2
Portfolio

Portfolios that are used in education contain evidence of how learners fulfil tasks and their competence is progressing. They may be digital or paper based and content may be prescribed or left to the learners' discretion. Despite variations in content and format, portfolios basically report on work done, feedback received, progress made, and plans for improving competence (Driessen et al., 2007b).

Since portfolios were introduced in medical education in the early 1990s (Royal College of General Practitioners, 1993), their use as an instrument for both assessment and encouraging professional growth has increased enormously (Snadden et al., 1999; Friedman Ben David et al., 2001). However, the evidence to date suggests that the introduction of portfolios for these purposes has met with mixed success (Driessen et al., 2007b; Tochel, et al., 2009, Buckley et al., 2009). Although potentially powerful instruments in education, the use of portfolios has proved to be vulnerable.

The aim of this AMEE Guide is to help medical teachers and educators to make full use of the possibilities that portfolios offer and prevent difficulties occurring. Based on an analysis of what portfolios help achieve, it is our purpose to provide practical clues about the design, implementation and use of portfolios in medical education.

Firstly, we will describe how portfolio content and structure relate to the various goals that they are designed to achieve. Next, we will focus on the use of portfolios as instruments that can encourage professional growth by stimulating learning from experience and subsequently, we will elaborate on the use of portfolios as instruments for assessment. Each of these goals requires specific content and organization of portfolios. Finally, we will focus on the factors that are important for the successful introduction of portfolios in (medical) education.

Portfolio goals, content, and organization

Portfolios as a multipurpose instrument

- **Portfolios for assessment**: When portfolios were originally introduced in education as instruments for authentic assessment, they closely resembled the portfolios of architects and artists that Lyons (1998) describes as a portable case for keeping, usually without folding, loose sheets of papers, drawings or photographs. Building on the principle of triangulation (Denzin, 1978; Denzin & Lincoln, 2000) all kinds of evidence can be brought

together in those portfolios that, in combination, give the possibility to draw valid conclusions about competence (Box 3).

BOX 3
Combining evidence to improve the quality of conclusions

In the literature, combining data from various sources with the aim to improve the quality of conclusions is often referred to as triangulation. The aim of triangulation is to avoid biases and problems, such as those related to the reliability and trustworthiness of data that are derived from one single source.

Procedures for multisource feedback or 360-degree feedback use a similar strategy by stimulating learners to gather feedback from different sources. Lockyer & Clyman (2008) describe a procedure involving a questionnaire survey among medical colleagues, nurses, and patients and their families to collect data about learners' specific competencies. The same questionnaire is completed by the learners themselves. By aggregating these data, reliability is improved.

However, in one of the first explorations of portfolios for teacher assessment, Bird (1990) wrote that the portfolio procedures for assessment might easily degenerate into exercises in amassing paper. He suggested that the evidence in a portfolio should be organised according to the competencies that the person compiling the portfolio wants to show. Both for the learner compiling the portfolio and for an assessor this would be helpful. Instructions starting with "Show how you..." might clarify for portfolio owners that they are asked to provide specific evidence about their performance. A portfolio organised by tasks or competencies might be helpful for assessors, because it indicates what the material in the portfolio is supposed to show. Based on initial experiments with portfolios, Collins (1991) suggested that captions should be attached to the evidence in the portfolio:

> *One essential component of the portfolio was the document caption. The caption is a little sheet attached to each document stating what the document is (...) and why it is valuable evidence. (...) Captions proved to be essential to the portfolio development process. Documents without captions were meaningless to the raters.* (p. 153)

...portfolio procedures for assessment might easily degenerate into exercises in amassing paper.

- **Portfolios for learning**: Soon after the introduction of portfolios in medical education, Snadden & Thomas introduced the term "portfolio learning" (Snadden & Thomas, 1998b):

  > *Portfolio learning is a method of encouraging adult and reflective learning for professionals. Derived from the graphic arts it is based on developing a collection of evidence that learning has taken place* (p. 192)

  They emphasise the importance the importance of supervision and critical reflection for portfolio learning:

  > *The system works well when it operates through the interaction of a learner and mentor using the material as a catalyst to guide further learning. It is essential that the portfolio does not become a mere collection of events seen or experienced, but contains critical reflections on these and the learning that has been made from them* (p.192).

Portfolio learning is a method of encouraging adult and reflective learning for professionals. Derived from the graphic arts it is based on developing a collection of evidence that learning has taken place.

A portfolio can also stimulate reflection...

A portfolio can also stimulate reflection, because collecting and selecting work samples, evaluations and other types of materials that are illustrative of the work done, compels learners to look back on what they have done and analyse what they have and have not yet accomplished.

In many cases, portfolios are assembled over a longer period of time. That is why they can also be used to support planning and monitoring in professional development. One way to do so is to include learning objectives in the portfolio as well as a document trail of related learning activities and accomplishments (Mathers et al. 1999; Oermann, 2002).

As a consequence, reflections and overviews of personal development have secured a prominent place in many portfolios. Portfolios that are primarily geared to assessment will remain organised around all kinds of materials that provide 'evidence' of competencies. In portfolios that are primarily used to monitor and plan learners' development, overviews will take centre stage. Portfolios whose primary objective is to foster learning by stimulating learners to reflect on and discuss their development will be organised around learners' reflections.

- **A multipurpose instrument**[1]: Inevitably, these developments have widened the applicability of the label *portfolio* to a broad range of instruments. Some portfolios might equally and aptly be labelled *Personal Development Plan or Reflective Essay*. Because of the tremendous variety in portfolios, careful and critical appraisal of the strengths and weaknesses of different portfolios is advisable before deciding which one to implement in a particular setting.

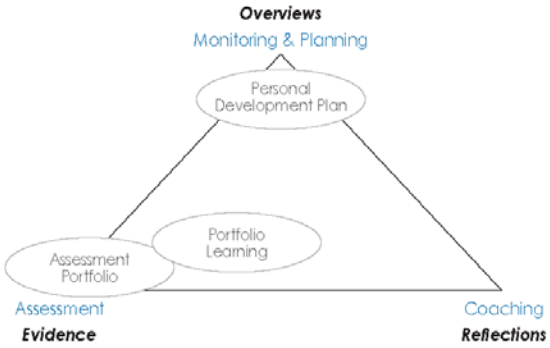
  The question to be answered is whether a certain portfolio is fit for its intended purpose. And just as someone else's shoes are unlikely to fit comfortably, portfolios tailored to one particular educational setting may not fit into the educational configuration(s) of other settings (Spandel, 1997). An ill-fitting portfolio will inevitably be discarded sooner or later. To assist in determining whether a portfolio is appropriate for its intended purpose the triangle in Figure 2 helps to define the nature of a portfolio. It does so by inviting positioning of a portfolio in the area of the triangle where it is most likely to achieve its intended principal objectives.

  Obviously, a portfolio can be used to achieve more than one goal. When a portfolio is to serve a combination of goals, its position in the triangle will shift towards the centre because its strengths have to be distributed more evenly over evidence, overviews and reflections. In practice, the majority of portfolios are not situated in one of the corners of the triangle (Buckley et al., in press). A controversial issue in the literature on educational portfolios is whether it is acceptable to have one portfolio for both assessment and reflection (Snyder et al. 1998). An argument against this dual function is that assessment may jeopardise the quality of reflection thereby detracting from the portfolio's effectiveness for mentoring purposes. Learners may be reluctant to expose their less successful efforts at specific tasks and to reflect on strategies for addressing weaknesses if

1 Parts of this section were published in the journal Quality in Higher Education (van Tartwijk, et al., 2007)

they believe they are at risk of having 'failures' turned against them in an assessment situation. Portfolios that are not assessed, on the other hand, do not "reward" learners for the time and energy they invest in them. As a result, learners are likely to take the portfolio and any associated learning activities less seriously. A recent BEME review showed that most portfolios were also assessed for summative purposes (Buckley et al., 2009).

FIGURE 2
Purposes and content of portfolios (van Tartwijk, et al., 2007)



An effective portfolio has a clear but flexible structure, giving individual learners opportunities to describe their own unique development (Pearson & Heywood, 2004; Driessen et al. 2005b; Grant et al. 2007). Clear instructions are important, but when the content of a portfolio is prescribed in detail, portfolios are often experienced as highly bureaucratic instruments (Davis et al., 2001; O'Sullivan et al. 2004; Pearson & Heywood, 2004; Kjaer et al. 2006). Portfolios meet with stronger appreciation when learners have a certain amount of freedom to determine the content of their own portfolios (Snadden & Thomas, 1998a; Driessen et al., 2005b).

An effective portfolio has a clear but flexible structure, giving individual learners opportunities to describe their own unique development.

Electronic portfolios

A growing number of medical schools use electronic portfolios (e-portfolios) instead of paper-based portfolios (Fung Kee Fung et al., 2000; Lawson et al., 2004; Woodward & Nanlohy, 2004; van Tartwijk et al., 2007; Driessen et al. 2007a). This preference is based on a number of considerations:

- In e-portfolios, hyperlinks can be inserted to make connections between evidence, overviews, and reflections. This can be useful, for instance, when learners want to illustrate reflections with evidence that is stored somewhere else in the portfolio, or want to illustrate a schematic overview of their development by making hyperlinks to materials and reflections. Hyperlinks can also be useful to make a table of contents of the portfolio. For instance by including a list of captions in the portfolio and making hyperlinks to related materials. Mentors or assessors can browse through this list of captions, obtain a quick overview of all the evidence in the portfolio, and just click on the evidence that is relevant to their specific purpose.

- A paper-based portfolio can be cumbersome because of its bulk. Imagine an assessor who needs to take 15 paper portfolios home! Furthermore, there is generally only one copy of a paper portfolio. Whenever learners hand their paper portfolios to their mentor or assessor, the portfolio is literally out of their hands. Not only do they run the risk of the portfolio getting lost, it is also more difficult for them to prepare to discuss the portfolio with their mentor or assessor. Another advantage of e-portfolios is that they are easier to keep up to date.

Of course there are disadvantages as well:

- Mentors who do not like to read a portfolio on screen will still have to print it. In most systems it is not possible to make notes on the portfolio itself (although making notes on the learner's paper portfolio might not be desirable as well).
- E-portfolios can only be used by learners and teachers who are sufficiently skilled in using the relevant software and hardware.
- An e-portfolio requires a stable and high quality information technology infrastructure that is not always available.

Nowadays, many dedicated portfolio systems are available, which are usually user-friendly (Dornan et al., 2002; www.eportfolioservice.nl). These systems can provide specific functionalities for specific portfolio goals: options to include work-based assessment instruments, such as multisource feedback or mini clinical evaluation exercises (mini-CEX) in portfolios for clinical training: to invite specific individuals to inspect the portfolio, either wholly or in part, while denying access to everyone else.

Apart from dedicated systems, learners can produce an e-portfolio using standard word-processors or HTML editors, preferably ones that they and their teachers are familiar with (Gibson & Barrett, 2003). The cost of dedicated portfolio software is not the only reason to support this choice: for many purposes the hyperlink functionality of generic software is all that learners need. Furthermore, generic software allows a learner to impart his or her own flavour to the portfolio. This can enhance the learners' motivation to work with the instrument. Another reason is that many portfolio systems are limited because they are built to accommodate no more than one or two portfolio types. Finally, portfolios built with dedicated software need to be accessible with generic software for later maintenance and presentation. This may well be the case after a learner has left the setting in which the portfolio was produced, or in the event that the vendor in question ceases to do business. In summary, standard software tools have disadvantages from the perspective of managing access to the portfolio using the internet or to include work-based assessment instruments, but they usually provide all the options learners need to produce a portfolio that works well and looks great.

...standard software tools have disadvantages from the perspective of managing access to the portfolio using the internet or to include work-based assessment instruments, but they usually provide all the options learners need to produce a portfolio that works well and looks great.

In a study comparing web-based and paper-based portfolios (Driessen et al., 2007a), not only did learners added more personal touches to content and form and invested more time in their portfolios, but mentors were unanimous in their appreciation of the greater ease of use of web-based portfolios compared to the more familiar paper-based ones. Information was

easy to locate without having to turn pages to find certain content and the portfolios could be accessed from different locations were two reasons cited for preferring web-based portfolios. Other authors have also reported on the user friendliness of electronic portfolios (Fung Kee Fung et al., 2000; Lawson et al., 2004). In these studies, tutors appreciated the easy electronic access and reduction in the amount of paper used. However, the same authors also reported certain situations that make web-based portfolios less user-friendly than paper-based portfolios. For instance, limited computer access in the clinical workplace cancels out the advantages of user-friendliness and may even have an opposite effect.

Portfolios and learning from experience

Research shows that the role of the mentor is crucial to the successful use of portfolios aimed at learning from experience (Finlay et al. 1998; Snadden & Thomas, 1998a Mathers et al., 1999; Pearson & Heywood, 2004; Driessen et al., 2005b; Grant et al., 2007). In this section, we focus on the strategies mentors can use to promote learning from experience with a portfolio.

Research shows that the role of the mentor is crucial to the successful use of portfolios aimed at learning from experience.

Theoretical background

The contemporary view of learning, based on constructivism, is that people "construct" new knowledge and understanding based on what they already know and believe (Bransford et al. 2000). What people know and believe can be represented as cognitive structures that guide their perception of reality. Evidently, a perception of reality based on individual cognitive structures does not afford an objective view of reality, but, by definition, an individual, idiosyncratic view. It is this personal perception of reality that guides a person's actions.

Reflection is an important concept in this framework, which relates to changing cognitive structures. Research has shown that meta-cognitive skills, such as reflection, increase the degree to which learners transfer what they have learned to new settings and events (Bransford et al., 2000). Despite considerable confusion about the precise definition of the term reflection (Hatton & Smith, 1995; Mann et al. 2007) all authors writing about reflection share the constructivist view that human behaviour is guided by mental structures that are not static but flexible, evolving, and changing in response to experiences. Based on this consensus view, reflection can be defined as the mental process of organising or reorganising cognitive structures that represent existing knowledge and beliefs and guide perceptions of experiences, situations, and problems (Korthagen et al. 2001). To put it in simpler terms: reflection means exploring and elaborating one's *understanding* of an experience (Eva & Regehr, 2008). Building on Van Manen's work (1977), Hatton & Smith (1995) distinguish three types or levels of reflection. The first type is concerned with the *means* to achieve certain ends. The second type is not only about means, but also about *goals*, the *assumptions* upon which they are based, and the actual *outcomes*. The third type of refection is referred to as *critical reflection*. Here, moral and ethical criteria are also taken into consideration. Judgements are made about whether professional activity is equitable, just, and respectful to persons or not. Hatton and Smith emphasise that these three types of reflection should

...meta-cognitive skills, such as reflection, increase the degree to which learners transfer what they have learned to new settings and events.

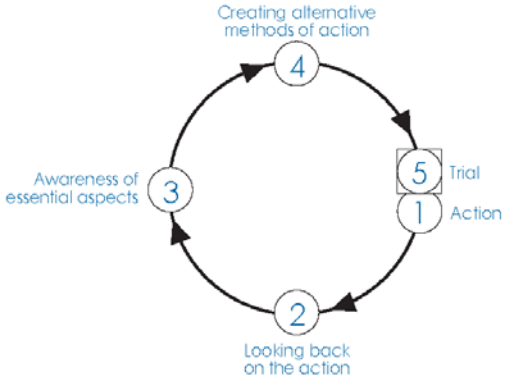
not be viewed as hierarchical. Different (educational) contexts and situations may lend themselves more to one kind of reflection than to another.

Reflection and professional development

For medical teachers who want to help learners learn from practice, the key question to answer is: "*How can I stimulate my learners to reflect on their experiences and learn from them?*" For this AMEE Guide the additional question is: "*... and how can a portfolio help to improve the quality of reflection?*".

Korthagen designed the **ALACT** model (Action, Looking back, Awareness, Creating alternative methods, Trial) (Figure 3) to describe the spiralling process that effective learners go through when faced with a situation for which no routine solution is available (Korthagen et al., 2001). This model resembles the three step model described by Snadden & Thomas (1998b) which focused on evaluation, reflection, and formulating a learning plan. We will describe the ALACT model, explain the potential contribution of working with a portfolio in each of the stages, and give suggestions for coaching strategies (Driessen et al., 2008).

FIGURE 3
ALACT model showing the phases of spiral professional development (Korthagen et al., 2001)



ALACT

A **Action:** The cycle starts with action undertaken for a specific purpose (e.g. for developing a specific competence). Learners can be helped to improve their existing routines and concurrently acquire new ones by pre-selecting experiences from which they can learn, for example a mixture of patients who are more or less easy to diagnose. Ericsson's research predicts that expertise will grow not just from the weight of experience but also from engaging in activities specifically designed or selected to improve performance (Ericsson, 2006).

Learners can be helped to improve their existing routines and concurrently acquire new ones by pre-selecting experiences from which they can learn.

L **Looking back on action: self-directed assessment seeking**: The ALACT cycle then moves to the stage where learners look back on a previous action, usually when that action was not successful or something unexpected happened. This looking back on action is assumed to be accompanied by an evaluation of whether the goals were realised and the learner's part in this. In many cases this can be regarded as a form of *self assessment*. Eva & Regehr (2008) write that most of the time self-assessment is conceptualised according to a "guess your grade" model of which the quality is generally poor (Davis et al., 2006). As an alternative they propose *self-directed assessment seeking*, which they describe as a process by which a learner takes personal responsibility for looking outward, explicitly seeking feedback and information from external sources of assessment data, to direct performance improvements that can help them to validate their self-assessment.

The role of the portfolio: Seeking and selecting evidence (documents, feedback, work-based assessments, etc.) for inclusion in a portfolio can be regarded as self-directed assessment seeking. To improve the quality of this process, it is important to use a variety of evidence from various sources. The validity of the results of self-directed assessment seeking will be maximised if the learner's self-reflections are consistent with all the information that is brought together in a portfolio.

Teaching strategies: Research has shown that a mentor can play a decisive role in determining whether the use of portfolios in education is successful or not (e.g. Driessen et al., 2007b). At the very least, learners may expect their mentors to pay serious attention to their portfolios, for after all they did spend a lot of time and energy to put their portfolio together. But even more importantly, careful scrutiny of their own performance may be confronting for learners. Effective mentors have an important role in this respect. In Box 4, we give suggestions for a number of strategies to be used by medical teachers in this phase, derived from the work by Korthagen and colleagues (Korthagen et al. 2002).

Seeking and selecting evidence (documents, feedback, work-based assessments, etc.) for inclusion in a portfolio can be regarded as self-directed assessment seeking.

A **Awareness of essential aspects: reflection**: After conclusions have been drawn about the quality of performance and the characteristics of the situation, the next step in the ALACT model is to foster awareness of essential aspects. In this phase, learners try to develop a new and better understanding of what has happened, i.e. they reflect on their performance.

They can focus on the *means* they used to achieve a goal and try to understand why their strategy was successful or not. They can also consider whether they had selected a suitable *goal* for this particular situation. And finally they may consider what they want to achieve from a *moral* or *ethical* perspective.

BOX 4
Strategies to stimulate self-directed assessment seeking

- Provide a safe environment by distinguishing between learners as individuals and their performance.
- Focus on description.
- Stimulate learners to be concrete in their reports. When learners give general evaluations about a situation and their performance, ask questions:
  - What went well?
  - What went wrong?
  - How did you solve this?
  - What effect did this have?
- Stimulate learners to carefully scrutinise all the information in their portfolio. Learners could be asked to go through all the available evidence and answer questions:
  - Which information in your portfolio supports your answers/evaluation?
  - Which information in your portfolio contradicts your answers/evaluation?
- Stimulate learners to take the perspective of other stakeholders. Ask questions:
  - What did you want? What do you think the patient/your colleague/the nurse wanted?
  - What did you think? What did the others think?
  - What did you do? What did the others do?
  - What emotions did you experience? What emotions did the other people involved experience?

The role of the portfolio: Language is important in supporting thinking. Writing things down can help to stimulate reflection (Korthagen et al., 2001). Written reflections were not a part of the original portfolios, like the ones in which artists presented a selection from their work, but almost immediately after the introduction of portfolios in education, written reflections became a fixture of portfolios (Paulson et al. 1991). Embedding a written reflection in a portfolio has the advantage that it can be built on the self-assessment that was validated by the evidence in the portfolio. This is a form of facilitated reflection (Conlon, 2003). The learner can also use the evidence to illustrate a reflection with a concrete example.

Language is important in supporting thinking. Writing things down can help to stimulate reflection.

Teaching strategies: To stimulate learners to reflect and learn from their experiences, mentors do not need to have all the right answers. The most important thing for them is to ask the right questions. In Box 5 we give a number of examples of questions that mentors can ask.

To stimulate learners to reflect and learn from their experiences, mentors do not need to have all the right answers. The most important thing for them is to ask the right questions.

BOX 5
Questions to stimulate reflection

Means
- Which strategies did you consider? Why did you select this strategy? Which are the advantages and disadvantages of the strategy you used?
- Which part of your strategy was effective and which part was not effective? Why was it effective / not effective?
- Would this strategy have been more /less effective in a different situation?

Goals, assumptions, outcomes
- What did you want to achieve? Were you successful? What do you consider successful?
- Why is this particular goal important?/Why did you pursue this goal?

Critical reflection
- Do you think patients / patients' families / medical colleagues / nurses / administrators are satisfied with these outcomes? What are their primary interests?

Confront with discrepancies
- I read in your portfolio that you are happy with the result, but when we talk about it, your face tells a different story.
- You write here that this is what you want to achieve, but you are pleased with your results even though they do not match your goals.
- You do not actually do what you say you want to do.

Generalize across experiences
- Which differences and similarities do you recognise between what is happening now and what happened in situations that you described in your portfolio?
- When do these things happen?
- Do you recognise a pattern?

C **Creating or identifying alternative methods of action:** change: Analysing previous actions may trigger a search for alternative strategies, or abandonment of original goals. It is important to explicate (new) goals and alternative strategies. A recent review showed that goal setting stimulates learning and that a mentor has an important role to play in this respect (Shute, 2008). Learners who work with a mentor set more specific goals and improve more than those who do not work with a mentor (Smither et al. 2003). Very often, agreement about what should be done differently and which goals should be achieved are written down in a document that is referred to as a Personal Development Plan (PDP).

Learners who work with a mentor set more specific goals and improve more than those who do not work with a mentor.

The role of the portfolio: In many portfolios, the central goal is to keep track of the learner's development. In these portfolios, PDPs can have an important place. Snadden & Thomas for instance, (Snadden & Thomas 1998b) propose that when a portfolio is used for professional development and to track progress, it is important to attach to the portfolio some kind of learning plan.

Teaching Strategies: Both mentors and learners should commit to the agreements in the PDP and it should be on the agenda of their next progress meeting. The plans in the PDP are often too vague. It is important that mentors stimulate learners to be very concrete. It can be helpful to keep in mind that the learning goals in the plan should be formulated in a SMART way (Box 6).

BOX 6
SMART

Specific	(Straightforward, not ambiguous)
Measurable	(It is clear under which conditions the goals are achieved)
Acceptable	(The goals should be acceptable to all stakeholders)
Realistic	(The learner should be able to achieve the goals)
Time-bound	(It should be clear when the goal is to be achieved)

T **Trial:** The last step in the ALACT cycle is trial. This is also the start of a new cycle in the spiral of professional development in this model.

Using portfolios as tools for assessment

In the introduction, we quoted Shulman (1998), who wrote that the reason for introducing portfolios in education as tools for assessment is that in a portfolio information can be brought together about how a person performs and how his or her competencies develop in his or her own complex working environment. From the perspective of assessment, the strength of the portfolio is also its weakness. The evidence held by a portfolio is often not standardised and its meaning often depends on the context from which it originates.

The evidence held by a portfolio is often not standardised and its meaning often depends on the context from which it originates.

Assessing non-standardised portfolios requires a different perspective on assessment than the traditional quantitative perspective that is best suited for analysing quantitative test scores or results from standardised observations. Authors like Snadden (1999) and Webb (2003) all come to the conclusion that we should not try to fit non-standardised portfolios to standardised psychometric assessment criteria. They point out that portfolio assessment is primarily concerned with interpreting various forms of qualitative information and suggest that assessment procedures should be developed that are based on methods used in qualitative research.

In the next section, we will translate the insights of this literature into recommendations for portfolio assessment. We will structure this section according to five questions that, according to Harden (1979), should always be asked and answered by medical teachers in relation to assessment:

- What is assessed?
- Why is this assessed?
- How is this assessed?
- Who assesses?
- When is this assessed?

What? Although portfolios are also used in undergraduate medical education to assess reflective ability or communication skills (Driessen et al. 2003), portfolios are particularly suited to work-based assessment. In other words, they have added value at the does level of Miller's pyramid (Miller 1990).

Many medical curricula are based on competency criteria developed by organisations such as the General Medical Council (GMC), the American Council of Graduate Medical Education (ACGME), and the Royal College of Physicians and Surgeons of Canada (RCPSC). More often than not, additional detail is required to fit the competency criteria to assessment procedures. In aligning competency descriptions with assessment procedures it is of the essence to strike the right balance between very concrete but also very detailed and long lists of "is able to" statements, on the one hand, and very global descriptions providing an overview but too little to support assessment, on the other hand. The extremes of this continuum have been referred to as an analytical versus a global approach. Both approaches have their pros and cons (Box 7).

BOX 7
Analytical versus global assessment

In an analytical assessment, various aspects of a competency are assessed separately. A formula is used to combine the partial assessments into one final score.

Because the criteria are explicitly defined and each partial competence is explicitly assessed, the result is very transparent and usually more reliable and more informative for the learner. Criteria are usually defined in terms of: "The candidate is able to...".

Problems that may occur are:

- Learners may adapt their learning activities to 'ticking' specified criteria. This may result in unnatural activities in the workplace where competencies are acquired.
- Analytical assessment is very labour intensive. It may be experienced as bureaucratic.
- It can be difficult for assessors to give a truly distinct assessment of each individual criterion ('halo effect').
- Assessors have limited freedom to take account of specific competencies or extremely good (or poor) performance: if it is not in the criteria, it is not assessed. The assessor may feel curtailed in his/her freedom by the criteria.

In a global assessment, the assessors study the entire portfolio and give an assessment based on their overall impression. A global assessment is far less labour intensive than an analytical assessment. It also enables assessors to take account of learners' special qualities.

Disadvantages are:

- It is less clear to learners on which criteria the assessment is based. The assessment may also be less reliable. As a result the assessment will be less acceptable to learners.
- Some assessors will feel less certain about their judgement. As a result they will study the material over and over again, which will take even more time than an analytical assessment.
- This type of assessment is relatively vulnerable to assessor preferences and sequence effects (the contrast with the previous candidate may influence the assessment).

A way to combine the best of both approaches is to use scoring rubrics. A *scoring rubric* is a global performance descriptor that lists the criteria for a competency and articulates a limited number of gradations of quality for each criterion. Gradations can be unsatisfactory, sufficient, good, and excellent. Scoring rubrics can be presented as tables, with the criteria in the rows and the grades in the columns. In each cell of this table, performance at that particular level of competence is described. Box 8 provides an example.

BOX 8
Rubrics used for the assessment for final year medical students (source Maastricht University)

	BELOW EXPECTATION	AS EXPECTED	ABOVE EXPECTATION
Clinical performance	Slow in taking a history and performing a physical examination. Considers irrelevant aspects. Slow in making a diagnosis. Misses important conclusions. Frequently unable to formulate management plan and needs considerable guidance.	Adequate speed in taking a history and performing a physical examination. Relevant aspects are considered. Adequate speed in making a diagnosis. Diagnosis contains important conclusions. Formulates an adequate management plan for simple clinical presentations. Needs some guidance. Achieves these goals in the second half of the internship.	Conducts an adequate and efficient history and physical examination. Arrives at an accurate diagnosis within adequate time. Formulates an adequate management plan for simple clinical presentations. Needs little guidance. Has achieved these goals at the start of the internship.
Professionalism (for instance as judged by 360 degree feedback)	Does not keep commitments. Occasionally fails to ask for supervision when this is necessary. Reacts defensively to feedback. Is unable to cope with stress Does not pay attention to his/her personal appearance. Frequently shows awkward behaviour or behaves disrespectfully.	Keeps commitments. Asks for supervision when this is necessary. Needs help in reflecting and considering alternatives and responds adequately to feedback. Occasionally needs help in coping with stress. Appropriate personal appearance; behaves respectfully.	Keeps commitments. Asks for supervision when this is necessary. Is able to reflect critically; responds adequately to feedback and is prepared to acknowledge errors. Is able to cope with stress adequately. Looks well cared for and behaves respectfully.
Has critically assessed his/her performance and formulated appropriate learning goals. This is evidenced by an adequate analysis of strengths and weaknesses and the development plan.	Incomplete, limited or one-sided description of strengths and weaknesses in performance (e.g. only strengths or only weaknesses, limited to one competency). No explanations only lists of facts or situations. No learning goals, learning goals do not match the analysis or are not specific.	A fair number of strengths and weaknesses are not explained or explanations are limited to external attributions (for instance mini-CEX at the wrong moment) Some of the learning goals are not specified.	Above expectation (authentic, recognizable, and well explained). A good analysis of strengths and weaknesses. Also internal attributions and references to evidence in the portfolio. Logical, detailed (based on the analysis) and attainable learning goals.

For learners and their mentors, scoring rubrics can be a roadmap for competence development. It can help them diagnose a learner's current level of competence and point the way to further development. Assessors should not use scoring rubrics as a checklist,

but as a list of arguments to underpin their assessment when they explain it to learners. Learners can also use scoring rubrics to organise their portfolio. They can organise the evidence in their portfolio in chapters corresponding to the different competencies to be assessed and use captions to explain what the evidence shows about a specific competency.

Why? Assessing competencies can be done for three reasons: selection, diagnosis, and certification.

Selection: Determining whether a person is suitable for a certain position. Assessments for selection purposes can take place before entering an educational programme, but also, for instance, before starting a new job.

Diagnosis: In the course of an education programme, the development of learners' competencies is assessed. The purpose of this type of assessment is to give feedback to learners and help them identify new learning goals. Sometimes, this assessment is also used to determine whether or not a learner is allowed to continue with a programme.

Certification: The goal of assessment at the end of an educational or training programme is to establish whether learners have attained the competencies required for graduation or certification. Obviously, the quality of any assessment is important. Poor quality of assessment for selection purposes, for instance, can harm the interests of prospective learners and waste talent. Similarly, poor quality of diagnostic assessment can cause frustration and delay in learners' development. Nevertheless, with graduation and certification decisions the quality of assessment is crucial. Learners who pass but should have failed will become (or continue to be) certified doctors and may become a risk to the community!

How? The quality of the assessment of competencies is crucially determined by the procedure that is used. In the introduction to this section about portfolio assessment, we wrote that the standard psychometric procedures that are used to determine the quality of tests and standardised observations are not very well suited to portfolios with their non-standardised content. In medical education, Webb and colleagues (2003) pointed out that portfolio assessment is primarily concerned with qualitative information and they introduced the idea to use routines developed for qualitative research. Guba & Lincolns' (1989) strategies to achieve *credibility* and *dependability* of assessment can be translated to portfolio assessment (Webb et al., 2003; Tigelaar et al. 2005). In Box 9, we discuss how these strategies can be used.

...the standard psychometric procedures that are used to determine the quality of tests and standardised observations are not very well suited to portfolios with their non-standardised content.

BOX 9

Strategies for portfolio assessment derived from the methodology of qualitative research

- Incorporate feedback cycles into the mentoring process that accompanies the portfolio to ensure that the mentor's final recommendation does not come as a(n) - unpleasant - surprise to the learner; this approach relates to the credibility strategies of prolonged engagement and member checking.
- Maintain a careful balance between the roles of the mentor as coach and assessor. The aim is to ensure that the person who knows the learner best provides the most relevant information while minimizing any damaging effect on the mentor-learner relationship by using an assessment committee to assess the portfolio; this approach relates to the credibility strategy of prolonged engagement.
- Involve the learner in the decision process to ensure commitment on the part of the learner and allow the learner to communicate a different point of view to that of the mentor; this approach relates to the credibility strategy of member checking.
- Use a sequential judgement procedure in which conflicting information necessitates more information gathering. This ensures the efficient use of resources by limiting the use of additional resources to cases where this is necessary to achieve reliable judgement. This approach relates to the credibility strategy of triangulation.
- Document the different steps of the assessment process. For example a formal assessment plan approved by the Examination Board; portfolio and assessment guidelines; overviews of the results per phase, and written assessment forms per learner. This approach relates to the dependability strategy of audit trail.

The major problem with qualitative research methods as well as with portfolio assessment is the required substantial time investment. At Maastricht University, we developed a portfolio assessment procedure that uses many of these strategies while at the same time aiming for optimal efficiency (Driessen et al., 2005a). This procedure is described in Box 10.

Who?

A problem that is much debated in the portfolio literature is the feasibility and acceptability of combining the roles of mentor and assessor into one person. Tigelaar et al. interviewed nine portfolio experts about their views on the use of portfolios in education (Tigelaar et al. 2004). While some of the experts agreed that the mentor is the most appropriate person to advise an assessment committee about a candidate, others argued that it is unethical for mentors to undertake the assessor role. The latter group argued that candidates must feel free to reflect on their professional development together with their mentors, knowing that the mentor will not pass any information on to others. For this reason, the majority of the experts were of the opinion that mentors should not be involved in summative assessment nor make recommendations to an assessment committee. However, there was a minority who agreed with Snyder and colleagues, who wrote that: "*The tension between assessment for support and assessment for high stakes decision making will never disappear. Still, that tension is constructively dealt with daily by teacher educators throughout the nation*" (Snyder, et al., 1998, p. 59).

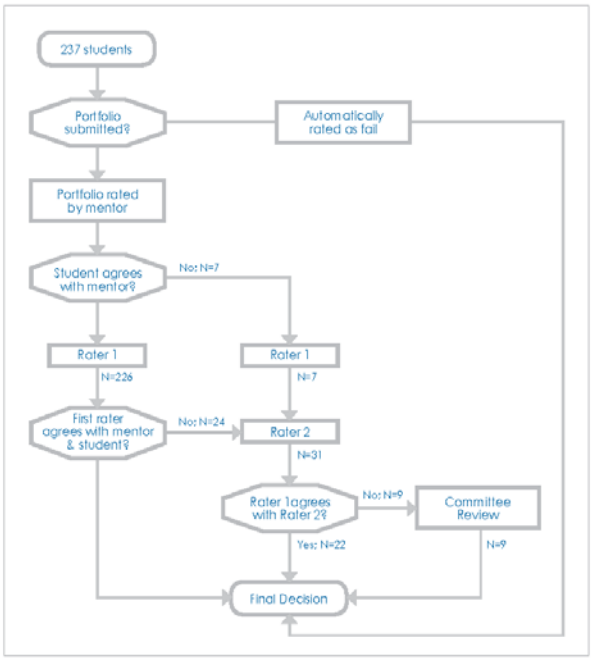
The tension between assessment for support and assessment for high stakes decision making will never disappear.

BOX 10
A procedure for portfolio assessment (Driessen et al., 2005a)

The students submits the portfolio to the mentor, who examines the portfolio and writes a recommendation regarding the grading of the portfolio to be submitted to the assessment committee.

In their final meeting of the academic year the student and the mentor discuss this recommendation. When student and mentor agree on the grade, the student signs the recommendation. If the student disagrees, he or she does not sign.

Subsequently, the portfolio is submitted to the assessment committee. This committee consists of all the mentors. The committee members do not grade the portfolios of the students they mentored themselves. Portfolios on which student and mentor agree are rated by one committee member, who does not study the portfolio in any great detail, but typically scans the work of the student and mentor and checks whether all procedures have been followed correctly. When rater and mentor agree on the grading, the recommendation becomes the final decision.



Striking the right balance between support and judgement is the challenge facing assessors/mentors with whom learners talk about their portfolios. A number of scenarios can be chosen in a procedure (Box 11). Which one is the most appropriate depends, amongst other things, on the educational context and the level of experience of the learners in question.

When? The answer to the question "when is this assessed?" depends on the answers to the other questions in this section.

Decisions about *selection* are made before the actual start of a programme or training period or after a first "trial" period, in which learners are observed and can prove themselves. The important question is whether a prospective learner matches the criteria for admission and whether this learner has the potential to finish an education or training programme.

Diagnostic assessment can be a frequent occurrence during an education or training programme. In fact, every time a mentor and a learner meet to discuss the learner's progress using information from the learner's portfolio, it can be qualified as diagnostic/ formative assessment. This implies that having easy access to a portfolio, for instance on-line, can be very helpful for mentors.

Decisions about *certification* are made when a learner's competencies match all the criteria or when the time available for a programme has run out. In an outcome based programme, this means that when the learner and his or her mentor conclude that the learner's competence meets all the criteria an assessment for certification purposes can take place. The logical consequence would be that if a person meets the competency criteria on entering an educational or training programme, he or she is exempt from training and awarded a certificate right away.

BOX 11
Portfolio assessors: scenarios

Combining the role of the mentor and assessor is often considered problematic. On the hand, most people will agree that the mentor is probably the person who is best informed about the learner's competencies. As a consequence, ignoring the mentor's opinion in assessing the portfolio can be considered as missing the chance to improve the validity of the assessment. On the other hand, combining the roles of assessor and mentor can put a strain on the relationship between mentor and learner, because learners may be reluctant to discuss any difficulties they are facing for fear of repercussions in the assessment. Below we use the metaphors of the mentor as teacher, PhD supervisor, driving instructor, and coach to distinguish between four (non exclusive) scenarios. When mentors are in the role of a teacher, their role of assessor is most prominent. When they are in the role of a coach, they do not assess at all.

The teacher: This is the most common assessment scenario in education. Just like most teachers in primary, secondary, and higher education, mentors discuss their learners' performance and progress and assess their level of competence at the end of a course.

PhD supervisor: In some scenarios the role of the mentors in the assessment procedure of portfolios can be compared with the role of supervisors of PhD students. In many countries, the formal assessment of theses/portfolios is the responsibility of a committee. Supervisors invite their peers to sit on the committee but they themselves are not a member of the committee. A negative assessment of the thesis/portfolio would harm their reputation among their peers. For this reason they are highly unlikely to invite their peers to sit on the committee unless they are convinced the portfolio meets the criteria. As a consequence, mentors and students have the same interest: to produce a thesis or portfolio that merits a positive judgment.

Driving license instructor: In this model the roles of the mentor and the assessor are strictly separated. The mentor/driving instructor mentors the learner in acquiring the required competencies, which are shown in the portfolio. If the mentor thinks the learner is competent, he invites an assessor from a professional body (i.e. the examiner from the Driver and Vehicle Licensing Agency) to assess the competence of the learner result. The learners can also approach the licensing agency themselves.

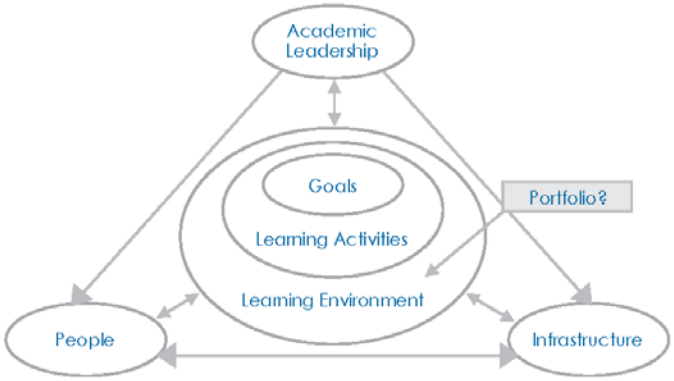
Coach: In this model, the learners themselves have the initiative. They can ask, for instance, a senior colleague to coach them until they have achieved the required level of competence. This scenario is likely, for instance, when a professional wants to acquire an additional qualification. The assessor would be someone from an external body.

Factors influencing the success of the introduction of a portfolio[2]

In the previous sections, we have argued that it is important to tailor portfolios to the intended purposes and to introduce portfolios only in situations in which they can serve a useful purpose. However, these conditions do not suffice to guarantee a successful introduction. In the literature on educational change, winning the hearts and minds of the people involved, both teachers and learners, as well as the quality of leadership are identified as key factors for lasting educational improvement (Martin et al. 2003; Hargreaves & Fink, 2004;).

Figure 4 presents a model in which portfolios are presented as part of the learning environment and in which three conditional factors are presented that influence whether an educational portfolio is introduced successfully or not: people (the teachers and learners), leadership, and infrastructure. The importance of these three conditional factors is discussed below.

FIGURE 4
Model of factors influencing the successful introduction of portfolios in education (van Tartwijk et al., 2007)



People

Educational innovations involving the use of portfolios usually imply a transfer from teacher-directed education with a strong focus on conveying knowledge, to education in which the development of students' competencies in the workplace is emphasised. In most cases, teachers are expected to invest more time and effort in coaching and assessment than they were used to. Almost inevitably, this change in roles and routines will cause uncertainty and evoke resistance (Hammerness et al., 2005). Not only does it imply that teachers need to rethink key ideas, practices, and values, but for many teachers it also means that they need to invest in developing new competencies for coaching and assessment.

Educational innovations involving the use of portfolios usually imply a transfer from teacher-directed education with a strong focus on conveying knowledge, to education in which the development of students' competencies in the workplace is emphasised.

2 Parts of this chapter were published before in Quality in Higher Education (van Tartwijk, et al., 2007)

In discussions about these innovations, the important questions are which educational problems need to be resolved and what is the most effective and efficient way to do that. Very often however, discussions concentrate on the portfolio, which becomes the visible "symbol" of the innovation. As a consequence, resistance to the innovation is likely to be projected onto the portfolio, while the important questions are not discussed.

Teachers are more likely to support and invest in educational changes if they acknowledge and subscribe to the educational value of the new learning approach, internalise and support the innovation, and are empowered to assume ownership of it. They are more likely to do so when it is clear to them how the innovation helps solve concrete problems that they have to cope with in their everyday teaching practice (Hargreaves et al. 1998). The risk that the important questions are not discussed can be reduced if teachers are involved in educational innovations at an early stage of decision-making. They are more likely to support and invest in working with a portfolio if the decision to work with this instrument was their own decision, based on their personal understanding and endorsement of the educational innovation and the role of the portfolio in it. From this perspective, the option should be kept of not using a portfolio when a better alternative is found. Teachers who have had a say in the decision to use a portfolio will feel a stronger commitment to it and will be more inclined to look for solutions and less likely to lay the instrument aside when faced with problems and inevitable design faults in the curriculum and the portfolio.

In the literature on educational change the importance of teachers as change agents is emphasised (Darling-Hammond et al., 2005) but the input of learners is crucial too. The successful introduction of a portfolio in education also depends on how much time and energy learners are willing to invest in their portfolios. In general, learners will only put effort into portfolios if this effort is rewarded in some way. The most obvious reward is that the portfolio is graded. In education, a very strong relationship exists between summative assessment and learning: assessment drives learning (Frederiksen, 1984; Driessen & van der Vleuten, 2000; van der Vleuten et al., 2000). Although assessment influences whether learners accept and put effort into a portfolio, assessment in itself is not enough. For learners, developing a portfolio implies putting a lot of effort into making their development visible. Thus, it is very frustrating for them if they discover that nobody takes a good look at the result of all their hard work. Mentors who take an interest in learners and their portfolios have been found to be a key factor in learners' appreciation of working with portfolios (Pearson & Heywood, 2004; Tigelaar et al. 2006).

Although assessment influences whether learners accept and put effort into a portfolio, assessment in itself is not enough. For learners, developing a portfolio implies putting a lot of effort into making their development visible.

A last condition for a successful introduction of portfolios related to learners and their mentors is their *understanding* of the portfolio and of what working with portfolios entails. Experience has shown that, although in theory portfolios can have a clear function in education, in practice the introduction of portfolios often leads to confusion and, consequently, frustration (Anderson & DeMeulle, 1998; Pearson & Heywood, 2004; Kjaer, et al., 2006; Davis et al. 2009). Most students who enrol in a medical curriculum are accustomed to teacher directed education. Self-assessment, asking for feedback, reflection and identifying personal learning needs, which are fundamental to portfolio learning (Snadden & Thomas, 1998b; Driessen et al. 2008), are perceived as

strange and sometimes even threatening by learners for whom education is synonymous with lectures and exams. Instructions are necessary that not only explain how to work with a portfolio, but also help learners and their mentors understand what a portfolio is and why it used in education. A study by Duque and colleagues (Duque et al., 2006) demonstrated that hands-on introduction with a proper briefing of learners by staff on the portfolio's purpose and procedures had a positive effect on portfolio scores and learner satisfaction with the portfolio. We have experimented with the use of the analogy between a portfolio and a CV to help learners better understand what a portfolio is and what working with a portfolio entails (van Tartwijk et al. 2008).

Academic leadership

Commitment by educational leaders is another vital condition for the successful introduction of portfolios. In a study on perceptions of leadership in academic contexts, Martin and her colleagues (2003) found that the quality of student learning is affected by the way leadership is constituted and experienced in academic contexts. A group of educational leaders was identified who were successful in stimulating teachers to adopt a student-focused approach to teaching. A characteristic of these educational leaders is that they discuss and negotiate these changes with the teachers. Similar findings are reported by Bland and her colleagues (2000), who reviewed the available literature with the aim to identify a set of characteristics that are associated with successful curricular change in medical education. They write that leadership comes up again and again as critical to the success of curricular change. The literature shows that successful and less successful leaders in medical education use organizational authority at about the same rate, but also that successful leaders more often seek input from others. When educational innovations ask teachers to change their roles and routines, these teachers must know that they can rely on educational leaders who support and value their commitment in every respect (Malden, 1994; van Veen et al. 2005). And finally, of course, commitment of the academic leaders is also reflected in the allocation of sufficient financial resources to ensure that the intended changes can actually be implemented.

Infrastructure

An increasing number of Faculties of Medicine are choosing to work with electronic rather than paper portfolios. In the section on e-portfolios, we described the reasons for this choice. We also wrote that research shows that adverse conditions like limited computer access in the workplace may cancel out the advantages of an e-portfolio. In general we conclude that e-portfolios are vulnerable to adverse conditions, because the demands of the technical infrastructure are large. If the electronic part of the portfolio system malfunctions, that is usually all the excuse that the adversaries of the use of portfolios need to drop the idea of a portfolio altogether, including the curriculum innovation for which the portfolio very often is a symbol.

Concluding remarks

In curricula with a strong focus on the development and assessment of competencies a portfolio can be a valuable instrument. They have the potential to make learning visible on the *Does* level of Miller's pyramid (Miller 1990), which describes independent performance in the workplace. However, portfolios are also vulnerable. Portfolio learning requires reflection by learners and investment in coaching by teachers. The quality of portfolio assessment depends on investing in the interpretation of and discussion about qualitative data. Not only does it require a new perspective on education from mentors and learners, many of whom are used to teacher-directed learning with a strong emphasis on the acquisition of knowledge, it also asks teachers and learners for a significant investment of time and energy. The literature shows that many conditions need to be fulfilled to enable successful introduction of a portfolio (Driessen et al., 2007b), and even then a portfolio is not a cure for all pains.

Portfolio learning requires reflection by learners and investment in coaching by teachers. The quality of portfolio assessment depends on investing in the interpretation of and discussion about qualitative data.

We conclude this Guide for using portfolios for assessment and learning by referring to Spandel once more (Spandel, 1997), who wrote:

> "..... *introducing portfolios is just like buying shoes: the best choice depends on purpose and comfort comes with wearing*".

"..... introducing portfolios is just like buying shoes: the best choice depends on purpose and comfort comes with wearing".

We would like to add that portfolios are like expensive shoes and even during the process of getting used to them, there will inevitably be times when one's toes are really hurting. However, for those owners who persist, the portfolio has the potential to become one of their best purchases.

References

ANDERSON RS & DEMEULLE L (1998). Portfolio use in twenty-four teacher education programs. *Teacher Education Quarterly*, 25: 23-32.

BIRD T (1990). The schoolteacher's portfolio: an essay on possibilities. In: J Millman & L Darling-Hammond (Eds), *The new handbook of teacher evaluation: Assessing elementary and secondary school teachers*, pp. 241-256 (Newbury Park, CA, Corwin Press, inc).

BLAND CJ, STARNAMAN S, WERSAL L, MOORHEAD-ROSENBERG L, ZONIA S & HENRY R (2000). Curricula change in medical schools: How to succeed. *Academic Medicine*, 75: 575-594.

BRANSFORD J, BROWN AL & COCKING RR (Eds) (2000). *How people learn: Brain, mind, experience, and school*. (Washington D.C., National Academy Press).

BUCKLEY S, ASHCROFT T, DAVIS J, KHAN KS, MORLEY D, POLLARD D, POPOVIC C, SAYERS J, SUSARLA R, THOMAS H & ZAMORA J (in press). The educational effects of portfolios on undergraduate student learning: A Best Evidence Medical Education systematic review, *Medical Teacher*.

COLLINS A (1991). Portfolios for biology teacher assessment. *Journal of Personnel Evaluation in Education*, 5: 147-167.

CONLON M (2003). Appraisal: The catalyst of personal development. *British Medical Journal*, 327: 389-391.

DARLING-HAMMOND L, PACHECO A, MICHELLI N, LEPAGE P, HAMMERNESS K & YOUNG P (2005). Implementing curriculum renewal in teacher education: managing organizational and policy change. In: L Darling-Hammond, J Bransford, P LePage, K Hammerness & H Duffy (Eds), *Preparing teachers for a changing world: What teachers should learn and be able to do*, pp. 442-479 (San Francisco, Jossey-Bass).

DAVIS DA, MAZMANIAN PE, FORDIS M, VAN HARRISON R, THORPE KE & PERRIER L (2006). Accuracy of physician self-assessment compared with observed measures of competence: a systematic review. *JAMA*, 296: 1094-1102.

DAVIS MH, FRIEDMAN BEN DAVID M, HARDEN RM, HOWIE P, KER J, MCGHEE C, et al. (2001). Portfolio assessment in medical students' final examinations. Medical Teacher, 23: 357-366.

DAVIS MH, PONNAMPERUMA GG, & KER JS (2009). Student perceptions of a portfolio assessment process. *Medical Education*, 43: 89-98.

DENZIN NK (1978). *Sociological Methods: A Sourcebook* (2nd ed.). New York: McGraw Hill.

DENZIN NK & LINCOLN YS (2000). *Handbook of qualitative research*. Thousand Oaks, CA: Sage.

DORNAN T, CARROLL C & PARBOOSHING J (2002). An electronic learning portfolio for reflective continuing professional development. *Medical Education*, 36: 767-769.

DREYFUS SE (2004). The five-stage model of adult skill acquisition. *Bulletin of Science Technology and Society*, 24: 117-181.

DRIESSEN EW, MUIJTJENS AMM, VAN TARTWIJK J & VAN DER VLEUTEN CPM (2007a). Web- or paper-based portfolios: is there a difference? *Medical education*, 41: 1067-1073.

DRIESSEN EW & VAN DER VLEUTEN CPM (2000). Matching student assessment to problem based learning: lessons from experience in a law faculty. *Studies in Continuing Education*, 22: 235-248.

DRIESSEN EW, VAN DER VLEUTEN CPM, SCHUWIRTH L, VAN TARTWIJK J & VERMUNT JD (2005a). Credibility of portfolio assessment as an alternative for reliability evaluation: a case study. *Medical Education*, 39: 214-220.

DRIESSEN EW, VAN TARTWIJK J & DORNAN T (2008). The self-critical doctor: Helping students become more reflective. *BMJ*, 336: 827-830.

DRIESSEN EW, VAN TARTWIJK J, OVEREEM K, VERMUNT JD & VAN DER VLEUTEN CPM (2005b). Conditions for successful reflective use of portfolios in undergraduate medical education. *Medical Education*, 39: 1230-1235.

DRIESSEN EW, VAN TARTWIJK J, VAN DER VLEUTEN CPM, & WASS V (2007b). Portfolios in medical education: Why do they meet with mixed success? A systematic review. *Medical Education*, 41: 1224-1233.

DRIESSEN EW, VAN TARTWIJK J, VERMUNT JD & VAN DER VLEUTEN CPM (2003). Use of portfolio in early undergraduate medical training. *Medical Teacher*, 25: 18-23.

DUQUE G, FINKELSTEIN A, ROBERT A, TABATABIA D, GOLD SL & WINER LR (2006). Learning while velauating: the use of an electronic evaluation portfolio in a geraitric medicine clerckship. *BMC Medical Education*, 6: 1-7.

ERICSSON KA (2006). The influence of experience and deliberate practice on the development of expert performance. In: KA Ericsson, N Charness, PJ Feltovich & RR Hoffman (Eds), *The Cambridge handbook of expertise and expert performance* (pp. 683-704). New York: Cambridge University Press.

EVA KW & REGEHR G (2008). "I'll never play professional football" and other fallacies of self-assessment. *Journal of Continuing Education in the Health Professions*, 28: 14-19.

FINLAY IG, MAUGHAN TS & WEBSTER DJ (1998). A randomized controlled study of portfolio learning in undergraduate cancer education. *Medical Education*, 32: 172-176.

FREDERIKSEN N (1984). The real test bias: Influences of testing on teaching and learning. American Psychologist, 39: 193-202.

FRIEDMAN BEN DAVID M, DAVIS MH, HARDEN RM, HOWIE PW, KER J & PIPPARD MJ (2001). *AMEE Medical Education Guide No. 24: Portfolios as a method of student assessment* (Dundee, Association for Medical Education in Europe).

FUNG KEE FUNG M, WALKER M, FUNG KEE FUNG K, TEMPLE L, LAJOIE F, BELLEMARE G, et al. (2000). An Internet-based learning portfolio in resident education: The KOALA-super (TM) multicentre programme. *Medical Education*, 34: 474-479.

GENERAL MEDICAL COUNCIL (2000). *Revalidating doctors: Ensuring standards, securing the future*. London: GMC.

GIBSON D & BARRETT H (2003). Directions in Electronic Portfolio Development. *Contemporary Issues in Technology and Teacher Education*, 2: 559-576.

GRANT AJ, VERMUNT JD, KINNERSLEY P & HOUSTON H (2007). Exploring students' perceptions of the use of a significant event analysis as part of a portfolio assessment process in general practice, as a tool for learning how to use ferlection in learning. *BMC Medical Education*: 7:5.

GUBA EG & LINCOLN YS (1989). Judging the quality of fourth generation evaluation. In: EG Guba & YS Lincoln (Eds), *Fourth Generation Evaluation* (London, Sage).

HAMMERNESS K, DARLING-HAMMOND L, BRANSFORD J, BERLINER DC, COCHRAN-SMITH M, MCDONALD M, et al. (2005). How teachers learn and develop. In: L Darling-Hammond, J Bransford, P LePage, K Hammerness & H Duffy (Eds), *Preparing teachers for a changing world: What teachers should learn and be able to do*, pp. 358-389 (San Francisco, Jossey-Bass).

HARDEN RM (1979). How to assess students: An overview. *Medical Teacher*, 1: 65-70.

HARGREAVES A & FINK D (2004). The seven principles of sustainable leadership. *Educational Leadership*, April 2004: 8-13.

HARGREAVES A, LIEBERMAN A, FULLAN M & HOPKINS D (Eds) (1998). *International handbook of educational change* (Dordrecht: Kluwer Academic Publishers).

HATTON N & SMITH D (1995). Reflection in teacher education: Towards definition and implementation. *Teaching and Teacher Education*, 11: 33-49.

KJAER NK, MAAGARD R & WIES S (2006). Using an online portfolio in postgraduate training. *Medical Teacher*, 28: 708-712.

KORTHAGEN FAJ, KESSELS J, KOSTER B, LAGERWERF B & WUBBELS T (2001). *Linking theory and practice: The pedagogy of realistic teacher education* (Mahwah, NY, Lawrence Erlbaum Associates).

KORTHAGEN FAJ, KOSTER B, MELIEF K & TIGCHELAAR A (2002). *Teach teachers to reflect: Systematic reflection in the training and coaching of teachers* [In Dutch: Docenten leren reflecteren: Systematische reflectie in de opleiding en begeleiding van leraren] (Soest, Uitgeverij Nelissen).

LAWSON M, NESTEL D & JOLLY B (2004). An e-portfolio in heath professional education. *Medical Education*, 38: 569-570.

LOCKYER JM & CLYMAN SG (2008). Multisource feedback (360-degree feedback). In: ES Holmboe & RE Hawkins (Eds), *Practical guide to the evaluation of clinical competence*, pp. 75-85 (Philadelphia, Pa, Mosby Elsevier).

LYONS N (1998). Reflection in teaching: Can it be developmental? A portfolio perspective. *Teacher Education Quarterly*, Winter 1998: 115-127.

MALDEN B (1994). The micropolitics of education: mapping the multiple dimensions of power relations in school policies. *Journal of Educational Policy*, 9: 147-167.

MANN K, GORDON J & MACLEOD A (2007). Reflections and reflective practice in health profession education: A systematic review. *Advanced Health Science Education*, (First published online November 2007): 1-27.

MARTIN E, TRIGWELL K, PROSSER M & RAMSDEN P (2003). Variations in the experience of leadership of teaching in higher education. *Studies in Higher Education*, 28: 247-259.

MATHERS NJ, CHALLIS MC, HOWE AC & FIELD NJ (1999). Portfolios in continuing medical education – effective and efficient? *Medical Education*, 33: 521-530.

MILLER GE (1990). The assessment of clinical skills/competence/performance. *Academic Medicine*, 65: S63-67.

NORCINI JJ & BURCH VC (Eds) (2007). *Workplace-based assessment as an educational tool*, AMEE Guide 31 (Dundee, UK, AMEE).

NORCINI JJ, HOLMBOE ES & HAWKINS RE (2008). Evaluation challenges in the era of outcome based education. In: ES Holmboe & RE Hawkins (Eds), *Practical guide to the evaluation of clinical competence*, pp. 1-9 (Philadelphia, PA, Mosby Elsevier).

O'SULLIVAN PS, RECKASE MD, MCCLAIN T, SAVIDGE MA & CLARDY JA (2004). Demonstration of portfolios to assess competency of residents. *Advances in Health Sciences Education*, 9: 1-15.

OERMANN MH (2002). Developing a professional portfolio in Nursing. *Orthopaedic Nursing*, 21: 73-78.

PAULSON FL, PAULSON PR & MEYER CA (1991). What makes a portfolio a portfolio? Eight thoughtful guidelines will help educators encourage self directed learning. *Educational Leadership*, February 1991: 60-63.

PEARSON DJ & HEYWOOD P (2004). Portfolio use in general practice vocational training: A survey of GP registrars. *Medical Education*, 38: 87-95.

ROYAL COLLEGE OF GENERAL PRACTITIONERS (1993). *Portfolio-based learning in general practice: Report of a working group on higher professional education*, Occasional paper 63 (London, Royal College of General Practitioners).

ROYAL COLLEGE OF PHYSICIANS AND SURGEONS OF CANADA (1996). *Canmeds 2000 Project: Skills for the New Millennium. Report on the societal needs working group* (Ottawa, The Royal College of Physicians and Surgeons of Canada).

SHULMAN LS (1998). Teacher portfolios: a theoretical activity. In: N Lyons (Ed), *With portfolio in hand: validating the new teacher professionalism*, pp. 23-38 (New York, Teachers College Press).

SHUTE VJ (2008). Focus on formative feedback. *Review of Educational Research*, 78: 153-189.

SMITHER JW, LONDON M, FLAUTT R, VARGAS Y & KUCINE I (2003). Can working with an executive coach improve multisource feedback ratings over time? A quasi-experimental field study. *Personal Psychology*, 56: 23-44.

SNADDEN D (1999). Portfolios – attempting to measure the unmeasurable? [Commentary]. *Medical Education*, 33(7): 478-479.

SNADDEN D, CHALLIS M, & THOMAS ML (1999). *AMEE Medical Education Guide No. 11: Portfolio-based learning and assessment* (Dundee, Association for Medical Education in Europe).

SNADDEN D & THOMAS ML (1998a). Portfolio learning in general practice vocational training - does it work? *Medical Education*, 32: 401-406.

SNADDEN D & THOMAS ML (1998b). The use of portfolio learning in medical education. *Medical Teacher*, 20: 192-199.

SNYDER J, LIPPINCOTT A & BOWER D (1998). The inherent tensions in the multiple uses of portfolios in teacher education. *Teacher Education Quarterly*, 25: 45-60.

SPANDEL V (1997). Reflections on portfolios. In: GD Phye (Ed), *Handbook of academic learning: Construction of knowledge* (pp. 573-591). San Diego: Academic Press.

STOOF A, MARTENS RL, VAN MERRIËNBOER J & BASTIAENS TJ (2002). The boundary approach of competence: a constructivist aid for understanding and using the concept of competence. Human resource development review, 1, pp. 345-365.

TIGELAAR DEH, DOLMANS DHJM, DE GRAVE WS, WOLFHAGEN HAP & VAN DER VLEUTEN CPM (2006). Participants opinions about the usefulness of a teaching portfolio. *Medical Education*, 40(4): 371-378.

TIGELAAR DEH, DOLMANS DHJM, WOLFHAGEN HAP & VAN DER VLEUTEN CPM (2004). Using a conceptual framework and the opinion of portfolio experts to develop a teaching portfolio prototype. *Studies in Educational Evaluation*, 30: 305-321.

TIGELAAR DEH, DOLMANS DHJM, WOLFHAGEN HAP & VAN DER VLEUTEN CPM (2005). Quality issues in judging portfolio: implications for organizing teaching portfolio assessment procedures. *Studies in Higher Education*, 30: 595-610.

TOCHEL C, HAIG A, HESKETH A, CADZOW A, BEGGS K, COLTHART L, et al. The effectiveness of portfolios for post-graduate assessment and education: a Best Evidence Medical Education systematic review. *Medical Teacher* (in press).

VAN DER VLEUTEN CPM, DOLMANS DHJM & SCHERPBIER AJJA (2000). The need for evidence in education. *Medical Teacher*, 22: 246-250.

VAN MANEN M (1977). Linking ways of knowing with ways of being practical. *Curriculum Inquiry*, 6: 205-228.

VAN TARTWIJK J, DRIESSEN EW, STOKKING K & VAN DER VLEUTEN CPM (2007). Factors influencing the successful introduction of portfolios. *Quality in Higher Education*, 13: 69-79.

VAN TARTWIJK J, VAN RIJSWIJK M, TUITHOF H & DRIESSEN EW (2008). Using an analogy in the introduction of a portfolio. *Teaching and Teacher Education*, 24: 927-938.

VAN VEEN K, SLEEGERS P, & VAN DE VEN P (2005). One teacher's identity, emotions, and commitment to change: A case study into the cognitive-affective processes of a secondary school teacher in the context of reforms. *Teaching and Teacher Education*, 21: 917-934.

WEBB C, ENDACOTT R, GRAY MA, JASPER MA, MCCULLAN M & SCHOLES J (2003). Evaluating portfolio assessment systems: What are the appropriate criteria? *Nurse Education Today*, 23: 600-609.

WOODWARD H & NANLOHY P (2004). Digital portfolios: Fact or fashion. *Assessment & Evaluation in Higher Education*, 29: 227-238.

Series 2

30 Peer Assisted Learning: a planning and implementation framework
Michael Ross & Helen Cameron (2007)
ISBN: 978-1-903934-38-8
Primarily designed to assist curriculum developers, course organisers and educational researchers develop and implement their own PAL initiatives.

31 Workplace-based Assessment as an Educational Tool
John Norcini & Vanessa Burch (2008)
ISBN: 978-1-903934-39-5
Several methods for assessing work-based activities are described, with preliminary evidence of their application, practicability, reliability and validity.

32 e-Learning in Medical Education
Rachel Ellaway & Ken Masters (2008)
ISBN: 978-1-903934-41-8
An increasingly important topic in medical education – a 'must read' introduction for the novice and a useful resource and update for the more experienced practitioner.

33 Faculty Development: Yesterday, Today and Tomorrow
Michelle McLean, Francois Cilliers & Jacqueline M van Wyk (2010)
ISBN: 978-1-903934-42-5
Useful frameworks for designing, implementing and evaluating faculty development programmes.

34 Teaching in the clinical environment
Subha Ramani & Sam Leinster (2008)
ISBN: 978-1-903934-43-2
An examination of the many challenges for teachers in the clinical environment, application of relevant educational theories to the clinical context and practical teaching tips for clinical teachers.

35 Continuing Medical Education
Nancy Davis, David Davis & Ralph Bloch (2010)
ISBN: 978-1-903934-44-9
Designed to provide a foundation for developing effective continuing medical education (CME) for practicing physicians.

36 Problem-Based Learning: where are we now?
David Taylor & Barbara Miflin (2010)
ISBN: 978-1-903934-45-6
A look at the various interpretations and practices that claim the label PBL, and a critique of these against the original concept and practice.

37 Setting and maintaining standards in multiple choice examinations
Raja C Bandaranayake (2010)
ISBN: 978-1-903934-51-7
An examination of the more commonly used methods of standard setting together with their advantages and disadvantages and illustrations of the procedures used in each, with the help of an example.

38 Learning in Interprofessional Terms
Marilyn Hammick, Lorna Olckers & Charles Campion-Smith (2010)
ISBN: 978-1-903934-52-4
Clarification of what is meant by Interprofessional learning and an exploration of the concept of teams and team working.

39 Online eAssessment
Reg Dennick, Simon Wilkinson & Nigel Purcell (2010)
ISBN: 978-1-903934-53-1
An outline of the advantages of on-line eAssessment and an examination of the intellectual, technical, learning and cost issues that arise from its use.

40 Creating effective poster presentations
George Hess, Kathryn Tosney & Leon Liegel (2009)
ISBN: 978-1-903934-48-7
Practical tips on preparing a poster – an important, but often badly executed communication tool.

41 The Place of Anatomy in Medical Education
Graham Louw, Norman Eizenberg & Stephen W Carmichael (2010)
ISBN: 978-1-903934-54-8
The teaching of anatomy in a traditional and in a problem-based curriculum from a practical and a theoretical perspective.

42 The use of simulated patients in medical education
Jennifer A Cleland, Keiko Abe & Jan-Joost Rethans (2010)
ISBN: 978-1-903934-55-5
A detailed overview on how to recruit, train and use Standardized Patients from a teaching and assessment perspective.

43 Scholarship, Publication and Career Advancement in Health Professions Education
William C McGaghie (2010)
ISBN: 978-1-903934-50-0
Advice for the teacher on the preparation and publication of manuscripts and twenty-one practical suggestions about how to advance a successful and satisfying career in the academic health professions.

44 The Use of Reflection in Medical Education
John Sandars (2010)
ISBN: 978-1-903934-56-2
A variety of educational approaches in undergraduate, postgraduate and continuing medical education that can be used for reflection, from text based reflective journals and critical incident reports to the creative use of digital media and storytelling.

45 Portfolios for Assessment and Learning
Jan van Tartwijk & Erik W Driessen (2010)
ISBN: 978-1-903934-57-9
An overview of the content and structure of various types of portfolios, including eportfolios, and the factors that influence their success.

46 Student Selected Components
Simon C Riley (2010)
ISBN: 978-1-903934-58-6
An insight into the structure of an SSC programme and its various important component parts.

47 Using Rural and Remote Settings in the Undergraduate Medical Curriculum
Moira Maley, Paul Worley & John Dent (2010)
ISBN: 978-1-903934-59-3
A description of an RRME programme in action with a discussion of the potential benefits and issues relating to implementation.

48 Effective Small Group Learning
Sarah Edmunds & George Brown (2010)
ISBN: 978-1-903934-60-9
An overview of the use of small group methods in medicine and what makes them effective.

To see the full list of guides available, and to order, see the website www.amee.org.

About AMEE

What is AMEE?

AMEE is an association for all with an interest in medical and healthcare professions education, with members throughout the world. AMEE's interests span the continuum of education from undergraduate/basic training, through postgraduate/specialist training, to continuing professional development/continuing medical education.

- **Conferences:** Since 1973 AMEE has been organising an annual conference, held in a European city. The conference now attracts over 2300 participants from 80 countries.
- **Courses:** AMEE offers a series of courses at AMEE and other major medical education conferences relating to teaching, assessment, research and technology in medical education.
- **MedEdWorld**: AMEE's exciting new initiative has been established to help all concerned with medical education to keep up to date with developments in the field, to promote networking and sharing of ideas and resources between members and to promote collaborative learning between students and teachers internationally.
- **Medical Teacher:** AMEE produces a leading international journal, Medical Teacher, published 12 times a year, included in the membership fee for individual and student members.
- **Education Guides:** AMEE also produces a series of education guides on a range of topics, including Best Evidence Medical Education Guides reporting results of BEME Systematic Reviews in medical education.
- **Best Evidence Medical Education (BEME):** AMEE is a leading player in the BEME initiative which aims to create a culture of the use of best evidence in making decisions about teaching in medical and healthcare professions education.

Membership categories

- **Individual and student members (£85/£39 a year):** Receive Medical Teacher (12 issues a year, hard copy and online access), free membership of MedEdWorld, discount on conference attendance and discount on publications.
- **Institutional membership (£200 a year):** Receive free membership of MedEdWorld for the institution, discount on conference attendance for members of the institution and discount on publications.

See the website (www.amee.org) for more information.

If you would like more information about AMEE and its activities, please contact the AMEE Office:
Association for Medical Education in Europe (AMEE), Tay Park House, 484 Perth Road, Dundee DD2 1LR, UK
Tel: +44 (0)1382 381953; Fax: +44 (0)1382 381987; Email: amee@dundee.ac.uk

www.amee.org

16 March 2017

OSCE

Kasana Raksamani, MD, MHPE
Department of Anesthesiology
Faculty of Medicine Siriraj Hospital

Objectives

- Describe the characteristics of OSCE
- Discuss the advantages and disadvantages of OSCE
- Develop and conduct an OSCE

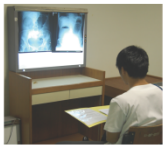
OSCE

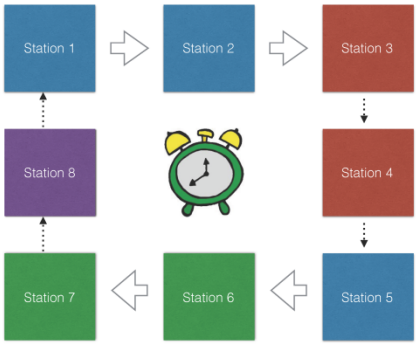
O bjective

S tructured

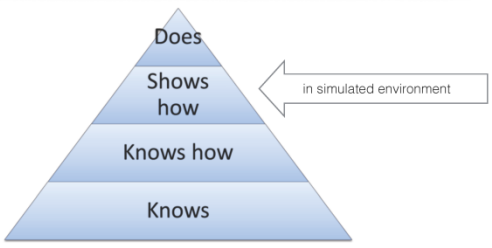
C linical

E xamination





Miller's Pyramid



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.

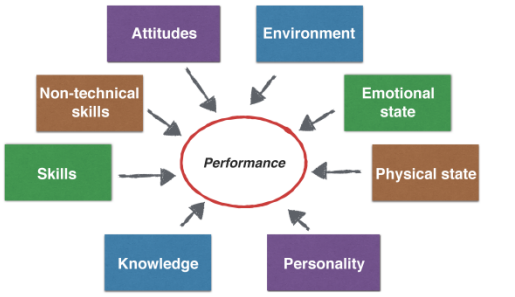
Competency based assessment of performance

- **Process** - various skills, attitudes
- **Product** – data interpretation, diagnosis, problem solving, report writing, order sheet writing, drug description
- **Mixed** – process and product

Advantage of OSCE

- Can assess clinical skills, technical skills, communication skills
- A realistic but safe environment
- Standardization and control the complexities of the case
- The encounter can be recorded, reviewed, and used for feedback
- Valid and reproducible

Factors influencing performance



Factors influencing performance. Modified from Khan & Ramachandran (2012)

Disadvantage of OSCE

- **Expensive**
- Time and resource consuming
- Difficult to administer
- Many potential sources of construct-irrelevant variance: SPs, raters, cases, scoring sheets
- Construct underrepresentation
  - Knowledge and skills are tested in compartments

Steps of developing an OSCE

The key to a successful OSCE = careful planning!

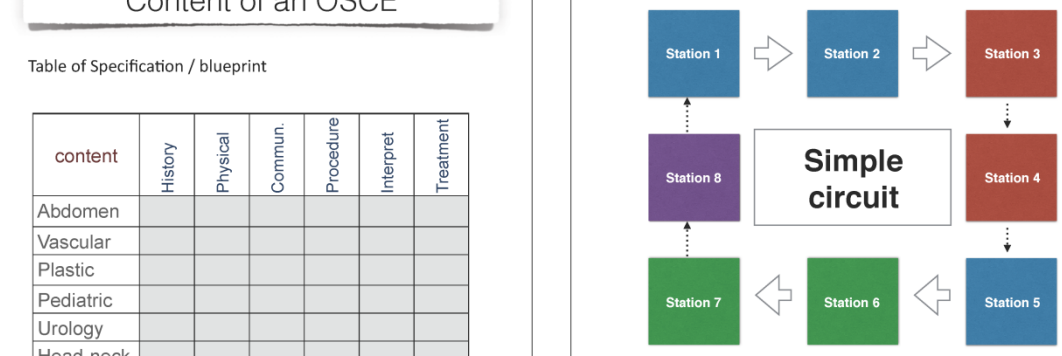
Steps of developing an OSCE

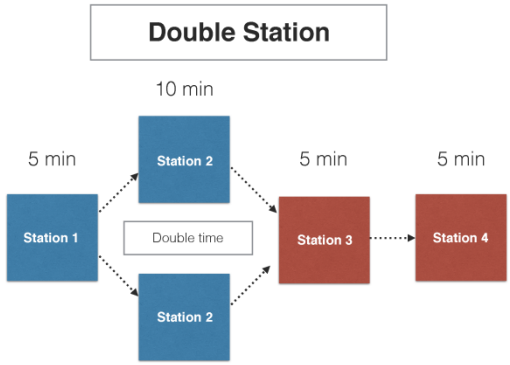
1. Define objectives
2. Choose the type of an OSCE item (process/product)
3. Write an item
4. Prepare instruments
5. SP recruitment and training
6. Item review committee
7. Item pool management

Content of an OSCE

Table of Specification / blueprint

content	History	Physical	Commun.	Procedure	Interpret	Treatment
Abdomen						
Vascular						
Plastic						
Pediatric						
Urology						
Head-neck						





Write an item

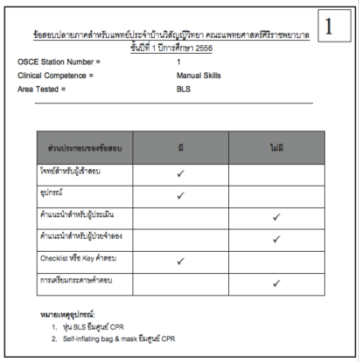
- ข้อสอบ OSCE 1 ข้อต้องมีอะไรบ้าง

Component of an OSCE

1. Scenario
2. Instruction for examinees
3. Instruction for SPs
4. Scoring rubric

Scenario

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



Scenario

- หัวข้อ : การตรวจร่างกายผู้ป่วยที่มีอาการปวดท้อง
- Objective : นักศึกษาแพทย์สามารถแสดงวิธีการตรวจร่างกายผู้ป่วยที่มีอาการปวดท้องเฉียบพลัน และให้การวินิจฉัยที่ถูกต้องได้
- ผู้สอบ: นักศึกษาแพทย์ชั้นปีที่ 6
- สถานการณ์: สมบูรณ์ อายุ 35 ปี มีอาการปวดท้องใต้ชายโครงด้านซ้าย 6 ชั่วโมง มีอาการหลังดื่มสุรา 2 ชั่วโมง ปวดตื้อๆตลอดเวลา ปวดร้าวไปที่กลางหลัง
- คำสั่ง : จงแสดงวิธีการตรวจหน้าท้องผู้ป่วย บรรยายสิ่งที่ตรวจพบและให้การวินิจฉัยโรคที่คิดถึงมากที่สุด 1 โรค
- เวลา : 5 นาที (ตรวจร่างกาย 4 นาทีครึ่ง บอกสิ่งตรวจพบและการวินิจฉัย ครึ่งนาที)

Scenario

Apparatus

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

Instruction for examinees

- ผู้ป่วยหญิงไทยคู่ อายุ 22 ปี มีอาการปวดท้อง 4 ชั่วโมงก่อนมาโรงพยาบาล จงซักประวัติผู้ป่วยรายนี้ และให้การวินิจฉัยโรคที่นึกถึงมากที่สุด (ซักประวัติ 4 นาทีครึ่ง บอกการวินิจฉัย ใน 30 วินาทีสุดท้าย)

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 : นักศึกษาจะทำการซักประวัติท่านเพื่อให้การวินิจฉัยโรค ให้ท่านให้ข้อมูลต่อไปนี้

ข้อมูลจากโจทย์ : ท่านเป็นผู้ป่วยชายไทยคู่ อายุ 40 ปี มีอาการปวดขาหนีบข้างขวา 1 วัน

การแต่งกาย: แต่งกายชุดลำลอง เป็นเสื้อ กางเกงที่สามารถเปิดหน้าท้องได้สะดวก

การตกแต่งบาดแผล -

ข้อมูลที่นักศึกษาจะซักถามจากท่าน

1. ตำแหน่งที่ปวดท้อง : ปวดบริเวณขาหนีบด้านขวา

2. ลักษณะของอาการปวด : ช่วงแรกปวดหน่วงๆ ตลอดเวลา

3. มีอาการปวดร้าวไปที่อื่นหรือไม่ : ไม่มี

4. ลักษณะของอาการปวดตอนเริ่มแรก เป็นอย่างไร เป็นทันทีทันใดหรือค่อยๆปวดเพิ่มขึ้นช้าๆ เป็นที่ตำแหน่งเดียวกันนี้หรือมีการย้ายที่ปวด : ค่อยๆปวดเพิ่มขึ้นช้าๆ ไม่มีการย้ายที่ปวด

5. มีปัจจัยใดที่ทำให้ปวดเปลี่ยนแปลงหรือไม่ : ปวดเพิ่มมากขึ้นในขณะยืนหรือไอ

Instruction for SPs

6.อาการร่วมอื่นๆ

6.1 ทั่วไป: มีไข้ต่ำๆ

6.2 ระบบทางเดินอาหาร: มีอาการปวดท้องบีบๆเป็นพักๆ ท้องอืดมากขึ้น คลื่นไส้และอาเจียน

7. ประวัติอดีต

7.1 ประวัติการมีก้อนที่ขาหนีบ

สังเกตมีก้อนที่ขาหนีบข้างขวามา 2 ปี

7.2 ประวัติการเปลี่ยนแปลงของก้อนที่ขาหนีบ

ขนาดก้อนเท่าๆเดิม จะโตมากเวลายืนหรือเบ่ง เวลานอนแล้ว เอามือดันก้อนจะยุบได้เอง

7.3 ประวัติอาการที่สัมพันธ์กับก้อน

มีปวดหน่วงเวลามีก้อนอื่น

7.4 โรคประจำตัว การผ่าตัด การใช้ยา การแพ้ยา: ไม่มี ไม่เคยพบแพทย์

8. ประวัติส่วนตัว : อาชีพ การสูบบุหรี่ การดื่มสุรา

- ทำงานเป็นเสมียน สูบบุหรี่วันละ 2 ซองมา 10 ปี ไม่ดื่มสุรา

Rehearsing the SPs

- Trainer role-plays several different student approaches
- Feedback to SPs' portrayal of the case
- Revision of the script
- Standardization of the portrayal

Scoring rubric

ขั้นตอนการประเมิน	สมบูรณ์	ไม่สมบูรณ์	ไม่ปฏิบัติ
1. การแนะนำตัว			
1.1 การแนะนำตัวเองอย่างสุภาพ	5	3	0
1.2 การถามชื่อผู้ป่วยอย่างสุภาพ	5	3	0
2. การถามประวัติ			
2.1 ถามตำแหน่งที่ปวด	10	-	0
2.2 ถามลักษณะของการปวด	10	6	0
2.3 ถามอาการปวดร้าวไปที่อื่น	10	-	0
...			
2.8 ถามประวัติประจำเดือน (ประจำเดือนครั้งสุดท้าย ความสม่ำเสมอ)	10	6	0
3. การวินิจฉัยโรค			
Ectopic pregnancy	10		
Acute appendicitis	6		

Item Review

- Item content review
- Scoring criteria
- Passing score

Define a passing score

- Test-centered method
- Examinee-centered method

Test-Centered Methods

- The judges set standards by reviewing the test items and provide judgments regarding the "just adequate" level of performance on these items.
- **Modified Angoff's method**
  - The judges review the scoring sheet and determine how a "borderline" examinee should perform on each item and how much he/she will score on each item.

Scoring Rubrics

การประเมิน	Complete	Incomplete	No
1.	10	6	0
2.	5	3	0
3.	5	3	0
4.	10	6	0
5.	5	3	0
6.	5	3	0
7.	10	6	0
8.	10	6	0
9.	10	6	0
10.	5	3	0
11.	5	3	0
12.	5	3	0
13.	10	6	0
14.	5	3	0

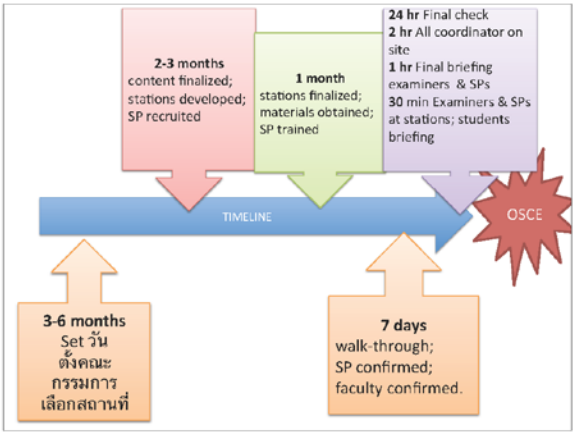
45

Examinee-Centered Methods

- The judges set a standard by reviewing the overall performance of examinees and determine who should pass and who should fail. The scores of examinees are reviewed and the passing score is set based on these judgments
  - Borderline-group method
  - Contrasting-groups method

Item pool management

- Item Pool Database
  - Content domain (specialty)
  - Skills category (History/Physical/Lab/etc.)
  - SP requirement
  - History of usage
    - Date
    - Average scores
    - Pass/fail rate



Common Problem in OSCE

- การออกข้อสอบ
- การจัดสถานีสอบ
- การคุมสอบ

การออกข้อสอบ

- เนื้อหาที่นำมาออกสอบไม่เหมาะสม
  - เน้นเนื้อหาที่เป็นปัญหาที่พบบ่อย ในเวชปฏิบัติ
  - หลีกเลี่ยงเนื้อหาความรู้ที่สามารถประเมินด้วยวิธีการอื่นได้
  - ตอบสนองเกณฑ์มาตรฐานผู้ประกอบวิชาชีพเวชกรรมของแพทยสภา
  - ให้มีการกระจายเนื้อหาครบทุกภาควิชา สาขาวิชา ทุกทักษะ (table of specification)
- โจทย์ขาดความชัดเจน อาจารย์สองท่านหรือนักศึกษาสองคนแปลโจทย์ต่างกัน
  - เขียนโจทย์แล้วควรมีการอ่านทบทวนหลังทิ้งไว้ระยะหนึ่ง
  - โจทย์ควรได้รับการพิจารณาปรับแก้โดยคณะกรรมการฯ
  - ระบุบริบท หรือ สถานที่ หรือขอบเขตของสิ่งที่ต้องปฏิบัติ ให้ชัดเจน

การออกข้อสอบ

- บทบาทของ SP ไม่ชัดเจน ผู้ป่วยจำลองสองคนปฏิบัติไม่เหมือนกัน
  - เขียนบทผู้ป่วยจำลอง ให้ครบทุกแง่มุม
  - มีการซักซ้อมการแสดงบทบาทก่อนทำการสอบจริง
  - จัดทำฐานข้อมูลคลังผู้ป่วยจำลอง

การจัดสถานีสอบ

- สถานที่สอบไม่เหมาะสม
  - เลือกสถานที่ที่เหมาะสม ไม่คับแคบเกินไป แต่ละสถานีมีฉากหรือประตูกั้น ไม่มีเสียงรบกวนระหว่างสถานี แสงสว่างเพียงพอ
- นักศึกษาเดินเข้าห้องผิด เดินผิดทิศทาง ข้ามสถานี
  - ทำป้ายบอกทิศทาง
  - มีเจ้าหน้าที่ยืนกำกับในตำแหน่งที่มีโอกาสเดินผิดพลาด
  - มีฉากหรือเชือกกั้นแนวทางเดิน
- นักศึกษาเหนื่อยจากการปฏิบัติหัตถการที่ยาก อย่างต่อเนื่อง
  - จัดสถานีสลับระหว่าง process กับ product
  - จัด ให้มีสถานีพักเป็นระยะๆ

การจัดสถานีสอบ

- อาจารย์ล้าจากการสังเกตพฤติกรรม และให้คะแนนอย่างต่อเนื่อง
  - จัดสถานีสอบ ให้มีจำนวนสถานีรวมมากกว่าจำนวนนักศึกษาที่เข้าสอบ
- เครื่องมือไม่พร้อม ไม่ครบ ไม่เหมาะสมกับหัตถการที่ต้องทำ
  - มีอาจารย์แพทย์ตรวจสอบรายการอุปกรณ์ที่ต้องใช้สำหรับข้อสอบแต่ละข้อ
  - จัดห้องสอบก่อนวันสอบหนึ่งวัน โดยมีอาจารย์เดินตรวจแต่ละสถานีถึงความพร้อมและเหมาะสมของเครื่องมือ
  - จัดอุปกรณ์สำรองสำหรับอุปกรณ์บางชนิด เช่น กล้องจุลทรรศน์, ตู้อ่าน film, slide blood smear, wet smear, ฯลฯ
  - เตรียม battery สำรอง (ophthalmoscope, laryngoscope)

การคุมสอบ

- สิ่งรบกวนสมาธิ (โทรศัพท์ เอกสารเฉลย)
  - ปิดโทรศัพท์มือถือ หรือตั้งเป็นระบบสั่น
  - แจ้งเลขาฯประจำภาควิชาหรือหน่วยฯ ให้ทราบถึงกำหนดการสอบ
  - รู้กำหนดการสอบล่วงหน้า และไม่นัดหมายซ้ำซ้อนกับภารกิจอื่น
- อาจารย์มาคุมสอบช้า
  - วางแผนการเดินทางให้ถึงห้องสอบก่อนเริ่มสอบ 15 นาที
  - จัดอาหารเช้า หรือเที่ยงที่สนามสอบสำหรับอาจารย์
  - จัด ให้มีอาจารย์กรรมการส่วนกลางที่พร้อมทำหน้าที่แทน
  - มีระบบการเตือนล่วงหน้าผ่าน SMS, โทรศัพท์

การคุมสอบ

- ผู้ป่วยจำลองมาสาย หรือไม่มา
  - มีระบบการแจ้งเตือนและยืนยันการมาเป็นผู้ป่วยจำลอง
  - จัดโต๊ะลงทะเบียนผู้ป่วยจำลอง ให้รายงานตัวก่อนเริ่มสอบ 30 นาที
  - เตรียมผู้ป่วยจำลองสำรองตามความเหมาะสม
  - มีเบอร์โทรศัพท์มือถือของผู้ป่วยจำลองทุกคนอยู่ประจำสนามสอบ
  - จัดอาหารและเครื่องดื่มบริการผู้ป่วยจำลองก่อนเริ่มสอบ 30 นาที
- ใบให้คะแนนนักศึกษาที่ไม่ระบุชื่อ
  - เน้นย้ำเรื่องการเขียนชื่อ และเลขที่กับอาจารย์ก่อนเริ่มสอบ
  - แจก sticker พิมพ์ชื่อ เลขที่ของนักศึกษาให้นักศึกษาก่อนเข้าสอบ
- เกิดความผิดพลาดในการจับเวลาแต่ละสถานี

การคุมสอบ

- มาตรฐานการให้คะแนนของอาจารย์แตกต่างกัน เกิดความผิดพลาดในการให้คะแนนจากอาจารย์ (Rater error)
- Leniency/Severity
  - difference in the levels of severity between raters
- Rater inconsistency
  - instability of the level of severity within each rater
- Halo
  - tendency to let the rating of one trait influence ratings on other traits
- Restriction of range
  - clustering of ratings around a particular point on the rating scale

How to Reduce Rater Errors

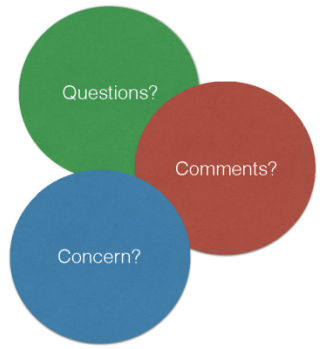
- Improving raters
  - Rater training
  - Rater monitoring
  - Rater feedback
- Improving the rating instrument

Summary

- Describe the characteristics of OSCE
- Discuss the advantages and disadvantages of OSCE
- Develop and conduct an OSCE

Steps of developing an OSCE

1. Define objectives
2. Choose the type of an OSCE item (process/product)
3. Write an item
4. Prepare instruments
5. SP recruitment and training
6. Item review committee
7. Item pool management



Line ID: Kasanarak

Medical Teacher, Vol. 12, No. 3/4, 1990

Twelve tips for organizing an Objective Structured Clinical Examination (OSCE)

R. M. HARDEN, *Centre for Medical Education, The University, Ninewells Hospital and Medical School, Dundee, Scotland*

The Objective Structured Clinical Examination or OSCE as it is more usually referred to, has been widely adopted as a tool to assess students' or doctors' competences in a range of subjects. The practical hints described in this article are aimed particularly at those who are undertaking for the first time the task of organizing the examination but they may be of interest also to a wider readership.

When used correctly, the OSCE can be highly successful as an instrument to assess competence in medicine and the approach has many advantages over more traditional methods. Careful organization and planning is necessary, however, if the potential of the technique is to be realised.

What is to be Assessed

Consider first what it is that you wish to assess. This should be related to the objectives of the course and may cover, for example, clinical methods, mastery of practical procedures, problem solving and clinical reasoning, and laboratory data interpretation. Is the emphasis on one, several, or all of these or indeed on other competencies not listed? Is the course concerned with the performance of the health professional in the hospital setting or with performance in a setting in the community? Is performance of the professional to be examined in isolation or as a member of a health care team? These are some of the issues that have to be addressed. It is important before planning proceeds with the OSCE to allow sufficient time to answer the question "What should be assessed?"

Tip 1 Produce a grid summarizing what is to be tested in your OSCE. Down the left hand side list the competencies such as history taking, physical examination, patient education etc. Along the top note the areas which should be represented

in the assessment. In an examination for medical students this might be classified in terms of body systems or specialist areas such as cardiology, respiratory medicine, endocrinology etc. or it may be broader subject areas such as orthopaedics, geriatrics, paediatrics, clinical pharmacology etc. Note in each of the squares on the grid the numbers of the stations where competence is assessed in that area. For example, a '1' in the square corresponding to history taking in endocrinology would reflect that at station one history taking ability is tested in a patient with hyperthyroidism. The number of entries in each of the horizontal rows will indicate the number of stations at which that competence, e.g. history taking, is assessed, while the number of entries in the vertical columns will indicate the number of stations at which the subject area, e.g. endocrinology, is assessed.

Duration of Station

A decision has to be made about the standard length of time to be allocated for each station. This will depend to some extent on the competencies to be assessed in the examination. Times ranging from 4 to 15 minutes have been reported in different examinations and a five minute station probably most frequently chosen. The use of linked stations extends the time available to complete a task. At a linked station students may be asked, for example, to assimilate information about a patient at one station and are then expected to act on this at the following station or they may take a history or examine a patient at one station and this is followed by a station at which they are asked to interpret their findings or to take further action with regard to the findings.

If necessary, you can include in the examination some stations which are allocated double the standard time. Such double stations will require to be duplicated in the examination.

Tip 2 Once the duration of stations has been fixed, make sure that the task expected of the student can be accomplished within the time. If in doubt reduce what is expected of the student. The exam should not be a race against time for the student unless this is one of the objectives being assessed. Where necessary use double or linked stations.

Number of Stations

The number of stations in an examination together with the time allocated for each station determines the time required to complete the whole examination. Twenty stations each of five minutes can be completed in 1 hour 40 mins, while 20 stations each of 10 minutes require 3 hrs 20 mins to complete. While some teachers organise examinations of this length others, concerned that the longer examination is perhaps as much a test of the candidates' endurance as it is of their competence in the areas assessed, would prefer to expose the students to two examinations each of 100 minutes.

One of the most frequently used formats for an OSCE is 20 stations, each of five minutes duration. This allows each student to be assessed on a range of competencies and for two groups of students to be passed through one circuit of the examination in the course of a morning or afternoon.

Tip 3 Within the time constraints include in the examination as many stations as possible as there is good evidence that the reliability of the examination is, in a large measure, dependent on the number of independent assessments of competence made during the examination.

Use of Examiners

In planning the examination it is important to consider the number of examiners available. Given that checklists for the examiners can be prepared and that the examiners can be rehearsed before the examination, it is possible to use junior as well as senior examiners in the OSCE. It is also possible—and many would say desirable—to use examiners from a range of specialties and disciplines. For example health education officers take part at a station on patient education, and dietitians at a station on nutrition. Use examiners for what they are best suited, that is observing students undertaking a task. Observation of how students process information and come to a conclusion may also be a legitimate use of an examiner where one is interested in the students' thought processes and the way in which they handle information.

Tip 4 Make sure that examiners are fully briefed prior to the examination, both about the procedure for the OSCE in general, and in particular with regard to the station at which they are examining. They should have an opportunity to comment on the brief for the student and on any checklists or rating marking scales to be used. Provide them with a list of the resources that will be available at their stations.

Range of Approaches

The OSCE offers a flexible approach to the assessment of clinical competence. Examinations vary in the number of stations, the duration of stations and the format of the stations. In the assessment of history taking, for example, a number of different approaches have been adopted, using the OSCE format. Workers in the field have varied in their design of the station, the brief for the students and the use of checklists and rating scales. Some exams are arranged in the clinical setting of the wards, others in outpatient clinics or offices and others in halls or seminar room accommodation. The range of possibilities is limited only by the imagination of the examiner.

Tip 5 When planning an OSCE for the first time, you will find it helpful to talk with a number of individuals from different settings who have previously made use of the technique. If possible visit and watch their examinations. Failing this read some of the articles that have been written about the subject so that you can get a better idea of the range of options available. Having done this, you should then make up an examination to suit your own needs.

New Stations

Pay particular attention when developing a station that the instructions are clear and unambiguous both for the student and the examiner, that the task can be undertaken in

the time available and that what is being tested is relevant to the objectives of the course.

Tip 6 Ideally, you should test a new station with one or more students before it is used in an examination. If this is not possible show it to colleagues who were not involved with the development of the station and ask them to review it. If there is an external examiner let him or her see the station. Following this review process, make any changes necessary; small changes in wording can make a big difference in clarity.

Organization of the Examination

If the OSCE is well organized in advance there should be no problems on the day of the examination that cannot be dealt with satisfactorily. Advance organization includes appropriate briefing of examiners and students to gain their co-operation, the assembly of resources and patients required for the OSCE along with reserves and completion of appropriate arrangements with regard to the site where the examination is to be held.

Tip 7 There should be a co-ordinator appointed well in advance of the examination who has responsibility for taking overall charge of the advance planning of the examination and for its implementation on the day. Where the exam is being run simultaneously at multiple sites, in the same building or in different building, there should be an additional co-ordinator in charge at each site. The co-ordinators have a key role to play in ensuring the smooth running of the examination.

Assigning Priority

Those involved with an OSCE must give it a high priority. All the resources for the examination including examiners and patients must be assembled at the site of the examination at the correct time. Examiners who are clinicians must ensure that alternative arrangements are made to cover their clinical commitments. The smooth running of the OSCE depends on everyone playing their part; it cannot proceed until all examiners and patients are present and all the stations set up.

Tip 8 Those in a position of authority should make it clear to their staff that the OSCE has a high priority. If they are not to take charge of the examination themselves they should invest the necessary authority in the co-ordinators to whom they delegate responsibility for the exam.

Resource Requirements

The resources required for an OSCE will depend on the nature of the examination and on the design of the stations. All the resources necessary for the examination should be identified prior to the examination and the ncessary steps taken to procure them.

Tip 9 It is helpful to produce a checklist of the resources required for each station in the examination. List what is required in terms of
i) examiners who are observers at the stations and examiners required to mark any written answers;
ii) patients, real and simulated;
iii) equipment, for example a sphygnomanometer, and furniture, for example, chairs, tables or beds;
iv) paperwork including checklists and instructions for the examiners, instructions or other sheets for students and a brief for simulated patients where required.

Plan of the Examination and Directions

Once the location of the examination has been decided and the stations fitted into the location, it is useful to prepare a plan of the layout on paper. On the plan note the position of each of the stations and indicate the path of the students from one station to the next. Give a copy of the plan to all examiners with their station marked, a copy to those setting up the examination and a copy to the students as part of the briefing on the morning of the examination, so that they get an overview of the circuit.

Ensure that the position of each station is marked clearly at the site of the examination and provide direction arrows to guide the student as he exits from one station and proceeds to the next one. Students should not be stressed during an examination by any difficulty in finding their way round the circuit. A student who moves inadvertently to the wrong station in the examination may cause chaos with a major disruption not only for himself but for the other students.

Tip 10 After the examination is set up and the direction signs in place, ask someone who has not been involved with setting up the examination to walk round the circuit to ascertain that they can find their way easily from one station to the next. In some situations, it may help to use tape on the floor as a direction guide for the student.

Change Signal

A range of methods are available to time the stations and provide the students and the examiners with an audible signal at five minute intervals or whatever time has been fixed for the station duration. This can range from a simple approach with a member of staff using a stop watch and a mechanical or electric bell, to devices available specifically for this purpose, or a computer appropriately programmed.

Tip 11 Before the examination check that the audible signal can be heard clearly at all locations on the examination, if necessary with doors to rooms closed and screens drawn. Allow for background noise on the day of the examination, e.g. the patient may be talking.

Records

It is likely that the OSCE will become, if it is not already, part of your toolkit as an

examiner. It always takes longer to organize the examination on the first occasion than on subsequent occasions, providing information is kept about the OSCE and the materials used in the OSCE retained.

Tip 12 Keep on OSCE file into which you put all the resources required for the OSCE such as station numbers, direction arrows, master sheets of instructions to students, checklists, rating scales, examples of correspondence notifying ward staff, patient, examiners etc., and a bank of questions used previously.

2013; 35: e1447–e1463 MEDICAL TEACHER

WEB PAPER
AMEE GUIDE

The Objective Structured Clinical Examination (OSCE): AMEE Guide No. 81. Part II: Organisation & Administration

KAMRAN Z. KHAN[1], KATHRYN GAUNT[2], SANKARANARAYANAN RAMACHANDRAN[2] & PIYUSH PUSHKAR[3]

[1]Manchester Medical School, UK, [2]Lancashire Teaching Hospitals NHS Trust, UK, [3]Aintree University Hospitals Trust, UK

Abstract

The organisation, administration and running of a successful OSCE programme need considerable knowledge, experience and planning. Different teams looking after various aspects of OSCE need to work collaboratively for an effective question bank development, examiner training and standardised patients' training. Quality assurance is an ongoing process taking place throughout the OSCE cycle. In order for the OSCE to generate reliable results it is essential to pay attention to each and every element of quality assurance, as poorly standardised patients, untrained examiners, poor quality questions and inappropriate scoring rubrics each will affect the reliability of the OSCE. The validity will also be influenced if the questions are not realistic and mapped against the learning outcomes of the teaching programme. This part of the Guide addresses all these important issues in order to help the reader setup and quality assure their new or existing OSCE programmes.

Introduction

This Guide is the second in the series of two Guides on the OSCE. The first Guide focuses on the historical background and educational principles of the OSCE; knowledge and understanding of these educational principles is essential before embarking upon designing and administering an OSCE. We would advise the reader to familiarise themselves with the contents of Part I prior to reading this part of the Guide.

In this second part we aim to assist the reader in applying the theoretical knowledge gained through Part 1 by outlining the practical steps required to design and run a successful OSCE, from preparation and planning through to implementation and post-OSCE considerations.

We have chosen to present Part II of this Guide as the evolving story of a fictional character Eva, an enthusiastic educationalist who is new to the concept of the OSCE. She is asked by the Dean of her institution to introduce the OSCE as a new form of assessment for the health care students graduating the following year. The knowledge and experiences she gains through this process are outlined in this Guide to assist others in implementing an OSCE for the first time or quality assuring their existing assessment processes.

Practice points

- The assessment team would need to adopt new roles and responsibilities when setting up a new OSCE programme.
- A nominated OSCE lead needs to have an oversight of all aspects of the OSCE programme.
- An OSCE question bank needs to be developed and maintained in order to have a pool of quality assured and peer reviewed stations for use in various examination sittings.
- Examiner and Standardised Patient training are important elements of quality assurance and standardisation process.
- Post-hoc psychometrics provide valuable data for further quality assuring the OSCE questions and the programme as a whole.

Preparation and planning

Organisational structure

Large numbers of personnel are required in order to successfully implement an OSCE programme (Cusimano et al. 1994). Within higher education institutions there is usually a team responsible for overseeing assessment procedures. Changes to an assessment programme, such as the implementation of new methods of assessment should be undertaken with the help of the assessment team. It may be worthwhile for a small sub-committee to be formed from the members of the Assessment Team to lead on the introduction of the OSCE in the existing assessment programme. Following a successful implementation, ongoing review and quality assurance

Correspondence: Dr Kamran Khan, Department of Anaesthetic, Mafraq Hospital, Mafraq, Abu Dhabi, UAE; email: Kamran.Khan950@gmail.com

ISSN 0142–159X print/ISSN 1466–187X online/13/091447–17 © 2013 Informa UK Ltd.
DOI: 10.3109/0142159X.2013.818635

procedures can be continued by the Assessment Team, as for all other methods of assessment.

Within this sub-committee it may be beneficial to assign a single key person (the OSCE lead) with the overall responsibility and accountability for overseeing the development, organisation and administration of the examination (McCoy & Merrick 2001). This person should have expert knowledge or prior experience in conducting the OSCE. If this is not the case the chosen lead should gather information by reviewing the literature, attending workshops and seeking guidance from experts at other centres.

Eva's Story

One sunny morning as I was walking towards my office for work I met the Dean, who had just returned from the AMEE conference the previous week. After a quick greeting he invited me into his office. I wondered what was on his mind.

Over freshly brewed coffee he told me that he had been learning a lot about the OSCE at the conference. He thought it would be a good idea to introduce the OSCE to assess our students who would be graduating the following year and asked me to lead on this. I am always up for a challenge but this was a different one.

I did not know much about the OSCE, except that it is an assessment tool. I accepted the challenge but openly admitted that I would need to do a lot of homework, and would report back to him in due course. He was delighted that I had agreed to help and I looked forward to expanding my repertoire of assessment techniques.

I returned to my office and began to do some research on the topic; when, why and how was the OSCE developed? After this background reading, I was starting to grasp the theoretical principles of the OSCE but what I needed now was some practical advice on how to establish our own examination. Who could I ask?

I remembered my colleague and friend George, a Paediatrician at a neighbouring University. They had been using the OSCE for the assessment of medical students for some time.

The following week I had a chance to visit George to find out more. He showed me around their facilities and explained that I would need quite a bit of help with the workload. He advised that I initially consider forming a small organisational group, ideally including colleagues already involved with our assessment procedures. I decided to approach our pre-existing Assessment Team for support.

Administrative support. Assessment of any kind inevitably creates a vast amount of administrative work. The OSCE is no exception to this and by ensuring there is adequate administrative support to meet these needs, the OSCE lead will have more time to address the academic considerations. Tasks such as the allocation of students to examination centres, distribution of examination paperwork and the handling of examination results should ideally be dealt by a dedicated administrative team.

Developing the larger team. Depending on the nature and format of the examination and the size of the institution there might be more than one site running the OSCE on the same day. At each site it may be helpful to develop a local organising team to oversee the practical aspects of the OSCE such as selecting local examiners and standardised patients, setting up the OSCE circuit and ensuring smooth running of the OSCE on the examination day. In smaller institutions members of the Assessment Team or their administrative support may perform such tasks.

Examination scheduling, rules and regulations

Setting the examination schedule. In any given academic year, there may be a need to schedule a number of OSCE sittings depending on the course curriculum requirements and the place of the OSCE within the broader assessment programme. It is common to run at least one OSCE for each year group of students per year. The exact timing of each examination should be primarily influenced by the institutional regulations and curriculum requirements, although venue and examiner availability should also be considered.

Setting an examination blueprint and examination length

Blueprinting and mapping. Blueprinting is the process of formally determining the content of any examination. In the case of an OSCE this involves choosing the spread of skills and the frequency with which each appears within an examination. Each blueprint for an OSCE should take into account the context of the examination, the content which needs to be assessed mapped to the curriculum and the need for triangulation, e.g. if any domains of assessment should be examined with the use of more than one assessment tool (Vargas et al. 2007). Part 1 of this Guide discusses the need to carefully match assessment methods to the skills, knowledge and attitudes being assessed, in more detail. In this way the OSCE should form only one part of the broader assessment programme.

The OSCE is primarily a competency-based assessment of performance in simulated environment (Khan & Ramachandran 2012), and therefore, principally assesses the skills-based learning outcomes. However, a detailed discussion of the learning domains which can be assessed by the OSCE is covered in Part 1 of the Guide. The blueprinting process should ensure that an appropriate sample of the skills-based curriculum is examined and it is mapped to the curriculum, i.e. the examination has adequate content validity. A blueprint normally consists of a two-dimensional matrix with one axis representing the generic competencies to be tested (e.g. history taking, communication skills, physical

Table 1. An example of an OSCE blueprint.

Topics	Procedural skills	Clinical examination skills	History taking	Total number of questions
Acute medicine	Q.1			1
ENT		Q.2 Q.3	Q.4 Q.5	4
Paediatrics	Q.6 Q.7			2
Geriatrics		Q.8 Q.9	Q.10	3
Total number of questions	3	4	3	10

examination, management planning, etc.) and the other axis representing the problems or conditions upon which the competencies will be demonstrated (Newble 2004). An example of a test blue print is shown in Table 1. Blueprinting can be done 'in-house' by the Assessment Team; however, in higher stakes examinations, a Delphi or other survey techniques may be used to agree on the topics to be included in the test blueprint. Questions can then be developed or chosen based upon the test blue-print.

Examination length (number of stations). In order to develop an examination blueprint, the examination length needs to be determined beforehand. This will depend on the number of stations within each OSCE and the length of each station. An OSCE station typically refers to one time-limited task given to the candidates generally lasting between 5 and 10 min. The reliability (reproducibility of the test results) and validity (the extent to which the test's content is representative of the actual skills learned or to be assessed) are both influenced by the number of stations and total length of the examination (Newble 2004). An appropriate and realistic time allocation for tasks at individual stations will improve the test validity. Whereas, increasing the breadth of the content, usually by ensuring an adequate number of stations per examination, improves reliability. In fact, the content specificity has been found to be a major contributor to poor reliability; hence competence testing across a large sample of cases is required before a reliable generalisation of candidates' performance can be made (Roberts et al. 2006).

The number of stations needed to generate a reliable score represented by either Cronbach's α or Generalisability (G) coefficient determines the examination length. A Cronbach's α or G value between 0.7 and 0.8 reflects an acceptable reliability for high stakes examinations. A detailed discussion on this topic is beyond the scope of this article and interested readers are advised to refer to AMEE guide 49 by Pell and colleagues (Pell et al. 2010) and AMEE guides 54 and 66 both by Tavakol & Dennick (2011, 2012). Work by Shavelson and Webb (2009) may also be of practical use. These concepts are also re-visited throughout the Guide.

For practical purposes decisions around test length generally need to balance reliability coefficients with feasibility and resource issues; but as a general recommendation, with well-constructed OSCE stations, an adequate reliability could be achieved with 14–18 stations each with 5–10 min duration (Epstein 2007).

Developing a bank of OSCE Stations

Before the stations are added to the bank they need to go through the processes of peer review and piloting. If available, psychometric data on individual stations might also provide useful information on station quality including its ability to discriminate between high-achieving and low-achieving candidates. These aspects are described in some detail below. A secure bank of robust and quality assured stations contributes significantly towards the better reliability and validity of the examination scores. The flow diagram presented here describes one approach to producing a bank of OSCE stations to meet the needs of the curriculum (Figure 1). It can be used as a step-by-step guide or adapted for individual requirements. A pre-existing bank of stations could also be updated or quality assured by following appropriate steps in the algorithm.

Choice of topics for new stations. In institutions where an OSCE is being set up for the very first time, the examination blueprint governed by the curriculum outcomes would act as a good starting point to identify the topics for writing new OSCE stations. In places where an OSCE station bank already exists, the OSCE lead or subject experts could review this to identify gaps in the assessment of certain skills or domains. The need for new stations could also arise if the curriculum is modified or learning objectives of the modules are changed. The assessment of competencies should always be aligned to the teaching and learning that has taken place as specified by the course curriculum. Occasionally assessment content is influenced by the regulatory authorities as in the case of General Medical Council (GMC) in the UK, which stipulates that medical graduates should be able to demonstrate certain competencies (GMC 2009).

Once the areas for assessment have been identified it is important to ensure that the clinical skills which are expected to be performed by the candidates can be realistically assessed using an OSCE format and in the limited time allocated for each station (Vargas et al. 2007). Part 1 of this Guide describes in some detail what the OSCE can assess most appropriately.

Choice of station writers. The OSCE lead has the responsibility of identifying appropriate people to design and write the OSCE stations. If a pool of trained examiners already exists, it would be an obvious choice to seek volunteers for question writing from this pool. Otherwise subject experts can be asked

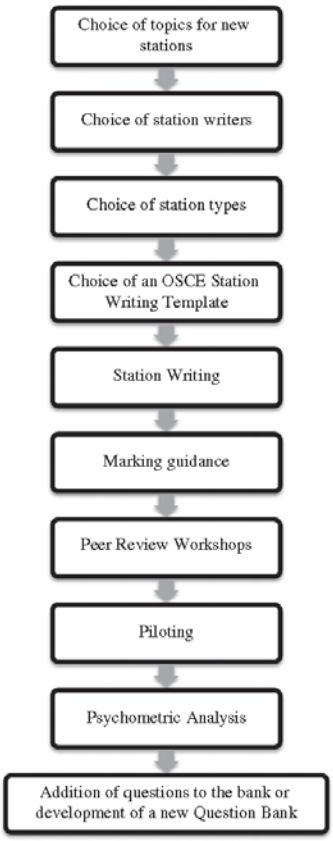


Figure 1. Flow diagram for the development of an OSCE question bank.

to help with writing. It is essential for station writers to be familiar with the underlying principles of the OSCE in order for them to produce appropriate work. Brief orientation sessions or written instructions could be developed for people new to this task.

Choice of station types. The OSCE lead or the person coordinating the station writing should advise the question writers about the type of new stations needed. An understanding of the different types of OSCE station formats is essential in the choice of appropriate station types for various assessment outcomes (Table 2).

The choice of OSCE station writing template. Once the type of station has been chosen, an appropriate template for station writing should be developed or used. A template helps authors to develop stations in a similar format to others within the bank. Such a standardisation prevents disadvantaging the candidates by posing questions in unfamiliar formats and helps to maintain the reliability of the scores. We have shown an example of an observed OSCE station writing template. This template is supplemented with an example of a station designed to assess the focussed history taking and respiratory examination of a patient with asthma (Appendix 1).

Station writing. The different sections of the template highlight the information that should be considered in order to write a successful OSCE station. Each of these sections is shown below with an explanation of the type of information required (Table 3).

Marking guidance. The marking guidance for each station would depend on the scoring rubric chosen as a standard for all OSCE examinations. There is a detailed discussion on the different types of scoring rubrics later in this Guide. If it is a checklist or a rating scale the author should develop it as they write the station. Such a checklist should reflect the outcomes being assessed. The same is applicable for rating scales if these are specific to the stations. If a global rating scale is to be used in isolation then the marking criteria for individual stations may not be required. AMEE guides 49, 54 and 66 discuss further the impact different scoring rubrics can have on OSCE outcomes (Pell et al. 2010; Tavakol & Dennick 2011; Tavakol & Dennick; 2012).

Peer review workshops. Running review workshops with examiners is one way of quality-assuring new OSCE stations. Once the examiners have written the new stations, they are invited to bring these to the workshops where delegates can review stations written by others, often in small groups. The presence of the authors for individual stations at the workshops ensures changes and clarifications are made more easily.

In addition to looking at the clinical accuracy and appropriateness of the tasks involved in the station the peer review process can help to identify validity issues as well. A simple questionnaire could be used for this purpose, an example of such a questionnaire is shown in Appendix 2.

Piloting. After the peer review process by the examiners, piloting of the stations helps to identify any issues with the practicality and allocation of time for the tasks. If required, changes can then be made to the stations to improve their quality (Whelan 1999). Initial psychometric analysis on reliability and station quality could also be done at this stage. In the case of any problems with a particular station it should be redesigned and then re-piloted. Piloting often takes place during mock or low-stakes examinations which may have the additional benefits of orientating candidates to the OSCE and providing them with immediate feedback on their performance. If individual stations are piloted within the circuit of a high stakes examination it is essential to inform the candidates about the inclusion of a pilot station and that its scores will not influence the overall examination results. In order to get valid and reliable data on the pilot stations included in real examinations, the identity of such stations is not disclosed beforehand.

Psychometric analysis. We have briefly discussed relevant aspects of psychometrics in the section on Examination Length earlier. With respect to development of new stations,

Table 2. Types of OSCE stations.

Station type	Description	Examples	Benefits	Limitations
Observed station	An examiner is present throughout the duration of testing	Communication skills Procedural skills Clinical examination	Direct observation allows assessment of the higher levels of the learning domains. Immediate feedback can be given for formative assessment.	Resource intensive with one examiner needed per station
Unobserved station	No examiner present throughout the testing period. Answers may be submitted on paper either after each station or following the completion of the examination	Interpretation of clinical information e.g. X-rays, pathology specimens, blood results. Prescribing skills. Information Technology Skills.	Examiners not required on such stations.	No direct observation of performance. OSCE station may not be necessary and alternative assessment tools may be as effective at assessing the cognitive skills in question.
Technology enhanced station	A station involving the use of technological advances such as part-task trainers or high-fidelity manikins to assess skills that would otherwise be difficult to assess in the OSCE format.	Intimate clinical examinations with use of part-task trainers, e.g. rectal examination. Complex decision making skills and the management of acutely unwell patients with the use of high-fidelity manikins.	Increases the scope of potential OSCE stations, allowing assessment of learning domains which could not be assessed effectively using traditional OSCE stations.	Personnel must be trained in the use of the equipment. Equipment failure. Initial cost of new equipment and maintenance costs.
Linked stations	Two consecutive stations are based upon the same clinical scenario or information. These may be observed or unobserved.	Observed examination of the respiratory system in the first of two stations. Unobserved documentation of findings and management plan in the second station.	Greater number of skills can be assessed per scenario Efficient use of examiner resources.	Needs careful circuit planning, as candidates can neither start on the second of a pair of linked stations nor end the examination on the first.

if a complete set of new questions is used in a mock OSCE then the psychometric analysis will indicate the overall reliability of the set of questions. Use of G theory will indicate the number of similar stations needed to achieve a good reliability by performing D or decision studies on the data (Shavelson & Webb 2009). Application of Item Response Theory (Downing 2003) will also be able to yield data highlighting the sources of variability or error. This theory can be used if one or more stations are piloted in real examinations. A detailed discussion on this topic is again beyond the scope of this Guide, but it is discussed in some detail in AMEE Guide No. 49 (Pell et al. 2010), and G theory is comprehensively covered in AMEE Guide No. 68 (Bloch & Norman 2012).

Choosing a scoring rubric and standard setting

Stevens and Levi (2005) have defined a scoring rubric as 'an assessment tool that delineates the expectations for a task or an assignment'. Various scoring rubrics are used to mark different types of assessment. There are two main types of scoring rubrics, analytical and holistic.

Analytical scoring (checklist scale). A checklist is a list of statements describing the actions expected of the candidates at the station. It is prepared in advance, following consultation with the team designing the OSCE stations and in line with the content and outcomes being assessed. Checklists could be 'binary', yes/no (performed/not performed), i.e. candidates are marked based on whether or not an action was performed, without any discrimination for the quality of the actions. Such checklists are may not be able to discriminate between lower and higher levels of performance. Alternatively, checklists can have 5–7 point rating scale, which allows the examiners to mark candidates based upon the quality of the actions. Such checklists with rating scales are different from global ratings (holistic scoring), which are described later.

Traditionally, a key strength of binary checklists has been their perceived ability to provide an objective assessment and thought to lead to greater *inter-rater reliability.* In fact, such checklists were originally used by Harden when he first developed OSCE techniques as shown in the first part of this Guide (Harden et al. 1975). There is, however a growing body of evidence which has called this view into question, showing that objectivity does not necessarily translate into greater reliability (Wilkinson et al. 2003). This is particularly applicable if expert examiners are used in an OSCE (Hodges & McIlroy 2003). An example of binary checklists and rating scales (which can be seen as mini global ratings as they rate one element of the overall consultation), is shown in Table 4.

Holistic scoring (global rating scale). Compared with checklists, which are task specific (Reznick et al. 1998), global rating scales allow the assessor to rate the whole process. Consider the performance of an expert who may not follow a pre-determined sequence of steps as outlined by a checklist, yet still performs the task to a high standard with fluidity and ease. In this situation an overall (global) assessment of the

Table 3. Guidance for completing the question writing template.

Template section	Information required
Station information	
Subject/Topic	Chosen as described at the beginning of this section.
Level of candidate	The competencies that need to be demonstrated should be appropriate for the level of training of the candidates.
Competencies to be assessed	As described in part one of this Guide.
Station duration	This should generally be standard for all stations as determined by the OSCE committee as described in the examination length section.
Information for the site organisers	
Standardised patient (SP) age and sex	This information is key for the organising team to find appropriate SPs for the stations.
Resources and equipment needed	The organising teams need this information to equip the station and to maintain standardisation across sites if an examination is taking place at multiple sites.
Setting up the station	This information is needed to ensure that the position of the chairs, tables and couches do not interfere with the tasks in questions. For example candidates should be able to approach the SP for examination from the right side and the examiners are able to view the whole process from where they are seated.
Instructions for candidates	
What is the scenario (briefly)?	This is the key information for the candidates about the station.
Who and where they (Candidates) are?	It is essential for the candidates to know in what capacity they are having a consultation with the patient, so that they can provide appropriate information to the patients and the examiners.
What has already happened	If the scenario is setup in such a way that a nurse has seen the patient and some information about the blood pressure and temperature etc. could be available to the candidates if they asked for it, then it should be made clear at this stage.
What are the candidates expected to do	These are the skills to be demonstrated by the candidates at a given station, e.g. take a focused history and examine the respiratory system
What are the candidates not expected to do?	If due to limitation of time candidates are not expected to perform General Physical Examination as a part of the respiratory system examination then it should be clarified here.
Supplemental data	If any investigations could be provided to the candidates then it should be included in this section. For instance, 'if you require the results of any relevant blood tests you can ask the examiners
Template section	Information required
Information for the examiner	
Brief background to the scenario	This section is similar to 'What is the scenario?' section above.
Examiner's role	It is very important to make clear here, in what way the examiners are allowed to interact with the candidates if at all. Are they supposed to observe only or ask some questions as well?
What are the objectives of the station	Again similar to 'What are the candidates expected to do?' section above.
What information they might be able to provide the candidate?	The examiners, if required, can provide information such as reports of blood tests or X-rays.
What information they should not provide the candidate?	The examiner should not volunteer any information candidates are expected to explore for themselves.
Clinical information relevant to the station	If examiners from disciplines other than the one in question are used for assessment then this information should be able to enable them to mark the candidates fairly. This becomes even more important, especially if generic marking schemes are to be used for assessment as described later in this Guide.
Simulated patient information (All the information in this section is essential in maintaining the standardisation of the station, especially if the examination is being conducted at split sites)	
Who are they?	As set out in the candidates' instructions earlier.
Their social/economic background (if applicable)	This needs to be brief, considering the station length and time allowed for the candidates to complete this task.
Presenting complaint and past history	Should be succinct and focused for the purposes of the station.
Details of current health problems and medications	Again this information should be relevant and to the point.
Details of their concerns/perceptions	If a station involves dealing with SP concerns then it is important to standardise this aspect.
What they should say (their agenda) and what they should not say	In certain stations this information would not be needed, for instance where the candidates are expected to examine only.
What they should ask (questions)	The SPs should be allowed to ask standard questions when talking to the candidates.
Specific standardisation issues (specific answers to specific questions, please stay with the script)	This information could be vital if the candidates are expected to ask some particular questions.

performance is required in order to accurately reflect the skill of the candidate. Global scales allow examiners to determine not only whether an action was performed, but also how well it was performed. This tool is therefore better for assessing skills where the quality with which it is performed is as important as performing it at all. An example might be the assessment of a candidate's ability to empathise with patients in communication skills stations. Hence holistic scales are more useful for assessing areas such as judgement, empathy, organisation of knowledge and technical skills (Morgan et al. 2001; Hodges & McIlroy 2003). Global ratings differ from checklist rating scales described above by the virtue that global ratings take a more holistic view of the overall performance at a station compared to a rating scales looking at one aspect alone.

Global ratings are being increasingly used over checklists for marking at OSCE stations, as there is now evidence to suggest that they show greater inter-station reliability, better

Table 4. Comparison of binary checklist and rating scale.

Station on history and examination of an asthmatic patient: two possible types of checklist for scoring the examination of the chest

Binary check list	Checklist using rating scales
Candidate performs an examination of the chest	Candidate performs an examination of the chest
1. Introduction	1 [] 2 [] 3 [] 4 [] 5 []
Yes [] No []	1. Unstructured approach
2. Obtaining consent	2. Structured approach but completes less than 50% of key steps.
Yes [] No []	3. Structured approach but completes more than 50% of key steps.
3. Appropriate exposure	4. Structured approach and completes a majority of key steps.
Yes [] No []	5. Structured approach and completes all key steps.
4. Professional approach	
Yes [] No []	Such a check list should then be accompanied by a list of key steps.
5. General physical examination	Uses alcohol rub before and after examination and, when appropriate uses gloves
Yes [] No []	Seeks permission to examine, and explains the nature of examination
6. Inspection	Offers/Asks for chaperone where appropriate
Yes [] No []	Asks the patient if any areas to be palpated or moved are painful
7. Palpation	Positions the patient correctly and comfortably, then uses a methodical, fluent and correct technique
Yes [] No []	
8. Percussion	Does not distress, embarrass or hurt the patient unduly
Yes [] No []	Examines, or suggests examining, all the relevant areas
9. Auscultation	Completes the task, covers up the exposed areas and thanks the patient
Yes [] No []	
10. Advising patient at the end to cover up the exposed areas and thank for cooperation.	
Yes [] No []	
Score awarded X/10	
Add the scores and award out of 10	

construct validity, and better concurrent validity compared to checklists (Turner & Dankoski 2008). Further information on the impact of scoring rubrics can be found in AMEE Guides 49, 54 and 66 (Pell et al. 2010; Tavakol & Dennick 2011; Tavakol & Dennick 2012).

Standard setting. Standard setting refers to defining the score at which a candidate will pass or fail. There are a number of methods that can be used for this purpose. In the Norm referenced methods the scores have meaning to each other and the pass/fail scores are determined by the relative scores of candidates, e.g. the Cohen method (Cohen-Schotanus & van der Vleuten 2010). In a norm referenced examination the standard set is based upon peer performance and can vary from cohort to cohort. It is, therefore, possible that in a 'poor' cohort, a candidate may pass an examination that they would have otherwise failed if they took the examination with a 'stronger' cohort. For this reason norm referencing is usually deemed unacceptable for clinical competency licensing tests, which aim to ensure that candidates are safe to practice. In this case a clear standard needs to be defined, below which a doctor would not be judged fit to practice. Such standards are set by Criterion referencing in which scores have absolute meanings to the domains of assessment. Angoff (1971) and Ebel (1972) are two commonly used methods for this purpose. The criterion methods of standard setting are performed before the examination by a group of experts who look at each test item to determine its difficulty and relevance. Although both Angoff and Ebel methods are well established, these were initially developed for tests of knowledge such as multiple-choice examinations and it may not always be appropriate to extrapolate these methods to tests of performance, i.e. the OSCE (PMETB 2007). Other absolute methods which may be of relevance include Borderline Group and Contrasting Groups Methods; readers can find more about these in the articles written by Kaufman (2000) and Kramer (2003). A detailed discussion of each of these methods and their pros and cons is beyond the scope of this Guide and interested readers are also referred to AMEE Guide No. 18 on Standard Setting in Assessment by Friedman (Friedman Ben-David 2000) and articles by Tavakol (2012) and Pell (2010).

Developing a pool of trained examiners

In maintaining the reliability of the scores in an OSCE examination, consistent marking by trained examiners plays a pivotal role. Examiner training is an ongoing process whereby new examiners are added to the pool and the existing examiners are provided with refresher training. This section deals with the process of examiner training and retention.

Identification of potential examiners. The reliability of the scores generated by the examiners not only depends upon the consistent marking by the examiners but also their clinical experience relevant to the OSCE station. It is common for doctors to assess doctors and nurses to assess nurses, however, skill-matching can add a degree of flexibility and has resource and financial implications. In order to lessen the burden of finding adequate numbers of doctors to act as examiners; there have been instances where non-physician examiners have been used in medical examinations. There is literature suggesting that simulated patient scores have a good correlation with physician scores (Mann et al. 1990;

Box 1. Outcomes for examiner training.

Learning outcomes for examiner training sessions

To understand the scope and principles of the OSCE examination
To maintain consistent professional conduct within the examination
To understand and use the scoring rubric in order to maintain standardisation
To provide written feedback on performance if required in summative examinations
To provide verbal feedback at the end of station in the formative examinations
To ensure confidentiality of the candidates' marking sheets
To understand the procedures for inappropriate or dangerous behaviour by candidates

Cohen et al. 1990). However one study suggests the agreement between physician and non-physician scores, although good for check-list scoring, does not extend to global scoring (Humphrey-Murto et al. 2005a). Whether expert or non-expert examiners are chosen, training all examiners will reduce examiner variation (Newble et al. 1980; van der Vleuten et al. 1989) and improve consistency in behaviour, which may improve exam reliability (Tan & Azila 2007).

Most physician examiners are sourced from local hospitals or community practices and it is helpful to approach those with a prior interest in medical education. In most professions the examiners are not financially remunerated as this is seen as a part of their responsibility towards teaching and training; however those who do volunteer describe an enhanced sense of duty and an insight into learners' skills (Humphrey-Murto et al. 2005b).

Examiner training workshops. Examiner training sessions should ideally take place well in advance of the examinations. The level of training will depend upon the background and ability of the examiners (Newble et al. 1980; van der Vleuten et al. 1989). As with any other teaching and learning activity the outcomes of the examiner training workshops should be explicit (Box 1).

These sessions can be organised in any format but generally include group discussions about some of the above topics, followed by the opportunity for the examiners to mark Mock OSCE or videos of real OSCE.

Although examiners tend to maintain and further develop their skills by regularly assessing, the need for refresher training can be driven by a change in the format of examination or scoring and also by changes in the requirements of the institutions or regulatory bodies. Such refresher training could be delivered using online resources or by further small group sessions.

Developing a pool of trained standardised patients

Patients form an integral part of an OSCE with many of the stations requiring active patient participation. Collins & Harden (1998) refer to the continuum of patients used in clinical examinations, from the real patient with clinical signs who receives no training to the rigorously trained simulated patient. The recruitment and training of each type of patient will differ depending upon their role within the examination.

Although the terms 'simulated patients' and 'standardised patients' are used interchangeably, a simulated patient is a usually a lay person who is trained to portray a patient with a specific condition in a realistic, and so standardised way (Cleland et al. 2009). Standardised Patient (SP) is an umbrella term for both a simulated patient and an actual patient trained to present their condition in a standardised way (Barrows 1993). Standardisation in the term 'standardised patient' relates to the consistent content of verbal and behavioural responses by the patient to stimuli provided by a candidate (Adamo 2003).

Recruitment of standardised patients. The type of patient required for each OSCE station will depend upon the desired outcomes of the station and the role expected to be played by them. If the station requires the candidate to elicit a specific clinical sign, e.g. a heart murmur, a real patient with the murmur in question must be used. However, if the focus of the station is to determine if the candidate can competently examine the cardiovascular system (regardless of any clinical abnormality) a 'healthy' volunteer can be used instead. Certain stations, such as history taking and communication skills stations will generally require the use of trained simulated patients. AMEE Guides Nos. 13 and 42 provide a detailed discussion on choosing the correct 'patient type' for the examination in question (Collins & Harden 1998; Cleland et al. 2009).

Patients can be recruited in a number of ways; real patients with clinical signs can be accessed through contacts with primary and secondary care physicians. A doctor previously known to the patient and responsible for their care may be the most appropriate person to make initial contact (Collins & Harden, 1998). Recruiting patients with common conditions that remain stable over time is easier than finding patients with rare and unstable disease and this should be taken into account at the time of blueprinting and station development. Healthy volunteers can be found through advertising in the local press, contacts with local educational institutions and by the word of mouth. Actors are commonly used for complex communication issues such as breaking bad news and for high-stakes examinations (Cleland et al. 2009). Highly trained professional actors are likely to incur significantly higher costs than volunteers and real patients who may be remunerated by the reimbursement of expenses alone.

In many large institutions a standardised/simulated patient co-ordinator is employed to undertake the selection process keeping in mind the ability, suitability and credibility of the SPs. Each of these areas is discussed in detail in AMEE Guide 42 and is beyond the scope of this guide (Cleland et al. 2009).

Standardised patient training. All standardised patients will require training, but real patients and simulated patients (actors) will require different levels of input. All will need to understand the importance of portraying the clinical conditions in question, reliably and repeatedly and the need for standardisation between candidates. In some cases the pre-examination briefing on the day may be adequate for this purpose; generally simulated patients for role play in more

complex scenarios will require dedicated training in advance of the examination.

In addition to their use in the OSCE, simulated patients are often used for teaching skills to the medical students outside the examination settings. It may be convenient and cost effective to train groups of simulated patients together to be used for a variety of purposes within the institution. In depth discussions on simulated patient training workshops can be found in AMEE Guides Nos. 13 (Collins & Harden 1998) and 42 (Cleland et al. 2009) and are not reproduced here. In addition there are associations dedicated to educating standardised patients such as the Association of Standardised Patient Educators, who provide leadership, education and structure to the training and assessment of standardised patients (Turner & Dankoski 2008). Although there is no real consensus in the literature as to the sufficient duration of training for each simulated patient, one estimate suggests it may take up to 15 h to adequately train a simulated patient dependent on the role, experience and adaptability of the person (Shumway & Harden 2003).

Once training is completed each standardised patient's performance needs to be quality assured before being used in a high stakes examination. Simulated patients may be videotaped and their performance evaluated by an independent group of trainers (Williams 2004). Alternatively new simulated patients could be used for the first time in mock OSCEs and feedback from candidates and examiners could be used to quality assure their performance (Stillman 1993).

If there is a standardised patient co-ordinator, they should hold a bank of trained patients who can be called upon for subsequent examinations. Ideally individuals within this bank should be trained to perform multiple roles, this will increase the flexibility and maximise the potential to find the right person for the right scenario (Whelan 1999).

Standardised patients are a valuable resource, it is important to keep them interested in the role by using them regularly, remunerating appropriately and always expressing thanks for their input (Cleland et al. 2009).

Running the OSCE

Administrative tasks

As previously described, any form of examination generates considerable administrative work. We describe here the key administrative activities that may need consideration in order to ensure the smooth running of an OSCE (Box 2).

All relevant information pertaining to the implementation of the OSCE could be held within a procedure manual for future reference. This may include lists of trained examiners, trained SPs, sources of equipment and catering facilities.

Choosing an OSCE venue

The OSCE venue should be booked well in advance bearing in mind the number of stations and candidates. In addition to housing the examination itself, the venue should ideally have the capacity for briefing rooms, administrative offices, waiting rooms for patients and examiners, quarantine facilities and refreshment areas. Stations may be accommodated in several small rooms similar to outpatient clinics or alternatively a larger room can be turned into 'station areas' with the use of dividing screens. Individual rooms have the advantage of increased confidentiality and low noise levels but may make the signposting of the circuit more challenging. Some institutions have a special site allocated specifically for the examinations.

Eva's Story (continued)

I have learnt so much about OSCEs in the past year! I had no idea so much preparation was going to be required in introducing this new examination. As the OSCE lead I was overseeing all of the academic considerations that I have just described to you.

It has taken some time but we now have a bank of questions designed to assess the final year medical students, these are blue-printed and mapped against the curriculum and have been quality assured at a peer-review workshop.

We have spent the last few months identifying and training our new OSCE examiners. George was a real help here, as he came along to describe to them how the OSCE worked at his University. Most of the new examiners were supportive of this change to assessment although there were a few who were quite resistant. Having the background knowledge of the advantages and disadvantages of the OSCE was really helpful in the debate that ensued.

We have managed to find some volunteers to act as patients in the OSCE and have identified some real patients with good clinical signs for our clinical examination stations.

We have decided to run a pilot examination with a group of the current interns about a month before the real examination of the final year students. In this way we can check the practicalities of the stations and ensure the examiners and patients are comfortable with their roles. There are still so many things to think about to ensure the smooth running on the big day. I have made a list of all the essential requirements for running an OSCE and share it with you now.

Setting up the OSCE circuit and equipment

The OSCE circuit. The circuit is the term used to describe the setup of stations for the seamless flow of candidates through the examination. Each candidate will individually visit every station within the circuit throughout the course of the examination. The number of candidates in each sitting should, therefore, be equal to the number of stations, unless rest stations are used as described below. Each candidate will be allocated a start station and move from station to station in the direction of the circuit until all stations have been

Box 2. Common administrative tasks for OSCE.

Common administrative tasks for the OSCE

Allocation of students to examination centres
- If examinations are to be held at multiple sites, planning is required to ensure that wherever possible examiners do not know the candidates and any candidates with disabilities are sent to centres with appropriate facilities.

Transport and reporting instructions
- Candidates must be provided with comprehensive instructions about where to report at the examination centre. In some circumstances transport may need to be arranged for large groups of candidates.

Distribution of paperwork
- Station information, candidates' lists and mark sheets need to be printed, collated and distributed to all examination sites. Mark sheets should be pre-populated with candidates' details to minimise time required during the examination.

Selection of standardised patients
- Once equipped with the station information it is necessary to identify appropriate SPs from the trained pool for all stations. Commonly, more than one SP for each station is identified, as fatigue may occur if the station is to be run several times in the day. In addition it is also advisable to invite a number of reserves. They should receive their scripts and reporting instructions in advance.

Selection of examiners
- Once the station information is known appropriate examiners must be selected from the trained pool, taking into consideration the decisions made regarding expert versus non-expert examiners. Reserve examiners should always be invited.

completed. The local organising team will usually be responsible for setting up the circuit.

Circuit with rest stations. The addition of rest stations allows a break for the candidates and examiners and may allow for the addition of an extra candidate if required (Humphris & Kaney 2001). Care should be taken to keep this station private, so that the candidate at this station cannot over hear what is being said at the other stations. It should be clearly marked, and candidates should be informed of its presence before the start of the examination, although ideally students will have had practice sessions to familiarise themselves with the examination circuit. It is extremely important to bear in mind that the circuit cannot start or finish with a rest station. If the rest stations are at the beginning or the end of a circuit a candidate will end up missing one or more stations. For this purpose the rest stations should be interspersed within the live stations.

Considerations for individual stations. In setting up individual stations, care must be taken to allocate space appropriate to the tasks, equipment and the personnel. For example, an unmanned station containing investigation results and some written questions would need just enough room for a table and chair, whereas a resuscitation station would need enough space for a manikin, a defibrillator and an examiner. The stations should provide an appropriate environment for the candidates to perform the procedures. For instance, adjustable lighting for Fundoscopy or a quiet area for auscultation of the chest should be provided as appropriate (McCoy & Merrick 2001). Some stations may also require power sockets for the equipment.

The equipment. The equipment required for each OSCE station is included in the documentation developed at the station writing stage. All equipment should be sourced well in advance of the OSCE, and checked to ensure that it is in good working order. There should be spare equipment and batteries available on the day in case of breakages or breakdowns.

Box 3. Examination day briefings.

Information for examination day briefings

Candidates briefing
- A description of the circuit including their start stations, rest stations and pilot stations.
- Reminders of rules and regulations
- Quarantine procedures
- Fire procedures

Examiners briefing
- The objective of the examination, e.g. formative/summative
- To check students' identity at the start of the station
- An overview of the scoring rubric and how to complete the mark sheets
- The importance of keeping stations and candidates' scores confidential
- Not to talk to the students any more than what is allowed in the script
- To treat all candidates equally
- The procedures for reporting concerns about candidates
- Completing feedback after the examination
- Fire and quarantine procedures

Standardised patient briefing
- The importance of standardisation between candidates
- Their role in providing feedback
- Rest-breaks and refreshment facilities
- Fire procedures

Decisions ought to be made about candidates' use of their own equipment during the examination. If candidates are expected to bring their own stethoscopes for instance, they should be informed of this.

If more advanced equipment is required such as high fidelity human patient simulators there must be personnel available who are able to programme and run these, as most examiners will not be familiar with such equipment.

Examination day briefings. On the day of the examination there should be separate briefing sessions for the candidates, examiners and SPs. If there has been prior training and if written instructions have been provided they need only be brief and should be kept succinct. Key information that may be included is outlined below (Box 3).

Table 5. Common problems and troubleshooting tips.

Problem	Potential solution
Variable performances by SPs affecting station standardisation	Occasionally SPs may change their behaviour between candidates or provide unsolicited information. Robust selection and training procedures should mininise these issues. Examiners should also be aware of this potential problem and be willing to intervene between candidates if necessary.
Equipment failure	There should always be spare equipment readily available at hand. If candidates lose a lot of time waiting for spare equipment it may be possible for them to retake the station at the end of the examination.
Unpredictable behaviour of candidates	Nervous candidates under stress can often act in unpredictable ways. In particular, getting lost in the venue or on the OSCE circuit. Adequate support staff should be available to help direct candidates and answer any queries. Examiners may have to prompt candidates to move on at the correct time if bells or voice commands are missed.
Removal of instructions or equipment from stations by candidates	Instructions can be firmly secured to a table or lectern. Examiners and support staff should be vigilant for candidates leaving stations with equipment
Removal of mark sheets or station information by examiners	This may preclude the station form being used in subsequent sittings and examiners should be warned that no documentation must leave the station. Support staff collecting documentation prior to examiners leaving the station can reduce the chance of this occurring.

Running the OSCE circuit and troubleshooting

Running the circuit. The movement of the candidates from one station to another can either be managed by ringing a bell manually or by using automated PowerPoint™ presentations set up with voice commands clearly instructing the candidates and the examiners. The OSCE starts with the command 'Start Preparation', during which time the candidates read the question, followed one minute later with instructions to 'enter the station'. The next instruction could be 'one minute left' and the station would end a minute later with the command 'move on'. During a formative examination an additional command 'start feedback' at an appropriate time interval could also be included. The cycle is repeated for the duration of the examination. This system may be preferable to the use of bells as it reduces the confusion as to what each ring of the bell signifies. However, if an automated system of commands is used, a back-up in case of technical failure is essential, which could be a simple stopwatch and a bell.

Once the examination is started there should be personnel available to ensure that the candidates move in the right direction. If any SPs need a break they should be replaced promptly with reserve SPs as described earlier. At the end of the examination the marking sheets are collected and the stations are reset for the following run of the circuit if needed.

Quarantine. Quarantine refers to separating those candidates who have completed the examination from those who have yet to take it on the same day. The same set of OSCE stations may be in use for both morning and afternoon sessions, allowing exchange of information if the morning candidates are allowed to leave prior to all the afternoon candidates arriving. This may lead to a perceived unfair advantage to the second set of candidates. To resolve this issue, candidates scheduled for the early circuits should be 'quarantined' in a separate room until all of the later candidates have arrived and registered. Mobile phones and other devices with the means for remote communication should not be permitted in the examination centres.

Trouble shooting. On the day of the examination a number of issues can arise, some common issues and their potential solutions are described below (some of this information is taken from the Queens University of Belfast's website on OSCE training available at http://www.med.qub.ac.uk/osce/background_Dilemma.html) (Table 5).

Eva's Story (continued II)

So we did it! What an exhausting day it was but we are very pleased with how it went. The candidates, patients and examiners all turned up and knew what to do. The meetings and planning we had been through were all worth it. The examination ran smoothly, except for a few hiccups with equipment failure, but we had anticipated it and were able to replace faulty tools with spares. The patients also became quite exhausted by the end of the day and I think we will recruit more reserves in future.

Now that we've got the OSCE itself out of the way, we can all let out a huge sigh of relief but there is still quite a lot of work to do. I was reminded of this as soon as I arrived in my office today to check my emails; there were a few from students asking when they would get their results. We need to collate all the marks and publish them. I am looking forward to analysing the candidates' results and the psychometrics of our OSCE; we should be able to extract some really valuable statistics to help us in improving things for next time. I've been in discussions with our psychometrician who has already helped a great deal and will now be invaluable.

The quality assurance process has been important to us from day one; we didn't just want to put on OSCE for the sake of it, we wanted a reliable and valid assessment that assessed skills not tested by our other tools. This process continues now, with feedback, evaluation and psychometrics. We can use all of this information to keep improving our OSCE for future students.

Post-OSCE considerations

Handling results

Following the examination the mark sheets are collected and cross-checked for accuracy and any missing scores. The examiners are contacted if any corrections need their

verification. These results are put in appropriate spreadsheets and cross-checked again in preparation for the examination boards for ratification as described below.

The examination boards and ratification

After compilation the results are made available to the examination boards for the purposes of ratification. The examination board ratifies the results and signs them off as accurate. In case of any doubts the results are verified again. In cases of poor performance or failure the penalties are decided at these meeting and later conveyed to the students.

Publication of results

After the ratification the publication of accurate results is the final responsibility of the Assessment Team. The results could be made available online as well as sent to the students as hardcopies.

Complaints and appeals

There may be mitigating circumstance appeals or complaints made by candidates or examiners that need to be dealt with fairly and promptly after each examination. There will often be institutional policies and procedures to follow under these circumstances. Valid complaints may help to inform changes to the examination as a part of the quality assurance process.

Quality assurance

The quality assurance of each examination is a continuous process repeated with each examination cycle. Although many quality assurance procedures take place following the OSCE, quality assurance by training examiners, peer reviewing stations and ensuring standardisation are also quality assurance measures that take place before the conduct of the examination. The Figure 2 highlights the factors contributing to quality assurance. Those that have not yet been addressed will be described in more detail (Figure 2).

External examiners

External examiners may be invited from different institutions to inform and comment on whether academic standards are being maintained between institutions and also to ensure the assessment process measures student achievement rigorously and fairly and is conducted in line with policies and regulations.

Post-hoc psychometrics

Post-hoc analysis of OSCE results allows the determination of the reliability of scores generated by the examination. This topic has been briefly dealt with in the sections on Station Length and Station Bank development. Although a detailed discussion is beyond the scope of this Guide, we would like to further address the concept of reliability at this stage for the sake of completeness.

The reliability of OSCE scores can be measured as Cronbach's α or G coefficient as mentioned earlier. Each of

Figure 2. Elements of OSCE quality assurance.

these coefficients represents the error in the scores generated by the OSCE. A coefficient of 1 means there is no error in the scores and all variance is true variance. A Cronbach's α or G coefficient of 0.7 to 0.8 is taken as an acceptable level of reliability for high stakes examinations. PMETB in the UK advocates a minimum reliability coefficient of 0.9 as a standard for high stakes Royal College examinations (PMETB 2007).

Application of Cronbach's α allows the detection of the OSCE stations which are main sources of error, by removing one station at a time from the analysis and looking at the reliability of the remainder. Application of G theory allows the identification of various other sources of error including the items, assessors and interaction of candidates with items and assessors etc. Item Response Theory also generates results somewhat similar to the G theory, but does not have the capacity to predict the reliability if the number of the stations was altered.

It is essential to perform psychometric analysis on the OSCE results and use the outcomes to enhance the quality of the examinations. Departments and institutions running OSCEs should seek help from Psychometricians in this respect. AMEE Guides 54 and 66 address post-hoc psychometrics (Tavakol & Dennick 2011, 2012).

Evaluation

The feedback on the examination process provided by the examiners can be used to improve the quality of the stations and organisation of the future examinations. Generally after each sitting of the OSCEs the examiners are invited to provide written comments on the individual stations they had examined on (Kowlowitz et al. 1991). Any issues such as undue difficulty of tasks, lack of clarity of instructions for the candidates and appropriateness of tasks for completion in the

allocated time are highlighted and addressed based on this information.

Candidates may also be invited to provide feedback on their experience of the examination as part of the quality assurance process (Williams 2004).

Conclusion

Part I of this Guide introduced the concept of the OSCE and explained the theoretical principles underlying its use as one part of a battery of assessment tools. Part II has focussed more on the organisational and practical factors for consideration while setting up an OSCE.

The key strength of OSCE is in its standardisation and reliability when compared to older forms of performance assessment, this reliability must not be compromised by poor planning or insufficient training of station-writers, station-examiners or standardised patients.

Organising and planning an OSCE from scratch is a huge task which requires a lot of logistical groundwork and training for all those involved. Good management and awareness of potential problems make the actual running of the OSCE easier. The quality assurance processes include post-hoc psychometrics to determine reliability and stations' quality. Together with the evaluation this psychometric data helps to improve the future examinations.

The instructions and advice in this two part Guide should help planners and faculty through every stage of the organisation of an OSCE, from understanding and application of the underlying theory to the administration, organisation, evaluation and quality assurance.

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

Notes on contributors

DR KAMRAN KHAN, MBBS, FRCA, FAcadMEd, MScMedEd, at the time of production of this manuscript was a Senior Teaching Fellow Manchester Medical School and Consultant Anaesthetist, LTHTR, Preston, UK. During his career Dr Khan has developed a passion for medical education and acquired higher qualification in this field alongside his parent speciality in anaesthetics. He has held clinical academic posts, first at the University of Oxford and then at the University of Manchester. He was Associate lead for 'Assessments' at the Manchester Medical School. He is currently working in the UAE as a Consultant Anaesthetist. He has extensively presented and published at the national and international levels in his fields of academic interest.

DR KATHRYN GAUNT, MBChB, MRCP, PGD(MedEd), at the time of production of this manuscript was a Medical Education and Simulation Fellow, LTHTR, Preston, UK. Dr Kathryn Gaunt graduated in 2002 from the University of Manchester Medical School, UK. She was a Specialty Registrar in Palliative Medicine and has a keen interest in medical education. She is currently pursuing a higher qualification in this field. She has recently taken time out of her training programme to work as a Medical Education and Simulation Fellow at the Lancashire Teaching Hospitals NHS Trust. Here she lectured, examined for OSCEs and was involved in research and development within the field. She has now returned to her training in Palliative Medicine.

DR SANKARANARAYANAN RAMACHANDRAN, MBBS, PgCert(MedEd), FHEA, is a Medical Education and Simulation Fellow, Lancashire Teaching Hospitals NHS Trust, Preston, UK. Dr Ramachandran qualified in India and underwent postgraduate training in General Medicine. He had been teaching and training undergraduate and postgraduate medical students throughout his career. He has completed a Postgraduate Certificate and pursuing a Master's degree in Medical Education at the University of Manchester. He has special interest in assessment, problem based learning and research in medical education.

Dr PIYUSH PUSHKAR, MBChB, works in Critical Care at the Aintree University Hospitals Trust, Liverpool, UK. Piyush qualified from the University of Edinburgh and has subsequently worked and trained in anaesthesia and critical care in Lancashire and Merseyside, as well as aeromedical retrieval medicine in Queensland, Australia. He also works as a volunteer doctor for Freedom from Torture in Manchester.

References

Adamo G. 2003. Simulated and standardized patients in OSCEs: Achievements and challenges 1992–2003. Med Teach 25:262–270.

Angoff WH. 1971. Scales, norms and equivalent score. In: Thorndike RL, editor. Educational measurement. 2nd ed. Washington DC: American Council on Education. pp 508–600.

Barrows SH. 1993. An Overview of the uses of standardized patients for teaching and evaluating clinical skills. Acad Med 68:443–451.

Bloch R, Norman G. 2012. Generalizability theory for the perplexed: A practical introduction and guide: AMEE Guide No. 68. Med Teach 34:960–992.

Cleland JA, Abe K, Rethans J-J. 2009. The use of simulated patients in medical education: AMEE Guide No 42. Med Teach 31:477–486.

Cohen-Schotanus J, Van Der Vleuten CP. 2010. A standard setting method with the best performing students as point of reference: Practical and affordable. Med Teach 32:154–160.

Cohen R, Reznick R, Taylor B, Provan J, Rothman A. 1990. Reliability and validity of the objective structured clinical examination in assessing surgical residents. Am J Surg 160:302–305.

Collins JP, Harden RM. 1998. AMEE Medical Education Guide No. 13: Real patients, simulated patients and simulators in clinical examinations. Med Teach 20(6):508–521.

Cusimano MD, Cohen R, Tucker W, Murnaghan J, Kodama R, Reznick R. 1994. A comparative analysis of the costs of administration of an OSCE (objective structured clinical examination). Acad Med 69:567–570.

Downing SM. 2003. Item response theory: Applications of modern test theory in medical education. Med Educ 37:739–745.

Ebel R. 1972. Essentials of educational measurement. New Jersey, NJ: Prentice-Hall.

Epstein RM. 2007. Assessment in Medical Education. N Eng J Med 356:387–396.

Friedman Ben-David M. 2000. Standard setting in student assessment. Association for Medical Education in Europe.

GMC. 2009. Tomorrow's Doctors [Online]. London: GMC. [Accessed 10 June 2012] Available from http://www.gmc-uk.org/static/documents/content/GMC_TD_09__1.11.11.pdf

Harden RM, Stevenson M, Downie WW, Wilson GM. 1975. Assessment of Clinical Competence using Objective Structured Examination. BMJ 1:447–451.

Hodges B, Mcilroy JH. 2003. Analytic global OSCE ratings are sensitive to level of training. Med Educ 37:1012–1016.

Humphrey-Murto S, Smee S, Touchie C, Wood TJ, Blackmore D. 2005a. A Comparison of Physician Examiners and Trained Assessors in a High-Stakes OSCE Setting. Acad Med 80:S59–S62.

Humphrey-Murto S, Wood TJ, Touchie C. 2005b. Why do physicians volunteer to be OSCE examiners? Med Teach 27:172–174.

Humphris GM, Kaney S. 2001. Examiner fatigue in communication skills objective structured clinical examination. Med Educ 35:444–449.

Kaufman DM, Mann KV, Muijtjens AM, Van Der Vleuten CP. 2000. A comparison of standard-setting procedures for an OSCE in undergraduate medical education. Acad Med 75:267–271.

Khan K, Ramachandran S. 2012. Conceptual Framework for Performance Assessment: Competency, Competence and Performance in the Context of Assessments in Healthcare – Deciphering the Terminology. Med Teach 34:920–928.

Kowlowitz V, Hoole AJ, Sloane PD. 1991. Implementing the objective structured clinical examination in a traditional medical school. Acad Med 66:345–347.

Kramer A, Muijtjens A, Jansen K, Dusman H, Tan L, Van Der Vleuten C. 2003. Comparison of a rational and an empirical standard setting procedure for an OSCE. Objective structured clinical examinations. Med Educ 37:132–139.

Mann KV, Macdonald AC, Nornici JJ. 1990. Reliability of objective structured clinical examinations: Four years of experience in a surgical clerkship. Teach Learn Med 2:219–224.

Mccoy JA, Merrick HW, editors. 2001. The Objective Structured Clinical Examination. Association for Surgical Education. Springfield, IL.

Morgan PJ, Cleave-Hogg D, Guest CB. 2001. A comparison of global ratings and checklist scores from an undergraduate assessment using an anesthesia simulator. Acad Med 76:1053–1055.

Newble D. 2004. Techniques for measuring clinical competence: Objective structured clinical examinations. Med Educ 38:199–203.

Newble DI, Hoare J, Sheldrake PF. 1980. The selection and training of examiners for clinical examinations. Med Educ 14:345–349.

Pell G, Fuller R, Homer M, Roberts T. 2010. How to measure the quality of the OSCE: A review of metrics - AMEE guide no. 49. Med Teach 32:802–811.

PMETB. 2007. Developing and maintaining an assessment system - a PMETB guide to good practice [Online]. London: PMETB. [Accessed 10 June 2012] Available from http://www.gmc-uk.org/assessment_good_practice_v0207.pdf_31385949.pdf

Reznick RK, Regehr G, Yee G, Rothman A, Blackmore D, Dauphinee D. 1998. Process-rating forms versus task-specific checklists in an OSCE for medical licensure. Medical Council of Canada. Acad Med 73:S97–S99.

Roberts C, Newble D, Jolly B, Reed M, Hampton K. 2006. Assuring the quality of high-stakes undergraduate assessments of clinical competence. Med Teach 28:535–543.

Shavelson R, Webb N. 2009. Generalizability theory and its contributions to the discussion of generalizability of research findings. In: Ercikan K, Roth W-M, editors. Generalizing from educational research: Beyond qualitative and quantitative polarization. New York, London: Routledge. pp 13–32.

Shumway JM, Harden RM. 2003. AMEE Guide No. 25: The assessment of learning outcomes for the competent and reflective physician. Med Teach 25:569–584.

Stevens DD, Levi A. 2005. Introduction to rubrics: An assessment tool to save grading time, convey effective feedback, and promote student learning. Sterling, VA: Stylus Pub.

Stillman PL. 1993. Technical Issues: Logistics. Acad Med 68:464–468.

Tan GP, Azila NM. 2007. Improving OSCE examiner skills in a Malaysian setting. Med Educ 41:517.

Tavakol M, Dennick R. 2011. Post Examination Analysis of Objective Tests: AMEE Guide 54. Med Teach 33:245–248.

Tavakol M, Dennick R. 2012. Post-examination interpretation of objective test data: Monitoring and improving the quality of high-stakes examinations: AMEE Guide No. 66. Med Teach 34:e161– e175.

Turner JL, Dankoski ME. 2008. Objective Structured Clinical Exams: A Critical Review. Fam Med 40:574–578.

Van Der Vleuten CPM, Van Luyk SJ, Van Ballegooijen AMJ, Swanson DB. 1989. Training and experience of examiners. Med Educ 23:290–296.

Vargas AL, Boulet JR, Errichetti A, Zanten MV, López MJ, Reta AM. 2007. Developing performance-based medical school assessment programs in resource-limited environments. Med Teach 29:192–198.

Whelan GP. 1999. Educational commission for Foreign Medical Graduates – clinical skills assessment prototype. Med Teach 21:156–160.

Wilkinson TJ, Frampton CM, Thompson-Fawcett M, Egan T. 2003. Objectivity in Objective Structured Clinical Examinations: Checklists Are No Substitute for Examiner Commitment. Acad Med 78:219–223.

Williams RG. 2004. Have Standardized Patient Examinations Stood the Test of Time and Experience? Teach Learn Med 16:215–222.

Appendix 1

Observed OSCE station Filled Template

Question Info

Author
AQ

Subject/Topic
Asthma

Level of the Candidate (choose at least 1)
[x] Year 3
[] Year 4
[x] Year 5

Competencies (Essential Field)
Please choose 1 to 3
[x] Clinical Examination Skills
[] Communication & Consultation Skills
[x] History Taking
[] Hand Over Skills
[] Procedural Skills
[] Mental Health Assessment
[] Professionalism
[x] Application of Knowledge
[] Prescription
[] Data Interpretation
[] Diagnostic Skills
[] Management Planning
[] Ethics

Station Duration (please do not modify this field)
8 min

Information for the site organisers

SP age & sex
Mark/Mary Freeman, age 32

Resources & Equipment needed
1. Paper and pen, in case the candidate wishes to make notes
2. Alcogel™
3. Couch for the patient
4. Stethoscope
5. Water bottle and glasses

Setting up the station
1. Examiner's chair should be positioned so that he/she can observe faces of both candidate and simulated patient.
2. No desk or table is necessary. If one is present, it should NOT form a barrier between the candidate and the patient.

Instructions for Candidates (outside the station)

You are a medical student, on your placement at the GP Practice
Mark/Mary Freeman is a 32-year-old patient who is attending for their annual asthma review. This is the first time that Mr/Mrs Freeman has attended this year.

(1) You are expected to take a brief, focused asthma history from this patient to assess his/her asthma control.
(2) Perform a focused respiratory system examination

Please do not take a detailed history
Please do not perform a general physical examination

Information for the examiner

Brief Background to the scenario
Mark/Mary Freeman is a 32-year-old patient who is attending for their annual asthma review

Examiner's Role
Your role is to observe the history taking process and to assess the candidate's examination of the respiratory system

What are the objectives of the station or what is expected of the candidate?
- The candidate is expected to take a brief, focused asthma history from this patient to assess his/her asthma control and make to suggestions as appropriate
- The candidate is also expected to examine this patient's chest

The candidate has NOT been asked to take a detailed history or perform a general physical examination

What information they might be able to provide the candidate?
- If the candidate attempts to perform a general physical examination, please ask them to move on to examining the chest
- If candidates attempt to perform tactile vocal fremitus and whispering pectoriloquy please ask them to omit these and move on

What information they should not provide the candidate?
- Please do not repeat instructions to the candidate
- Please do not interrupt or prompt candidate or ask any questions

Clinical information relevant to the station

Key points in Asthma history taking;
1. Onset
2. Duration
3. Family History
4. Occupation
5. Smoking history
6. Housing, soft toys
7. Aggravating and relieving factors, e.g. exercise, weather, allergies, etc.
8. Cough
9. Wheeze
10. Perennial versus Seasonal
11. Diurnal variation/night time symptoms
12. Affect on lifestyle
13. Hospital admissions
14. Treatment including steroids

Key Points in Examination of the Respiratory System;

(A good student would complete the examination at the front before moving to the back)

1. Proper exposure
2. Inspection for symmetry and shape
3. Palpation of trachea, chest expansion (vocal fremitus not necessary in this case)
4. Percussion, front and back including the clavicles
5. Auscultation at front and back, (whispering pectoriloquy not necessary in this case)

British Thoracic Society (BTS) guidelines

BTS/Sign Guidance on pets in asthma: There are no controlled trials on the benefits of removing pets from the home. If you haven't got a cat, and you've got asthma, you probably shouldn't get one.

BTS/Sign Guidance on smoking and asthma: Direct or passive exposure to cigarette smoke adversely affects quality of life, lung function, need for rescue medications and long-term control with inhaled steroids.

BTS/Sign Guidance on control of asthma: The aim of asthma management is control of the disease. Complete control is defined as:

- no daytime symptoms
- no night time awakening due to asthma
- no need for rescue medication
- no exacerbations
- no limitations on activity including exercise
- normal lung function (in practical terms FEV1 and/or PEF >80% predicted or best)
- minimal side effects from medication

Simulated Patient Information

Who they are?
Mark/Mary Freeman, age 32

Their social/economic background if applicable?

- You are currently working in a telesales call centre (trying to sell fitted kitchens)
- You live in your own small terraced house, with your partner and one daughter
- You bought your daughter a cat 4 months ago (for her birthday)
- Your partner smokes 20 a day and you smoke occasionally – when you go out socialising. You know you shouldn't, but you don't think an occasional one harms your chest. You've been doing this since you were 16
- You have 2–3 drinks (pints of beer or large glasses of wine) when out socialising – which is once a week or less often
- You go to the gym very occasionally
- Your mood is normal (not anxious or depressed)

History

- You have had Asthma most of your life – you think it was diagnosed when you were about 5 or 6
- It seemed to be much worse as a child – you had been under a hospital clinic, but you have never had to stay in hospital
- As a teenager, it seemed to go much better – you rarely seemed to need treatment
- In your adult years, it hasn't troubled you too much, though you have always liked to have a blue inhaler (salbutamol) handy, just in case
- If asked, you have never taken a course of steroid tablets by mouth and you do not own your peak flow meter
- One inhaler usually lasted you several months, as you only needed it every few days (but see below for current situation). When you use it, you take 2 puffs at a time (as instructed), which relieves the coughing and wheezing noises within 5 min or so)
- You have an elder brother with asthma, a your daughter has mild eczema and your father has had hay fever all his life
- Your asthma is not seasonal
- It does not vary by day or night

Details of their concerns/perceptions

- You also wonder if the cat might be to blame. On the other hand, you had a dog when you were young, which didn't affect you, so you don't think you are allergic to animal fur
- You are not really worried about your asthma – you don't think it's anything serious, but you're happy to come for a check-up

What they should say (their agenda) & what they should not say
You think the reason why your asthma is worse is because of working longer hours. There is a fair amount of stress at work, as the company's sales are not doing well. You don't feel that you are getting over-stressed; it's just the general atmosphere at work

What they should ask (questions)
Please do not ask any questions

Specific Standardisation issues (specific answers to specific questions, please stay with the script)
If the candidate asks for your thoughts about why your asthma has worsened very early on (before making any effort to establish rapport), say you believe it's the stress at work. But if the candidate asks the same question when you have established some rapport, mention your belief about it being the stress at work, but also say: *It's been ever since my daughter's birthday, I suppose*. If they probe why you think that might be, mention the cat.

Appendix 2

OSCE QA Questionnaire

Question Title/Number

Feasibility Question

How easy will it be to find an SP for this question?	Very easy	Relatively easy	Difficult

Validity Questions

Are the tasks in this station achievable in 8 Minutes?	Yes	No
Is this topic taught in curriculum?	Yes	No
Where is it taught in curriculum?		

The attributes being tested are expected of a FY1 doctor.	Yes	No
Would this question discriminate between good and poor students?	Yes	No
Would this station be able to recreate an atmosphere close to real patient encounter?	Yes	No
If a candidate passes this station would that be able to extrapolated to competence in workplace?	Yes	No
Does this question test what it is supposed to test – as outlined in primary and secondary competences?	Yes	No

Supplemental Questions

Would this question benefit from piloting if unsure about any of the above?	Yes	No
Can this question be used in other years as well?	Yes	No
If yes please indicate which year		

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

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

เชิดศักดิ์ ไอรมณีรัตน์

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

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

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

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

เชิดศักดิ์ ไอรมณีรัตน์

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

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

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

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

เชิดศักดิ์ ไอรมณีรัตน์

ในบทความนี้จะขอเสนอเกร็ดความรู้เกี่ยวกับการให้คะแนนในการสอบ OSCE

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

17 March 2017

Clinical Performance Ratings

เชิดศักดิ์ ไชยมณีรัตน์
ภาควิชาศัลยศาสตร์ คณะแพทยศาสตร์ศิริราชพยาบาล
มหาวิทยาลัย มหิดล

Competence and Performance

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

Clinical Performance Ratings

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

Clinical Performance Ratings

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

Reducing Rater Errors

- Improving raters
- Improving a rating instrument

Improving Raters

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

Cherdsak.ira@mahidol.ac.th

Rating Instrument

- Item
- Scale

Writing Effective Items

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

Characteristics of A Good Scale

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

Key Points: Performance Ratings

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

แบบประเมินการปฏิบัติงาน นักศึกษาแพทย์ชั้นปีที่ 6

นศพ. ชื่นบาน แช่มชื่นใจ รหัสประจำตัวนักศึกษา 6001999

ให้อาจารย์ทำเครื่องหมายกากบาท (X) ในระดับคะแนนที่เหมาะสม

ข้อ	10 (ดีมาก)	9	8	7	6 (พอใช้)	5	4	3	2	1 (ไม่ดีอย่างยิ่ง)
1. ความรู้พื้นฐานทางการแพทย์										
2. ทักษะการซักประวัติ และตรวจร่างกาย										
3. ทักษะการวินิจฉัยโรค										
4. ความสามารถในการเลือกการตรวจค้นเพิ่มเติมที่เหมาะสม										
5. การเลือกการรักษาที่เหมาะสม										
6. การคิดอย่างมีเหตุผล										
7. การสื่อสาร										
8. มนุษยสัมพันธ์										
9. บุคลิกภาพ และการแต่งกาย										
10. ความเป็นวิชาชีพแพทย์										

แบบประเมินการปฏิบัติงานของนักศึกษาแพทย์ปี 6
คณะแพทยศาสตร์ศิริราชพยาบาล

นศ.พ. รหัส กลุ่ม

ฝึกปฏิบัติงานที่ ภาควิชา/แผนก

หอผู้ป่วย ช่วงเวลาปฏิบัติงาน ถึง

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

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

ผู้ประเมิน .....
(.....)
วันที่ .....

หมายเหตุ กรุณาใส่คะแนนในช่องสีเหลี่ยมหน้าช่องที่ต้องการ (ไม่มีจุดทศนิยม), NA = ไม่สามารถประเมินได้

2007; 29: 855–871 MEDICAL TEACHER

AMEE GUIDE

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

JOHN NORCINI[1] & VANESSA BURCH[2]

[1]Foundation for Advancement of International Medical Education and Research, Philadelphia, USA, [2]University 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 time 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

ISSN 0142–159X print/ISSN 1466–187X online/07/09-100855–17 © 2007 Informa UK Ltd.
DOI: 10.1080/01421590701775453

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 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).

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.

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	☐	☐	☐	☐	☐	☐	☐
2. Physical Examination Skills	☐	☐	☐	☐	☐	☐	☐
3. Communication Skills	☐	☐	☐	☐	☐	☐	☐
4. Clinical Judgement	☐	☐	☐	☐	☐	☐	☐
5. Professionalism	☐	☐	☐	☐	☐	☐	☐
6. Organisation/Efficiency	☐	☐	☐	☐	☐	☐	☐
7. Overall clinical care	☐	☐	☐	☐	☐	☐	☐

*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): M M / Y Y

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*
1. Demonstrates understanding of indications, relevant anatomy, technique of procedure	☐	☐	☐	☐	☐	☐	☐
2. Obtains informed consent	☐	☐	☐	☐	☐	☐	☐
3. Demonstrates appropriate preparation pre-procedure	☐	☐	☐	☐	☐	☐	☐
4. Appropriate analgesia or safe sedation	☐	☐	☐	☐	☐	☐	☐
5. Technical ability	☐	☐	☐	☐	☐	☐	☐
6. Aseptic technique	☐	☐	☐	☐	☐	☐	☐
7. Seeks help where appropriate	☐	☐	☐	☐	☐	☐	☐
8. Post procedure management	☐	☐	☐	☐	☐	☐	☐
9. Communication skills	☐	☐	☐	☐	☐	☐	☐
10. Consideration of patient/professionalism	☐	☐	☐	☐	☐	☐	☐
11. Overall ability to perform procedure	☐	☐	☐	☐	☐	☐	☐

*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): M M / Y Y

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

Case-based Discussion (CbD) - F2 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: Pain ☐ Airway/ Breathing ☐ CVS/ Circulation ☐ Psych/ Behav ☐ Neuro ☐ Gastro ☐ Other

Focus of clinical encounter: Medical Record Keeping ☐ Clinical Assessment ☐ Management ☐ Professionalism ☐

Complexity of case: Low ☐ Average ☐ High ☐ Assessor's position: Consultant ☐ SpR ☐ GP ☐

Please grade the following areas using the scale below:	Below expectations for F2 completion 1	2	Borderline for F2 completion 3	Meets expectations for F2 completion 4	Above expectations for F2 completion 5	6	U/C*
1 Medical record keeping	☐	☐	☐	☐	☐	☐	☐
2 Clinical assessment	☐	☐	☐	☐	☐	☐	☐
3 Investigation and referrals	☐	☐	☐	☐	☐	☐	☐
4 Treatment	☐	☐	☐	☐	☐	☐	☐
5 Follow-up and future planning	☐	☐	☐	☐	☐	☐	☐
6 Professionalism	☐	☐	☐	☐	☐	☐	☐
7 Overall clinical judgement	☐	☐	☐	☐	☐	☐	☐

*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:

Trainee satisfaction with CbD (Not at all) 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ 10 ☐ (Highly)

Assessor satisfaction with CbD 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ 10 ☐

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

Time taken for discussion: (in minutes)

Assessor's Signature: .....

Date: / /

Time taken for feedback: (in minutes)

Assessor's Surname

Assessor's GMC Number

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

2466400642

Figure 3. Case-based assessment form. Source: www.mmc.nhs.uk.

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	☐	☐	☐	☐	☐	☐	☐
2 Ability to formulate appropriate management plans	☐	☐	☐	☐	☐	☐	☐
3 Awareness of their own limitations	☐	☐	☐	☐	☐	☐	☐
4 Ability to respond to psychosocial aspects of illness	☐	☐	☐	☐	☐	☐	☐
5 Appropriate utilisation of resources e.g. ordering investigations	☐	☐	☐	☐	☐	☐	☐
Maintaining good medical practice							
6 Ability to manage time effectively / prioritise	☐	☐	☐	☐	☐	☐	☐
7 Technical skills (appropriate to current practice)	☐	☐	☐	☐	☐	☐	☐
Teaching and Training, Appraising and Assessing							
8 Willingness and effectiveness when teaching/training colleagues	☐	☐	☐	☐	☐	☐	☐
Relationship with Patients							
9 Communication with patients	☐	☐	☐	☐	☐	☐	☐
10 Communication with carers and/or family	☐	☐	☐	☐	☐	☐	☐
11 Respect for patients and their right to confidentiality	☐	☐	☐	☐	☐	☐	☐
Working with colleagues							
12 Verbal communication with colleagues	☐	☐	☐	☐	☐	☐	☐
13 Written communication with colleagues	☐	☐	☐	☐	☐	☐	☐
14 Ability to recognise and value the contribution of others	☐	☐	☐	☐	☐	☐	☐
15 Accessibility/Reliability	☐	☐	☐	☐	☐	☐	☐
16 Overall, how do you rate this doctor compared to a doctor ready to complete F1 training?	☐	☐	☐	☐	☐	☐	☐

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:	☐ British	☐ Bangladeshi
	☐ Irish	☐ Other Asian Background
	☐ Other White Background	☐ White and Black Caribbean
	☐ Caribbean	☐ White and Black African
	☐ African	☐ White and Asian
	☐ Any other Black background	☐ Any other mixed background
	☐ Indian	☐ Chinese
	☐ Pakistani	☐ Any other ethnic group
Which environment have you primarily observed the doctor in? **(Please choose one answer only)**	☐ Inpatients	☐ Intensive Care
	☐ Outpatients	☐ Theatre
	☐ Both In and Out-patients	☐ General Practice
	☐ A&E/Admissions	☐ Other (Please specify)
	☐ Community Speciality	
	☐ Laboratory/Research	
Your position:	☐ Consultant ☐ SASG	☐ SpR ☐ Foundation/PRHO
	☐ Nurse ☐ SHO	☐ Allied Health Professional
	☐ GP	
	☐ Other (Please specify)	

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	Feedback techniques to be adopted
Creating a disrespectful, unfriendly, closed, threatening climate	Creating a respectful, open minded, non-threatening climate
Not eliciting thoughts or feelings before giving feedback	Eliciting thoughts and feelings before giving feedback
Being judgemental	Being non-judgemental
Focusing on personality	Focusing on behaviours
Basing feedback on hearsay	Basing feedback on observed facts
Basing feedback on generalizations	Basing feedback on specifics
Giving too much/too little feedback	Giving the right amount of feedback
Not suggesting ideas for improvement	Suggesting ideas for improvement
Basing feedback on unknown, non-negotiated goals	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, Trotshcel 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 Ame 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 elf-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, Home 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, Scherpbier 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 evalaution exercie). 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 recommened 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 studnets' clinical skills and behaviours in medical school. Acad Med 74:841–849.

Kirch W, Schafii 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 evalaution 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 117:757–765.

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 Inter 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 medcial 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.

17 March 2017

Setting a Passing Standard:
How to decide who should fail an exam?

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

1

Standard

- A score that is set to be a boundary between those who perform well enough on the test (pass) from those who do not (fail).
- Standard = cutpoint

2

Objectives

- เมื่อสิ้นสุดการบรรยายแล้ว ผู้เข้าอบรมสามารถ
  - บอกถึงความสำคัญของการตั้งเกณฑ์ผ่านได้ถูกต้อง
  - บอกถึงขั้นตอนของการตั้งเกณฑ์ผ่านได้ถูกต้อง
  - ยกตัวอย่างวิธีการตั้งเกณฑ์ผ่านได้อย่างน้อยสามวิธี
  - ยกตัวอย่างเกณฑ์ผ่านการสอบ MCQ ด้วยวิธีการ modified Angoff method ในการสอบที่ตนเกี่ยวข้องได้อย่างเหมาะสม

3

Outline

- Basic concepts
- Steps in setting standards
  - The type of standard
  - The method
  - Selecting judges
  - Standard setting meeting
  - Calculate the standards
  - Checking the standards

4

Basic Concepts

- A standard is an answer to the question, "How much is enough?"
- The classification of examinees into two groups can result in two types of wrong decisions
  - False positive: Passing an examinee who should fail the exam
  - False negative: Failing an examinee who should pass the exam

5

Judgment

1. Made by qualified judges
2. Meaningful to the persons who are making the decision
3. Made in a way that takes into account the purpose of the test

6

cherdsak.ira@mahidol.ac.th

1

Steps in Setting Standards

1. Deciding on the type of standard
2. Deciding on the method for setting standards
3. Selecting judges
4. Holding the standard setting meeting
5. Calculating the standards
6. Checking the standards after test

7

1. Types of Standards

- Absolute standard
- Relative standard

8

Absolute Standard

- The standard is fixed, based on specific criteria of performance, but may undergo periodic re-evaluation of the standard
- Strengths
  - A standard is known in advance
  - A stable performance level is required to pass the examination => content-related standard
  - Provide clear feedback to examinees
  - Nobody has to fail the exam if their knowledge/skills is adequate for the purpose of the exam.
  - Promote a collaborative learning environment.

9

Relative Standard

- The standard is set in reference to the group of examinees. The resulting standard may be reasonable providing a representative heterogeneous group.
- Strengths
  - The failure rate is stable, which in someway is easy for curriculum management

10

2.Methods for Setting Standards

1. Test-centered methods
2. Examinee-centered methods
3. Compromised methods

11

Test-Centered Methods

- The judges set standards by reviewing the test items and provide judgments regarding the "just adequate" level of performance on these items.
  - Angoff's method
  - Nedelsky's method
  - Ebel's method

12

Modified Angoff's Method

- The judgment
  - The probability that a borderline examinee would answer the test item correctly
- The passing score
  - The sum of all the probability of correct answers for all items on the exam

13

Nedelsky's Method

- The judgment
  - How many options a borderline examinee can eliminate from choosing in an item
- The passing score
  - The probability of correct answer for an item = 1/(the number of options not eliminated)
  - The passing score of the test = the sum of all the probability of correct answers of all items on the test

14

Ebel's Method

- The judgment
  - What is the level of difficulty of an item?
    - Easy/Medium/difficult
  - What is the level of importance of that content in clinical practice?
    - Essential/Important/Acceptable/Questionable
  - The probability that a borderline examinee will answer an item in each category correctly
- The passing score
  - The sum of all the probability of correct answers for all items on the exam

15

Examinee-Centered Methods

- The judges set a standard by reviewing the overall performance of examinees and determine who should pass and who should fail. The scores of examinees are reviewed and the passing score is set based on these judgments
  - Borderline-group method
  - Contrasting-groups method

16

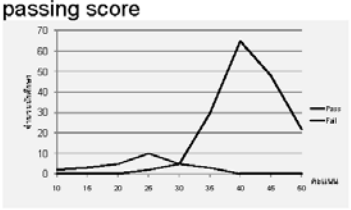
Borderline-Group Method

- The judgment
  - Identify examinees who are "borderline"
- The passing score
  - The median score of this "borderline group"

17

Contrasting-Groups Method

- The judgment
  - Identify examinees who should "pass" and those who should "fail"
- The passing score



18

Compromised Method

- Combining relative and absolute standard setting methods
  - Hofstee method

Hofstee Method

- The judgment
  - Minimum failure rate
  - Maximum failure rate
  - Minimum passing score
  - Maximum passing score
- The passing score
  - The intersection of test scores curve with diagonal line drawn from upper left to lower right corner

3. Selecting Judges

- The number of judges
- The qualification of judges

4. Standard Setting Meeting

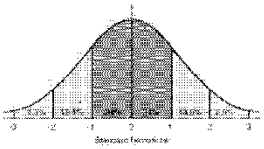
- Discussion of the purpose of the test, the characteristics of examinees, and the nature of competence.
- Explanation of the method and practice before the real standard setting procedure.

5. Calculating Standard

- Outliers
- Errors of the cutpoint

Do we have to care about error?

- 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

25

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

26

6. Checking Standard

- Stakeholders' acceptance of the results
- Relationship with other markers of competence
- Prediction of future performance

27

Summary

- Steps in setting up a standard

1. Deciding on the type of standard
2. Deciding on the method for setting standards
3. Selecting judges
4. Holding the standard setting meeting
5. Calculating the standards
6. Checking the standards after test

28

Questions & Comments

Cherdsak Iramaneerat
CherdsakIramaneerat@gmail.com

29

"Maybe the most any of us can expect of ourselves isn't perfection but progress."

Michelle Burford

30

กระดาษบันทึก

กระดาษบันทึก

กระดาษบันทึก

กระดาษบันทึก

กระดาษบันทึก

กระดาษบันทึก

Question & Comments

หน่วยพัฒนาแพทยศาสตรศึกษาและวิจัยการศึกษา
ฝ่ายการศึกษา คณะแพทยศาสตร์ศิริราชพยาบาล
สำนักงาน: ตึกอดุลยเดชวิกรม ชั้น 6 (ห้อง 656)
Tel. 02 419 9978 Fax. 02 412 3901

: www.si-merd.com
: merd.project@gmail.com
: MERD
: MERD FC