



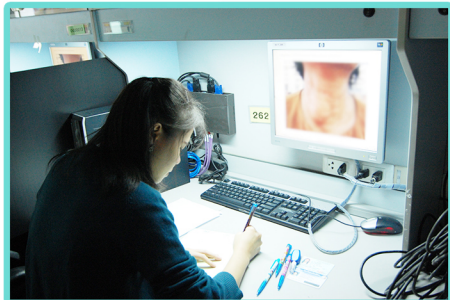
มหาวิทยาลัยมหิดล
คณะแพทยศาสตร์
ศิริราชพยาบาล

ศูนย์ความเป็นเลิศด้านการศึกษาวិทยาศาสตร์สุขภาพ (ศศว)
คณะแพทยศาสตร์ศิริราชพยาบาล มหาวิทยาลัยมหิดล

Assessment workshop for clinical teachers

การวัดและประเมินผลนักศึกษาชั้นคลินิก

วัดผลนักศึกษาอย่างไร
ให้ถูกต้อง เทียบตรง และเป็นธรรม



ระหว่างวันที่ 14 - 16 มีนาคม พ.ศ. 2561
ณ ห้องบรรยาย 3A01 อาคารศรีสวรินทิรา ชั้น 3A
คณะแพทยศาสตร์ศิริราชพยาบาล



	หน้า
กำหนดการ	1
รายชื่อผู้ร่วมอบรม	3
เอกสารประกอบการอบรม (วันที่ 14 มีนาคม 2561)	9
หัวข้อ : What is good assessment?	11
(วิทยากร : รศ.ดร. นพ.เชิดศักดิ์ ไธรมณีรัตน์)	
หัวข้อ : How to choose assessment method?	19
(วิทยากร : รศ.ดร. นพ.เชิดศักดิ์ ไธรมณีรัตน์)	
หัวข้อ : Validity	23
(วิทยากร : ผศ. พญ.กษณา รักขมณี)	
หัวข้อ : Reliability	31
(วิทยากร : รศ.ดร. นพ.เชิดศักดิ์ ไธรมณีรัตน์)	
หัวข้อ : Standard setting	35
(วิทยากร : ผศ. พญ.กษณา รักขมณี)	
หัวข้อ : Grading.....	39
(วิทยากร : รศ.ดร. นพ.เชิดศักดิ์ ไธรมณีรัตน์)	
หัวข้อ : Summary	43
(วิทยากร : รศ.ดร. นพ.เชิดศักดิ์ ไธรมณีรัตน์)	
เอกสารประกอบการอบรม (วันที่ 15 มีนาคม 2561)	45
หัวข้อ : Multiple-choice questions item development	47
(วิทยากร : ศ. พญ.บุญมี สถาปัตยวงศ์)	
หัวข้อ : Multiple-choice questions item analysis	71
(วิทยากร : ผศ. นพ.ตรีภพ เลิศบรรณพงษ์)	
หัวข้อ : Constructed response item development	97
(วิทยากร : ผศ. นพ.สุประพัฒน์ สนใจพานิชย์)	
หัวข้อ : Summary	141
(วิทยากร : รศ.ดร. นพ.เชิดศักดิ์ ไธรมณีรัตน์)	
เอกสารประกอบการอบรม (วันที่ 16 มีนาคม 2561)	143
หัวข้อ : OSCE item development	145
(วิทยากร : รศ.ดร. นพ.เชิดศักดิ์ ไธรมณีรัตน์)	
หัวข้อ : Long case examination	153
(วิทยากร : รศ. พญ.ปวพรรณ ภูมานะชัย)	
หัวข้อ : Portfolio	157
(วิทยากร : ผศ. นพ.ตรีภพ เลิศบรรณพงษ์)	
หัวข้อ : Clinical performance ratings	221
(วิทยากร : รศ.ดร. นพ.เชิดศักดิ์ ไธรมณีรัตน์)	
หัวข้อ : Workplace-based assessment	225
(วิทยากร : ผศ. พญ.กษณา รักขมณี)	
หัวข้อ : Summary	229
(วิทยากร : รศ.ดร. นพ.เชิดศักดิ์ ไธรมณีรัตน์)	
กระดาษบันทึก	231
ช่องทางการติดต่อสื่อสาร	237



กำหนดการอบรมเชิงปฏิบัติ เรื่อง Assessment workshop for clinical teachers

วันที่ 14 - 16 มีนาคม พ.ศ. 2561 ณ ห้องบรรยาย 3A01 อาคารศรีสวรินทิรา ชั้น 3A คณะแพทยศาสตร์ศิริราชพยาบาล

วันที่ 14 มีนาคม พ.ศ.2561 (Part 1 : หลักการพื้นฐานของการวัดผล)		วิทยากรหลัก	วิทยากรร่วม
08.30 – 09.30 น.	What is good assessment?	รศ. ดร.นพ.เชิดศักดิ์ ไอรมณีรัตน์	
09.30 – 10.00 น.	How to choose assessment methods?	รศ. ดร.นพ.เชิดศักดิ์ ไอรมณีรัตน์	
10.15 – 11.30 น.	Validity	ผศ. พญ.กษณา รักขมณี	
11.30 – 12.00 น.	Reliability	รศ. ดร.นพ.เชิดศักดิ์ ไอรมณีรัตน์	
12.00 – 13.00 น.	รับประทานอาหารกลางวัน		
13.00 – 14.30 น.	Standard setting	ผศ. พญ.กษณา รักขมณี	
14.45 – 15.45 น.	Grading	รศ. ดร.นพ.เชิดศักดิ์ ไอรมณีรัตน์	
15.45 – 16.00 น.	Summary	รศ. ดร.นพ.เชิดศักดิ์ ไอรมณีรัตน์	
วันพฤหัสบดีที่ 15 มีนาคม พ.ศ.2561 (Part 2 : การสอบข้อเขียน)		วิทยากรหลัก	วิทยากรร่วม
08.30 – 10.00 น.	Multiple-choice questions item development	ศ. พญ.บุญมี สถาปัตยกรรมศาสตร์	
10.15 – 11.00 น.	Multiple-choice questions item review	ศ. พญ.บุญมี สถาปัตยกรรมศาสตร์	รศ. นพ.สุพจน์ พงศ์ประสพชัย รศ. ดร.นพ.เชิดศักดิ์ ไอรมณีรัตน์ ผศ. พญ.กษณา รักขมณี ผศ. นพ.สุประพัฒน์ สนใจพานิชย์ ผศ. นพ.ตรีภพ เลิศบรรณพงษ์
11.15 – 12.00 น.	Multiple-choice questions item analysis	ผศ. นพ.ตรีภพ เลิศบรรณพงษ์	
12.00 – 13.00 น.	รับประทานอาหารกลางวัน		
13.00 – 14.30 น.	Constructed response item development	ผศ. นพ.สุประพัฒน์ สนใจพานิชย์	
14.45 – 15.45 น.	Constructed response item review	ผศ. นพ.สุประพัฒน์ สนใจพานิชย์	รศ. นพ.สุพจน์ พงศ์ประสพชัย รศ. ดร.นพ.เชิดศักดิ์ ไอรมณีรัตน์ ผศ. พญ.กษณา รักขมณี ผศ. นพ.ตรีภพ เลิศบรรณพงษ์ อ. นพ.อนุภ จิตต์เมือง
15.45 – 16.00 น.	Summary	รศ. ดร.นพ.เชิดศักดิ์ ไอรมณีรัตน์	
วันศุกร์ที่ 16 มีนาคม พ.ศ.2561 (Part 3 : การสอบปฏิบัติ)		วิทยากรหลัก	วิทยากรร่วม
08.30 – 10.00 น.	OSCE item development	รศ. ดร.นพ.เชิดศักดิ์ ไอรมณีรัตน์	อ. นพ.อนิรุต วรวาท
10.15 – 11.30 น.	OSCE item review	อ. นพ.อนิรุต วรวาท	รศ. นพ.สุพจน์ พงศ์ประสพชัย รศ. ดร.นพ.เชิดศักดิ์ ไอรมณีรัตน์ ผศ. นพ.สุประพัฒน์ สนใจพานิชย์ ผศ. นพ.ตรีภพ เลิศบรรณพงษ์ ผศ. พญ.กษณา รักขมณี
11.30 – 12.00 น.	Long case examination	รศ. พญ.พรพรรณ กุ้มานะชัย	
12.00 – 13.00 น.	รับประทานอาหารกลางวัน		
13.00 – 13.45 น.	Portfolio	ผศ. นพ.ตรีภพ เลิศบรรณพงษ์	
13.45 – 14.45 น.	Clinical performance ratings	รศ. ดร.นพ.เชิดศักดิ์ ไอรมณีรัตน์	
15.00 – 15.45 น.	Workplace-based assessment	ผศ. พญ.กษณา รักขมณี	
15.45 – 16.00 น.	Summary	รศ. ดร.นพ.เชิดศักดิ์ ไอรมณีรัตน์	

หมายเหตุ: กำหนดการอาจมีการเปลี่ยนแปลงตามความเหมาะสม

รายชื่อผู้ร่วมอบรม

14 March 2018 Part 1 : Basic concept of assessment

รายชื่อผู้เข้าร่วมโครงการอบรมเชิงปฏิบัติการ เรื่อง "Assessment workshop for clinical teachers" Part 1 : หลักการพื้นฐานของการวัดผล
วันที่ 14 มีนาคม 2561 ณ ห้องบรรยาย 3A01 อาคารศรีสวรินทิรา ชั้น 3A คณะแพทยศาสตร์ศิริราชพยาบาล

กลุ่มที่ 1					
ลำดับ	คำนำหน้า	ชื่อ	สกุล	สังกัด	หน่วยงาน/ภาควิชา
1	อ. นพ.	พิเชษฐ์	วัฒนาประสิทธิ์	โรงพยาบาลยะลา	ออร์โธปิดิกส์
2	อาจารย์	เกวลี	สีหราช	คณะสหเวชศาสตร์ มหาวิทยาลัยพะเยา	สาขาวิชากายภาพบำบัด
3	อาจารย์	ชัชฎาภรณ์	ใจเย็น	คณะสหเวชศาสตร์ มหาวิทยาลัยพะเยา	สาขาวิชากายภาพบำบัด
4	อ. พญ.	อุทัยวรรณ	เล็กยิ่งยง	โรงพยาบาลตำรวจ	กลุ่มงานเวชศาสตร์ฟื้นฟู
5	ผศ.พญ.	จีรดา	พลอยเพชร	คณะแพทยศาสตร์ศิริราชพยาบาล	ภาควิชาเวชศาสตร์ฟื้นฟู
6	อ. นพ.	ทวีศักดิ์	สุตรภาษานนท์	โรงพยาบาลสวรรค์ประชารักษ์	กลุ่มงานเวชกรรมฟื้นฟู
7	อ. พญ.	ลิษา	จุฬาโรจน์มนตรี	คณะแพทยศาสตร์ศิริราชพยาบาล	ภาควิชาจิตวิทยา
8	ผศ. พญ.	สุพรรณิษา	วโรทัย	คณะแพทยศาสตร์ศิริราชพยาบาล	ภาควิชาจิตวิทยา

กลุ่มที่ 2					
ลำดับ	คำนำหน้า	ชื่อ	สกุล	สังกัด	หน่วยงาน/ภาควิชา
1	นางสาว	ศิริประภา	ฤชัย	โรงพยาบาลบำรุงราษฎร์	Clinical Learning
2	อาจารย์	นาฏนภา	อารยะศิลปธร	วิทยาลัยพยาบาลบรมราชชนนีนครพนม ม.นครพนม	กลุ่มวิชาการพยาบาลในคลินิก
3	นางสาว	สุรัชณา	เกษตรเสริมวิริยะ	คณะแพทยศาสตร์ศิริราชพยาบาล	โรงเรียนผู้ช่วยพยาบาล
4	นาง	ลัดดาวัลย์	ปิยะทรงสุทธิ์	คณะแพทยศาสตร์ศิริราชพยาบาล	โรงเรียนผู้ช่วยพยาบาล
5	นางสาว	อทิญา	สียงนอก	โรงพยาบาลตำรวจ	แพทยศาสตร์ศึกษา
6	อ. พญ.	สิริพร	สาสกุล	คณะทันตแพทยศาสตร์ มหาวิทยาลัยเวสเทิร์น	ฝ่ายวิชาการ
7	อ. ดร.	วิรัชพัชร	สกุลสันติพร	วิทยาลัยพยาบาลบรมราชชนนี สรรพสิทธิประสงค์	งานทะเบียนประเมินผล
8	อ. พญ.	พิมพ์ขวัญ	จารุอำพรพรรณ	คณะแพทยศาสตร์ศิริราชพยาบาล	ภาควิชาจักษุวิทยา
9	อ. พญ.	ปณตคม	เง่ายากร	คณะแพทยศาสตร์ศิริราชพยาบาล	ภาควิชาจักษุวิทยา

กลุ่มที่ 3					
ลำดับ	คำนำหน้า	ชื่อ	สกุล	สังกัด	หน่วยงาน/ภาควิชา
1	อาจารย์	อารยา	เชียงใหม่	คณะพยาบาลศาสตร์เกื้อการุณย์ มหาวิทยาลัยนวมินทราชธิราช	ภาควิชาสาธารณสุขศาสตร์และเวชศาสตร์เขตนเมือง
2	อ. ดร.	สรายศ	รุ่งเรืองใจ	คณะสหเวชศาสตร์ มหาวิทยาลัยพะเยา	สาขาเทคนิคการแพทย์
3	ผศ.ดร.	สิทธิชัย	ปัญญาใส	คณะสหเวชศาสตร์ มหาวิทยาลัยพะเยา	สาขาเทคนิคการแพทย์
4	อ. ดร.	ศิริวัฒน์	วรรณตุง	คณะเทคนิคการแพทย์ มหาวิทยาลัยเวสเทิร์น	
5	อาจารย์	รุ่งกาญจน์	สังข์รักษ์	คณะเทคนิคการแพทย์ มหาวิทยาลัยเวสเทิร์น	
6	อ. พญ.	รุจิลักษณ์	โรจน์ธำรงค์	โรงพยาบาลพระนั่งเกล้า	กลุ่มงานรังสีวิทยา
7	ผศ. พญ.	กอบกุล	เมืองสมบูรณ์	คณะแพทยศาสตร์ศิริราชพยาบาล	ภาควิชารังสีวิทยา
8	อ. ดร.	น้ำฟ้า	เสริมแก้ว	สำนักวิชาเภสัชศาสตร์ มหาวิทยาลัยวลัยลักษณ์	
9	ผศ.ดร.	สุทธิพร	ภัทรชยากุล	คณะเภสัชศาสตร์ มหาวิทยาลัยสงขลานครินทร์	ภาควิชาเภสัชกรรมคลินิก

รายชื่อผู้เข้าร่วมโครงการอบรมเชิงปฏิบัติการ เรื่อง "Assessment workshop for clinical teachers" Part 1 : หลักการพื้นฐานของการวัดผล
วันที่ 14 มีนาคม 2561 ณ ห้องบรรยาย 3A01 อาคารศรีสวรินทิรา ชั้น 3A คณะแพทยศาสตร์ศิริราชพยาบาล

กลุ่มที่ 4					
ลำดับ	คำนำหน้า	ชื่อ	สกุล	สังกัด	หน่วยงาน/ภาควิชา
1	อ. พญ.	รุ่งทิพย์	ชัยพรโกติน	โรงพยาบาลพระนั่งเกล้า	กลุ่มงานสูติ-นรีเวชกรรม
2	อ. พญ.	เจติยา	สุรารักษ์	โรงพยาบาลพระนั่งเกล้า	กลุ่มงานสูติ-นรีเวชกรรม
3	อ. นพ.	สรายุทธ์	แท่นนิล	โรงพยาบาลพระนั่งเกล้า	กลุ่มงานสูติ-นรีเวชกรรม
4	อ. พญ.	ชนกานต์	มุสิกวงค์	โรงพยาบาลเจ้าพระยาอภัยภูเบศร	กุมารเวชกรรม
5	อ. นพ.	ปราการ	ตอวิเชียร	คณะแพทยศาสตร์ศิริราชพยาบาล	ภาควิชากุมารเวชศาสตร์
6	อ. พญ.	พัชนี	เบ็ญจสุพัฒน์นันท์	โรงพยาบาลเจริญกรุงประชารักษ์	แผนกกุมารเวชกรรม
7	อ. พญ.	กิตติยา	เศรษฐไกรสิงห์	โรงพยาบาลเจริญกรุงประชารักษ์	แผนกกุมารเวชกรรม
8	อ. พญ.	ปิยนุช	บูรณพร	โรงพยาบาลสงขลา	กลุ่มงานโสต ศอ นาสิก
9	อ. พญ.	นันทน์ภัส	ประคองเดชา	สำนักวิชาแพทยศาสตร์ มหาวิทยาลัยเทคโนโลยีสุรนารี	แผนกโสต ศอ นาสิก

กลุ่มที่ 5					
ลำดับ	คำนำหน้า	ชื่อ	สกุล	สังกัด	หน่วยงาน/ภาควิชา
1	อ. พญ.	จันทน์จิรา	วิมาลา	โรงพยาบาลพุทธชินราช	ภาควิชาวิสัญญีวิทยา
2	ผศ. พญ.	สุกัญญา	เดชอาคม	คณะแพทยศาสตร์ศิริราชพยาบาล	ภาควิชาวิสัญญีวิทยา
3	พ.ต.อ. นพ.	ณัฐพงษ์	กุลสิทธิจินดา	โรงพยาบาลตำรวจ	กลุ่มงานศัลยกรรม
4	อาจารย์	ปิยาภรณ์	เยาวเรศ	คณะพยาบาลศาสตร์ มหาวิทยาลัยมหิดล	ภาควิชาการพยาบาลศัลยศาสตร์
5	อ. นพ.	ชัยธวัช	หาญคุณากร	คณะแพทยศาสตร์ศิริราชพยาบาล	ภาควิชานิติเวชศาสตร์
6	อ. นพ.	กมลพันธ์	ลิ้มเล็ก	คณะแพทยศาสตร์ศิริราชพยาบาล	ภาควิชานิติเวชศาสตร์
7	อ. พญ.	อักษร	พูนิตีพร	โรงพยาบาลขอนแก่น	วิสัญญีวิทยา
8	อ. พญ.	รัชยากร	ลิ้มอภิชาติ	โรงพยาบาลขอนแก่น	วิสัญญีวิทยา

กลุ่มที่ 6					
ลำดับ	คำนำหน้า	ชื่อ	สกุล	สังกัด	หน่วยงาน/ภาควิชา
1	อ. นพ.	เจริญรัตน์	ชัยเจริญธนพร	โรงพยาบาลพุทธชินราช	แผนกอายุรกรรม
2	อ. พญ.	รุจิรา	ลีธนาภรณ์	โรงพยาบาลสงขลา	กลุ่มงานอายุรกรรม
3	พ.ต.ต.หญิง พญ.	ดวงภา	เบญจวงศ์เสถียร	โรงพยาบาลตำรวจ	กลุ่มงานอายุรกรรม
4	อ. นพ.	เอกพันธ์	ครุพงศ์	คณะแพทยศาสตร์ศิริราชพยาบาล	ภาควิชาอายุรศาสตร์
5	อ. พญ.	นารารพร	ประยูรวิวัฒน์	คณะแพทยศาสตร์ศิริราชพยาบาล	ภาควิชาอายุรศาสตร์
6	อ. พญ.	ดำรงศิริ	ไพอุรีย์	โรงพยาบาลบุรีรัมย์	แผนกอุบัติเหตุและฉุกเฉิน
7	พ.ต.ต.หญิง พญ.	พลอยแก้ว	ตัมภ์แสงงาม	โรงพยาบาลตำรวจ	กลุ่มงานเวชศาสตร์ฉุกเฉิน
8	อ. พญ.	สุมาลิน	ชุมคช	โรงพยาบาลสงขลา	กลุ่มงานอุบัติเหตุและฉุกเฉิน

รายชื่อผู้ร่วมอบรม

15 March 2018 Part 2 : MCQ and constructed response items

รายชื่อผู้เข้าร่วมโครงการอบรมเชิงปฏิบัติการ เรื่อง "Assessment workshop for clinical teachers" Part 2 : การสอบข้อเขียน
วันที่ 15 มีนาคม 2561 ณ ห้องบรรยาย 3A01 อาคารศรีสวรินทิรา ชั้น 3A คณะแพทยศาสตร์ศิริราชพยาบาล

กลุ่มที่ 1					
ลำดับ	คำนำหน้า	ชื่อ	สกุล	สังกัด	หน่วยงาน/ภาควิชา
1	อาจารย์	รุ่งกาญจน์	สังข์รักษ์	คณะเทคนิคการแพทย์ มหาวิทยาลัยเวสเทิร์น	
2	อ. ดร.	ถิรวัฒน์	วรรณตุง	คณะเทคนิคการแพทย์ มหาวิทยาลัยเวสเทิร์น	
3	อ. ดร.	สรายศ	รุ่งเรืองใจ	คณะสหเวชศาสตร์ มหาวิทยาลัยพะเยา	สาขาเทคนิคการแพทย์
4	ผศ.ดร.	สิริรัชชัย	ปัญญาใส	คณะสหเวชศาสตร์ มหาวิทยาลัยพะเยา	สาขาเทคนิคการแพทย์
5	อ. ดร.	น้ำฟ้า	เสริมแก้ว	สำนักวิชาเภสัชศาสตร์ มหาวิทยาลัยวลัยลักษณ์	
6	ผศ.ดร.	สุทธิพร	ภัทรชยากุล	คณะเภสัชศาสตร์ มหาวิทยาลัยสงขลานครินทร์	ภาควิชาเภสัชกรรมคลินิก
7	ผศ. พญ.	กอบกุล	เมืองสมบูรณ์	คณะแพทยศาสตร์ศิริราชพยาบาล	ภาควิชารังสีวิทยา
8	อ. พญ.	รุจิลักษณ์	โรจน์อำรงค์	โรงพยาบาลพระนั่งเกล้า	กลุ่มงานรังสีวิทยา

กลุ่มที่ 2					
ลำดับ	คำนำหน้า	ชื่อ	สกุล	สังกัด	หน่วยงาน/ภาควิชา
1	นางสาว	สุรัชนา	เกษตรเสริมวิริยะ	คณะแพทยศาสตร์ศิริราชพยาบาล	โรงเรียนผู้ช่วยพยาบาล
2	นาง	ลัดดาวัลย์	ปิยะทรงสุทธิ์	คณะแพทยศาสตร์ศิริราชพยาบาล	โรงเรียนผู้ช่วยพยาบาล
3	อาจารย์	นาฏนภา	อารยะศิลปธร	วิทยาลัยพยาบาลบรมราชชนนีนครพนม ม.นครพนม	กลุ่มวิชาการพยาบาลในคลินิก
4	นางสาว	ศิริประภา	ฤกษ์ชัย	โรงพยาบาลบำรุงราษฎร์	Clinical Learning
5	นางสาว	อภิญญา	สียานอก	โรงพยาบาลตำรวจ	แพทยศาสตร์ศึกษา
6	อ. ดร.	วีรลพัชร	สกุลสันติพร	วิทยาลัยพยาบาลบรมราชชนนี สรรพสิทธิประสงค์	งานทะเบียนประเมินผล
7	อ. พญ.	พิมพ์ขวัญ	จารุอำพรพรรณ	คณะแพทยศาสตร์ศิริราชพยาบาล	ภาควิชาจักษุวิทยา
8	อ. พญ.	ปณตคม	เง่ายุทธากร	คณะแพทยศาสตร์ศิริราชพยาบาล	ภาควิชาจักษุวิทยา

กลุ่มที่ 3					
ลำดับ	คำนำหน้า	ชื่อ	สกุล	สังกัด	หน่วยงาน/ภาควิชา
1	อาจารย์	เกวลี	สีหราช	คณะสหเวชศาสตร์ มหาวิทยาลัยพะเยา	สาขาวิชากายภาพบำบัด
2	อาจารย์	ชัชฎาภรณ์	ใจเย็น	คณะสหเวชศาสตร์ มหาวิทยาลัยพะเยา	สาขาวิชากายภาพบำบัด
3	อ. นพ.	พิเชษฐ์	วัฒนาประสิทธิ์	โรงพยาบาลยะลา	ออร์โธปิดิกส์
4	ผศ.พญ.	ธีรดา	พลอยเพชร	คณะแพทยศาสตร์ศิริราชพยาบาล	ภาควิชาเวชศาสตร์ฟื้นฟู
5	อ. พญ.	อุทัยวรรณ	เล็กยิ่งยง	โรงพยาบาลตำรวจ	กลุ่มงานเวชศาสตร์ฟื้นฟู
6	อ. นพ.	ทวีศักดิ์	สุดรภาษานนท์	โรงพยาบาลสวรรค์ประชารักษ์	กลุ่มงานเวชกรรมฟื้นฟู
7	ผศ. พญ.	สุเพ็ญญา	วโรทัย	คณะแพทยศาสตร์ศิริราชพยาบาล	ภาควิชาจักษุวิทยา
8	อ. พญ.	ลีนา	จุฬาโรจน์มนตรี	คณะแพทยศาสตร์ศิริราชพยาบาล	ภาควิชาจักษุวิทยา

รายชื่อผู้เข้าร่วมโครงการอบรมเชิงปฏิบัติการ เรื่อง "Assessment workshop for clinical teachers" Part 2 : การสอบข้อเขียน

วันที่ 15 มีนาคม 2561 ณ ห้องบรรยาย 3A01 อาคารศรีสุรินทรา ชั้น 3A คณะแพทยศาสตร์ศิริราชพยาบาล

กลุ่มที่ 4					
ลำดับ	คำนำหน้า	ชื่อ	สกุล	สังกัด	หน่วยงาน/ภาควิชา
1	อ. พญ.	อินทิพร	โมชิตานฤทธิ์	คณะแพทยศาสตร์ มหาวิทยาลัยขอนแก่น	ภาควิชาวิสัญญีวิทยา
2	อ. พญ.	จันทน์จิรา	วิมาลา	โรงพยาบาลพุทธชินราช	ภาควิชาวิสัญญีวิทยา
3	ผศ. พญ.	สุกัญญา	เดชอุดม	คณะแพทยศาสตร์ศิริราชพยาบาล	ภาควิชาวิสัญญีวิทยา
4	อ. พญ.	อักษร	พูนิติพร	โรงพยาบาลขอนแก่น	วิสัญญีวิทยา
5	อ. พญ.	รัชยากร	ลิ้มอภิชาติ	โรงพยาบาลขอนแก่น	วิสัญญีวิทยา
6	อ. พญ.	ดาร์สศิริ	โพอุบรี	โรงพยาบาลบุรีรัมย์	แผนกอุบัติเหตุและฉุกเฉิน
7	อ. พญ.	สุมาลิน	ชุมคช	โรงพยาบาลสงขลา	กลุ่มงานอุบัติเหตุและฉุกเฉิน
8	พ.ต.ท.หญิง พญ.	พลอยแก้ว	ดัมพ์แสงงาม	โรงพยาบาลตำรวจ	กลุ่มงานเวชศาสตร์ฉุกเฉิน
9	พ.ต.อ. นพ.	ณัฐพงษ์	กุลสิทธิจินดา	โรงพยาบาลตำรวจ	กลุ่มงานศัลยกรรม

กลุ่มที่ 5					
ลำดับ	คำนำหน้า	ชื่อ	สกุล	สังกัด	หน่วยงาน/ภาควิชา
1	พ.ต.ท.หญิง พญ.	ดวงนภา	เบญจวงศ์เสถียร	โรงพยาบาลตำรวจ	กลุ่มงานอายุรกรรม
2	อ. พญ.	รุจิรา	สีธนาภรณ์	โรงพยาบาลสงขลา	กลุ่มงานอายุรกรรม
3	อ. นพ.	เจริญรัตน์	ชัยเจริญธร	โรงพยาบาลพุทธชินราช	แผนกอายุรกรรม
4	อ. นพ.	เอกพันธ์	ครุพงศ์	คณะแพทยศาสตร์ศิริราชพยาบาล	ภาควิชาอายุรศาสตร์
5	อ. พญ.	นารารพร	ประยูรวิวัฒน์	คณะแพทยศาสตร์ศิริราชพยาบาล	ภาควิชาอายุรศาสตร์
6	อ. พญ.	ปิยนุช	บุรณพร	โรงพยาบาลสงขลา	กลุ่มงานโสต ศอ นาสิก
7	อ. พญ.	นันทน์ภัส	ประจวบเดชา	สำนักวิชาแพทยศาสตร์ มหาวิทยาลัยเทคโนโลยีสุรนารี	แผนกโสต ศอ นาสิก
8	อ. พญ.	อารีรัตน์	สิริพงศ์พันธ์	โรงพยาบาลมหาวิทยาลัยเทคโนโลยีสุรนารี	แผนกจิตเวช

กลุ่มที่ 6					
ลำดับ	คำนำหน้า	ชื่อ	สกุล	สังกัด	หน่วยงาน/ภาควิชา
1	อ. พญ.	รุ่งทิพย์	ชัยพรโกคิน	โรงพยาบาลพระนั่งเกล้า	กลุ่มงานสูติ-นรีเวชกรรม
2	อ. พญ.	เจติยา	สุรารักษ์	โรงพยาบาลพระนั่งเกล้า	กลุ่มงานสูติ-นรีเวชกรรม
3	อ. นพ.	สรายุทธ์	แท่นนิล	โรงพยาบาลวชิระภูเก็ต	กลุ่มงานสูติ-นรีเวชกรรม
4	อ. นพ.	กมลพันธ์	ลิ้มเล็ก	คณะแพทยศาสตร์ศิริราชพยาบาล	ภาควิชานิติเวชศาสตร์
5	อ. นพ.	ชัยธวัช	หาญคุณากร	คณะแพทยศาสตร์ศิริราชพยาบาล	ภาควิชานิติเวชศาสตร์
6	อ. พญ.	ชนกานต์	มุสิกวงศ์	โรงพยาบาลเจ้าพระยาอภัยภูเบศร	กุมารเวชกรรม
7	อ. นพ.	ปราการ	ตอวิเชียร	คณะแพทยศาสตร์ศิริราชพยาบาล	ภาควิชากุมารเวชศาสตร์
8	อ. พญ.	พัชนี	เป็ญจสุพัฒน์นันท์	โรงพยาบาลเจริญกรุงประชารักษ์	แผนกกุมารเวชกรรม

รายชื่อผู้ร่วมอบรม

16 March 2018 Part 3 : Practical examination

รายชื่อผู้เข้าร่วมโครงการอบรมเชิงปฏิบัติการ เรื่อง "Assessment workshop for clinical teachers" Part 3 : การสอบปฏิบัติ
วันที่ 16 มีนาคม 2561 ณ ห้องบรรยาย 3A01 อาคารศรีสวรินทิรา ชั้น 3A คณะแพทยศาสตร์ศิริราชพยาบาล

กลุ่มที่ 1					
ลำดับ	คำนำหน้า	ชื่อ	สกุล	สังกัด	หน่วยงาน/ภาควิชา
1	อาจารย์	อารยา	เชียงของ	คณะพยาบาลศาสตร์เกื้อการุณย์ มหาวิทยาลัยนวมินทราชิราช	ภาควิชาสาธารณสุขศาสตร์และเวชศาสตร์เขตเมือง
2	นางสาว	สุรัชณา	เกษตรเสริมวิริยะ	คณะแพทยศาสตร์ศิริราชพยาบาล	โรงเรียนผู้ช่วยพยาบาล
3	นาง	ลัดดาวัลย์	ปิยะทรงสุทธิ์	คณะแพทยศาสตร์ศิริราชพยาบาล	โรงเรียนผู้ช่วยพยาบาล
4	อาจารย์	นาฏนภา	อารยะศิลปธร	วิทยาลัยพยาบาลบรมราชชนนีนครพนม ม.นครพนม	กลุ่มวิชาการพยาบาลในคลินิก
5	อาจารย์	มนสมรณ์	วิฑูรเมธา	คณะพยาบาลศาสตร์ มหาวิทยาลัยสวนดุสิต	
6	อาจารย์	ศิริพร	นันทเสนีย์	คณะพยาบาลศาสตร์ มหาวิทยาลัยสวนดุสิต	
7	นางสาว	อพิญญา	สียงนอก	โรงพยาบาลตำรวจ	แพทยศาสตรศึกษา
8	อ. พญ.	อารีรัตน์	สิริพงศ์พันธ์	โรงพยาบาลมหาวิทยาลัยเทคโนโลยีสุรนารี	แผนกจิตเวช

กลุ่มที่ 2					
ลำดับ	คำนำหน้า	ชื่อ	สกุล	สังกัด	หน่วยงาน/ภาควิชา
1	อ. พญ.	อุทัยวรรณ	เล็กยิ่งยง	โรงพยาบาลตำรวจ	กลุ่มงานเวชศาสตร์ฟื้นฟู
2	อ. นพ.	ทวีศักดิ์	สุตรภาษานนท์	โรงพยาบาลสวรรค์ประชารักษ์	กลุ่มงานเวชกรรมฟื้นฟู
3	ผศ.พญ.	ธีรดา	พลอยเพชร	คณะแพทยศาสตร์ศิริราชพยาบาล	ภาควิชาเวชศาสตร์ฟื้นฟู
4	อาจารย์	เกวลี	สีหราช	คณะสหเวชศาสตร์ มหาวิทยาลัยพะเยา	สาขาวิชากายภาพบำบัด
5	อาจารย์	ชัชฎาภรณ์	ใจเย็น	คณะสหเวชศาสตร์ มหาวิทยาลัยพะเยา	สาขาวิชากายภาพบำบัด
6	อ. นพ.	พิเชษฐ์	วัฒนาประสิทธิ์	โรงพยาบาลยะลา	ออโรโรดิคส์
7	อ. พญ.	พิมพ์ขวัญ	จารุอำพรพรรณ	คณะแพทยศาสตร์ศิริราชพยาบาล	ภาควิชาจักษุวิทยา
8	อ. พญ.	ปณตคม	เง่ายุทธกร	คณะแพทยศาสตร์ศิริราชพยาบาล	ภาควิชาจักษุวิทยา

กลุ่มที่ 3					
ลำดับ	คำนำหน้า	ชื่อ	สกุล	สังกัด	หน่วยงาน/ภาควิชา
1	อ. นพ.	เอกพันธ์	ครุพงศ์	คณะแพทยศาสตร์ศิริราชพยาบาล	ภาควิชาอายุรศาสตร์
2	พ.ต.ต.หญิง	พดุงนภา	เบญจวงศ์เสถียร	โรงพยาบาลตำรวจ	กลุ่มงานอายุรกรรม
3	อ. พญ.	รุจิรา	สีธนาภรณ์	โรงพยาบาลสงขลา	กลุ่มงานอายุรกรรม
4	อ. นพ.	เจริญรัตน์	ชัยเจริญธนพร	โรงพยาบาลพุทธชินราช	แผนกอายุรกรรม
5	ผศ. พญ.	สุเพ็ญญา	วโรทัย	คณะแพทยศาสตร์ศิริราชพยาบาล	ภาควิชาตจวิทยา
6	อ. พญ.	ลิษา	จุฬาโรจน์มนตรี	คณะแพทยศาสตร์ศิริราชพยาบาล	ภาควิชาตจวิทยา
7	อ. พญ.	นาราพร	ประยูรวิวัฒน์	คณะแพทยศาสตร์ศิริราชพยาบาล	ภาควิชาอายุรศาสตร์
8	อ. นพ.	ชัยรัช	หาญคุณากร	คณะแพทยศาสตร์ศิริราชพยาบาล	ภาควิชานิติเวชศาสตร์
9	อ. นพ.	กมลพันธ์	ลิ้มเล็ก	คณะแพทยศาสตร์ศิริราชพยาบาล	ภาควิชานิติเวชศาสตร์

รายชื่อผู้เข้าร่วมโครงการอบรมเชิงปฏิบัติการ เรื่อง "Assessment workshop for clinical teachers" Part 3 : การสอบปฏิบัติ
วันที่ 16 มีนาคม 2561 ณ ห้องบรรยาย 3A01 อาคารศรีสวรินทิรา ชั้น 3A คณะแพทยศาสตร์ศิริราชพยาบาล

กลุ่มที่ 4					
ลำดับ	คำนำหน้า	ชื่อ	สกุล	สังกัด	หน่วยงาน/ภาควิชา
1	อ. ดร.	สรายศ	ร่ำเรืองใจ	คณะสหเวชศาสตร์ มหาวิทยาลัยพะเยา	สาขาเทคนิคการแพทย์
2	ผศ.ดร.	สิทธิชัย	ปัญญาใส	คณะสหเวชศาสตร์ มหาวิทยาลัยพะเยา	สาขาเทคนิคการแพทย์
3	ผศ. พญ.	กอบกุล	เมืองสมบูรณ์	คณะแพทยศาสตร์ศิริราชพยาบาล	ภาควิชารังสีวิทยา
4	ผศ.ดร.	สุทธิพร	ภัทรชยากุล	คณะเภสัชศาสตร์ มหาวิทยาลัยสงขลานครินทร์	ภาควิชาเภสัชกรรมคลินิก
5	อ. ดร.	น้ำฟ้า	เสริมแก้ว	สำนักวิชาเภสัชศาสตร์ มหาวิทยาลัยวลัยลักษณ์	
6	ดร.ภญ.	จรรยาพร	พงศ์เวชรักษ์	คณะเภสัชศาสตร์ มหาวิทยาลัยธรรมศาสตร์	สาขาการบริหารทางเภสัชกรรม
7	นางสาว	ศิริประภา	ฤชัย	โรงพยาบาลบำรุงราษฎร์	Clinical Learning
8	อ. ดร.	วิรัชพัชร	สกุลสันติพร	วิทยาลัยพยาบาลบรมราชชนนี สรรพสิทธิประสงค์	งานทะเบียนประเมินผล

กลุ่มที่ 5					
ลำดับ	คำนำหน้า	ชื่อ	สกุล	สังกัด	หน่วยงาน/ภาควิชา
1	อ. พญ.	อินทิพร	โฆษิตานฤฤทธิ์	คณะแพทยศาสตร์ มหาวิทยาลัยนเรศวร	ภาควิชาวิสัญญีวิทยา
2	อ. พญ.	จันทน์จิรา	วิมาลา	โรงพยาบาลพุทธชินราช	ภาควิชาวิสัญญีวิทยา
3	ผศ. พญ.	สุกัญญา	เดชะอคม	คณะแพทยศาสตร์ศิริราชพยาบาล	ภาควิชาวิสัญญีวิทยา
4	อ. พญ.	อักษร	พูนดิพร	โรงพยาบาลขอนแก่น	วิสัญญีวิทยา
5	อ. พญ.	รัชยากร	ลิ้มอภิชาติ	โรงพยาบาลขอนแก่น	วิสัญญีวิทยา
6	พ.ต.อ. นพ.	ณัฐพงษ์	กุลสิทธิจินดา	โรงพยาบาลตำรวจ	กลุ่มงานศัลยกรรม
7	อ. พญ.	สุมาลิน	ชุมชช	โรงพยาบาลสงขลา	กลุ่มงานอุบัติเหตุและฉุกเฉิน
8	อ. พญ.	คำร์ลศิริ	ไพฑูรี	โรงพยาบาลบุรีรัมย์	แผนกอุบัติเหตุและฉุกเฉิน

กลุ่มที่ 6					
ลำดับ	คำนำหน้า	ชื่อ	สกุล	สังกัด	หน่วยงาน/ภาควิชา
1	อ. นพ.	ปราการ	ตอวิเชียร	คณะแพทยศาสตร์ศิริราชพยาบาล	ภาควิชากุมารเวชศาสตร์
2	อ. พญ.	พัชนี	เบ็ญจสุพัฒน์นันท์	โรงพยาบาลเจริญกรุงประชารักษ์	แผนกกุมารเวชกรรม
3	อ. พญ.	กิตติยา	เศรษฐไกรสิงห์	โรงพยาบาลเจริญกรุงประชารักษ์	แผนกกุมารเวชกรรม
4	อ. พญ.	ชนกานต์	มุสิกวงค์	โรงพยาบาลเจ้าพระยาอภัยภูเบศร	กุมารเวชกรรม
5	อ. พญ.	ปิยนุช	บูรณพร	โรงพยาบาลสงขลา	กลุ่มงานโสต ศอ นาสิก
6	อ. พญ.	นันทน์ภัส	ประจวบเดชา	สำนักวิชาแพทยศาสตร์ มหาวิทยาลัยเทคโนโลยีสุรนารี	แผนกโสต ศอ นาสิก
7	อ. พญ.	รุ่งทิพย์	ชัยพรโกดิน	โรงพยาบาลพระนั่งเกล้า	กลุ่มงานสูติ-นรีเวชกรรม
8	อ. นพ.	สรายุทธ์	แท่นนิล	โรงพยาบาลวชิระภูเก็ต	กลุ่มงานสูติ-นรีเวชกรรม

เอกสารประกอบการอบรม



14 March 2018

Part 1 : Basic concept of assessment

รศ.ดร. นพ.เชิดศักดิ์ ไอรมณีรัตน์

หัวข้อ : What is good assessment?

What is Good Assessment?

นพ. เชิดศักดิ์ ไอรมณีรัตน์
ภาควิชาศัลยศาสตร์ คณะแพทยศาสตร์ศิริราชพยาบาล
มหาวิทยาลัยมหิดล

*“Purposeful assessment
drives instruction and affects
learning.”*

Wisconsin's guiding principles for teaching and learning

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.

Assessment

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

Assessment drives instruction.

Outline

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

A Research Study

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

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

The Benefit of Testing

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

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

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

- การทดสอบหัวข้อเดียวกันกับนักศึกษาในระดับชั้นเรียนเดียวกัน ที่จัดสอบกันต่างเวลา ได้คะแนนที่เทียบเคียงกันได้

15

4. Feasibility

ความเป็นไปได้ของการจัดสอบ

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

- Time
- Money
- Expertise
- Administration

5. Educational Effect

- การประเมินผลนั้นกระตุ้นให้ผู้เรียนมีการเรียนรู้ในเรื่องที่ควรเรียนรู้ ... educational benefit

6. Catalytic Effect

- การประเมินผลก่อให้เกิดการนำผลของการสอบไปใช้ให้ feedback เพื่อสร้าง หรือส่งเสริม หรือสนับสนุนการเรียนรู้ของนักศึกษา

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 (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.

7. Acceptability

- ผู้เกี่ยวข้อง (stakeholders) ทั้งหมดเชื่อถือผลการประเมิน

Guidelines for Effective Assessment (1)

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

Guidelines for Effective Assessment (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

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

สิ่งไม่พึงประสงค์ในการสอบ

เชิดศักดิ์ ไอรมนรัตน์

ในบทความนี้ผมจะขอกล่าวถึงสิ่งอันไม่พึงประสงค์ในการสอบ (Validity threats) ที่เราต้องคำนึงถึงในการจัดสอบ ดังที่ได้กล่าวในบทความก่อนหน้านี้แล้วว่า Validity นั้นคือการประเมินคุณค่าของการแปลผลและการนำผลสอบไปใช้ ดังนั้น สิ่งอันไม่พึงประสงค์ในการสอบ หรือ validity threats ก็คือสิ่งใดก็ตามที่เข้ามาบรบกวนการแปลผลสอบ 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 หรือโจทย์คำถามบอกรูปให้นักเรียนตอบถูกโดยไม่ต้องใช้ความรู้ ข้อสอบเก่าที่รั่วไหลออกจากคลังข้อสอบทำให้นักเรียนที่รู้ข้อสอบมาก่อนสามารถตอบได้โดยไม่ต้องคิด
 - นักเรียนที่ทุจริตในการสอบ ลอกข้อสอบของเพื่อน หรือใช้วิธีการอื่นในการได้มาซึ่งคำตอบโดยที่ไม่ได้ใช้ความรู้ในเรื่องที่ทำการสอบ
 - อาจารย์ที่บอกข้อสอบให้นักเรียนในการสอน ทำให้นักเรียนที่ท่องคำตอบเข้าไปสอบ ทำข้อสอบได้โดยไม่ต้องคิด สำหรับการสอบในรูปแบบอื่นที่ต้องใช้กรรมการให้คะแนน เช่น OSCE การสอบข้อสอบบรรยาย หรือการสอบปากเปล่า นั้นจะมีปัจจัยที่เกี่ยวข้องเกี่ยวกับกรรมการผู้ให้คะแนนเข้ามาบรบกวนการแปลผลคะแนนสอบได้ด้วย เช่น
 - ความไม่เสมอภาคของอาจารย์ในเกณฑ์การให้คะแนน นักเรียนที่สอบกับอาจารย์ที่กดคะแนน เสียเปรียบนักเรียนที่สอบกับอาจารย์ที่ใจดี และปล่อยคะแนน
 - ความไม่สม่ำเสมอของอาจารย์ในการให้คะแนน อาจารย์บางท่านมีแนวโน้มจะให้คะแนนต่ำลงในกลุ่มนักเรียนที่สอบตอนท้าย เนื่องด้วยความเหนื่อยล้า ในขณะที่อาจารย์บางท่านมีแนวโน้มจะให้คะแนนสูงขึ้นในตอนท้ายของการสอบ เนื่องจากได้เห็นความสามารถของนักเรียนจำนวนหนึ่งแล้วพบว่าเกณฑ์ที่ตั้งเป้าไว้นั้นสูงเกินความสามารถของนักเรียนส่วนใหญ่จึงปรับเกณฑ์การให้คะแนนให้ง่ายขึ้น ทำให้นักเรียนในกลุ่มหลังได้คะแนนง่ายขึ้น
 - การจำกัดช่วงของคะแนน ที่พบบ่อยคืออาจารย์บางท่านนิยมเดินสายกลาง ไม่ว่านักเรียนจะทำดีมากหรือน้อยเพียงใด ก็มักจะให้คะแนนอยู่ในเกณฑ์ปานกลาง ไม่กล้าให้คะแนน 0 ในรายที่ทำไม่ได้ แต่ก็ไม่กล้าให้คะแนนเต็มในนักเรียนที่ทำได้ดี
- ปัจจัยต่างๆ เหล่านี้ เป็นสิ่งที่ผู้จัดสอบต้องคำนึงถึงเสมอในการจัดสอบและตั้งมาตรฐานเพื่อควบคุมและกำจัดปัจจัยรบกวนเหล่านี้จากการสอบ เพื่อให้ได้ผลการสอบที่มีความเที่ยงตรง เป็นธรรม และสามารถใช้อธิบายความรู้ 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 ไชยมณีรัตน์

ในบทความนี้ผมจะกล่าวถึงการประเมินความแม่นยำของคะแนนสอบ (Reliability) การตรวจสอบความแม่นยำของคะแนนสอบเป็นการตอบคำถามว่า หากทำการสอบซ้ำนักเรียนจะได้คะแนนเท่าเดิมหรือไม่ ในการสอบทั่วไปมักรายงานความแม่นยำของคะแนนสอบด้วยค่า reliability coefficient ซึ่งมีค่าได้ตั้งแต่ 0 ถึง 1 โดยค่ายิ่งสูงบ่งบอกว่าผลสอบมีความน่าเชื่อถือมาก ค่า reliability coefficient = 0 บอกถึงคะแนนสอบที่ขาดความแม่นยำโดยสิ้นเชิง เทียบได้กับการให้คะแนนนักเรียนโดยการสุ่มตัวเลขให้ ส่วนค่า reliability coefficient = 1 บอกถึงคะแนนสอบที่มีความแม่นยำมาก หากให้นักเรียนสอบซ้ำก็จะได้คะแนนเท่าเดิม เพื่อขยายความเข้าใจผมจะกล่าวถึงคุณลักษณะที่สำคัญของ 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 ต้องทำความเข้าใจว่าค่าดังกล่าวบ่งบอกถึงความแม่นยำชนิดใด โดยทั่วไปในการวัดความแม่นยำของคะแนนสอบ multiple-choice examination จากการสอบครั้งเดียว มักเป็นการประเมิน internal consistency reliability ซึ่งบ่งบอกว่าข้อสอบทุกข้อที่ใช้ในการสอบนักเรียนกลุ่มหนึ่งๆทำการวัดความรู้ในเรื่องเดียวกันหรือไม่

3. Reliability เป็นปัจจัยที่สำคัญเพียงปัจจัยหนึ่งในการประเมินคุณค่าของผลสอบ ผลสอบที่ไม่มีความแม่นยำนั้นเป็นผลสอบที่มีคุณค่าต่ำไม่สามารถให้ข้อมูลที่เป็นประโยชน์เกี่ยวกับนักเรียนผู้สอบได้ แต่ผลสอบที่มีความแม่นยำสูงนั้นก็ไม่ว่าจะเป็นผลสอบที่เราสามารถนำไปใช้ประโยชน์ได้เสมอไป จำเป็นต้องพิจารณาปัจจัยร่วมอื่นๆ อีกหลายอย่าง เช่น หากมีนักเรียนทุจริตในการสอบ คะแนนสอบที่ได้ก็อาจมีค่า reliability coefficient สูง แต่ผลสอบนั้นก็จะเป็นผลสอบที่บิดเบือน ไม่สามารถบอกได้ว่านักเรียนที่ได้คะแนนสูงเป็นนักเรียนที่มีความรู้ หรือเป็นนักเรียนที่ไม่มีความรู้แต่ลอกข้อสอบเพื่อน

ประเด็นที่ได้รับความสนใจกันมากคือ ค่า reliability coefficient ต้องสูงแค่ไหนจึงจะเพียงพอที่จะนำผลสอบไปใช้ได้ โดยทั่วไปนั้นจำเป็นต้องพิจารณาควบคู่ไปกับการนำผลสอบไปใช้ หากผลสอบนั้นนำไปใช้ในการตัดสินใจที่สำคัญ เมื่อตัดสินใจไปแล้วผลเป็นที่สุดไม่สามารถเปลี่ยนแปลงได้ และส่งผลยาวนาน โดยเฉพาะการตัดสินใจที่ส่งผลต่อตัวบุคคล มักต้องการคะแนนสอบที่มีค่า reliability coefficient สูงมาก ในทางกลับกัน หากผลสอบนั้นใช้ในการตัดสินใจที่ไม่ค่อยสำคัญ มีผลระยะสั้น และการตัดสินใจผลอาจเปลี่ยนแปลงได้หลังจากการสอบนี้โดยพิจารณาจากการสอบอื่นที่จะจัดตามมาภายหลัง 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 การสอบวุฒิปดฺรผู้เชี่ยวชาญเฉพาะทาง มักต้องให้ reliability coefficient สูงกว่า 0.9

อีกประเด็นหนึ่งที่มีความสำคัญคือ มีปัจจัยใดบ้างที่ส่งผลต่อค่า reliability coefficient สิ่งเหล่านี้มีความสำคัญมากเมื่อเราต้องการอธิบายว่าเหตุใดคะแนนสอบที่ได้จึงไม่แม่นยำ และเราต้องทำอะไรจึงจะทำให้คะแนนสอบมีความแม่นยำมากขึ้น โดยทั่วไปปัจจัยที่สำคัญที่ส่งผลต่อความแม่นยำของคะแนนสอบมีด้วยกัน 4 ปัจจัย คือ

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

รศ.ดร. นพ.เชิดศักดิ์ ไอรณรัตน์

หัวข้อ : How to choose assessment method?

How to Choose Assessment Methods?

เชิดศักดิ์ ไอรณรัตน์
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.

6

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.

8

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

9

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

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

13

OSCE

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

OSCE

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

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

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
- Can be used to assess attitudes and personal development
- Provide vital information for student diagnosis

Davis MH, Ponnamparuma GG. Portfolio assessment. JMME 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, Ponnamparuma GG. Portfolio assessment. JMME 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)
 - Multisource 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, April 2010.

WPBA: Limitations

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

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

The metric of medical education

Validity: on the meaningful interpretation of assessment data*Steven M Downing*

Context All assessments in medical education require evidence of validity to be interpreted meaningfully. In contemporary usage, all validity is construct validity, which requires multiple sources of evidence; construct validity is the whole of validity, but has multiple facets. Five sources – content, response process, internal structure, relationship to other variables and consequences – are noted by the *Standards for Educational and Psychological Testing* as fruitful areas to seek validity evidence.

Purpose The purpose of this article is to discuss construct validity in the context of medical education and to summarize, through example, some typical sources of validity evidence for a written and a performance examination.

Summary Assessments are not valid or invalid; rather, the scores or outcomes of assessments have more or less evidence to support (or refute) a specific interpretation (such as passing or failing a course). Validity is approached as hypothesis and uses theory, logic and the scientific method to collect and assemble data to

support or fail to support the proposed score interpretations, at a given point in time. Data and logic are assembled into arguments – pro and con – for some specific interpretation of assessment data. Examples of types of validity evidence, data and information from each source are discussed in the context of a high-stakes written and performance examination in medical education.

Conclusion All assessments require evidence of the reasonableness of the proposed interpretation, as test data in education have little or no intrinsic meaning. The constructs purported to be measured by our assessments are important to students, faculty, administrators, patients and society and require solid scientific evidence of their meaning.

Keywords Education, Medical, Undergraduate/ *standards, Educational measurement, Reproducibility of results.

Medical Education 2003;37:830–837

Introduction

The purpose of this paper is to discuss validity in the context of assessment in medical education and to present examples of the five types of validity evidence typically sought to support or refute the valid interpretations of assessment data.¹ This essay builds on and expands the older and more traditional view of test validity expressed in the first article in this series² and extends the validity discussion into state-of-the-art 21st century educational measurement.

Validity refers to the evidence presented to support or refute the meaning or interpretation assigned to assessment results. All assessments require validity

evidence and nearly all topics in assessment involve validity in some way. Validity is the *sine qua non* of assessment, as without evidence of validity, assessments in medical education have little or no intrinsic meaning.

Validity is always approached as hypothesis, such that the desired interpretative meaning associated with assessment data is first hypothesized and then data are collected and assembled to support or refute the validity hypothesis. In this conceptualization, assessment data are more or less valid for some very specific purpose, meaning or interpretation, at a given point in time and only for some well-defined population. The assessment itself is never said to be ‘valid’ or ‘invalid’ rather one speaks of the scientifically sound evidence presented to either support or refute the proposed interpretation of assessment scores, at a particular time period in which the validity evidence was collected.

In its contemporary conceptualization,^{1,3–14} validity is a unitary concept, which looks to multiple sources of

Department of Medical Education (MC 591), College of Medicine, University of Illinois at Chicago, Chicago, Illinois, USA

Correspondence: S M Downing, University of Illinois at Chicago, College of Medicine, Department of Medical Education (MC 591), 808 South Wood Street, Chicago, Illinois 60612-7309, USA. Tel.: +1 312 996 6428; Fax: +1 312 413 2048, E-mail: sdowning@uic.edu

Key learning points

Validity is a unitary concept, with construct validity as the whole of validity.

Assessments are not valid or invalid, rather assessment scores have more (or less) validity evidence to support the proposed interpretations.

Validity requires multiple sources of evidence to support or refute meaningful score interpretation.

Validity is always approached as hypothesis.

Validation research uses theory, data and logic to argue for or against specific score interpretations.

evidence. These evidentiary sources are typically logically suggested by the desired types of interpretation or meaning associated with measures. All validity is construct validity in this current framework, described most eloquently by Messick⁸ and embodied in the current *Standards of Educational and Psychological Measurement*.¹ In the past, validity was defined as three separate types: content, criterion and construct, with criterion-related validity usually subdivided into concurrent and predictive depending on the timing of the collection of the criterion data.^{2,15}

Why is construct validity now considered the sole type of validity? The complex answer is found in the philosophy of science⁸ from which, it is posited, there are many complex webs of inter-related inference associated with sampling content in order to make meaningful and reasonable inferences to a domain or larger population of interest. The more straightforward answer is: Nearly all assessments in the social sciences, including medical education, deal with *constructs* – intangible collections of abstract concepts and principles which are inferred from behavior and explained by educational or psychological theory. *Educational achievement* is a construct, usually inferred from performance on assessments such as written tests over some well-defined domain of knowledge, oral examinations over specific problems or cases in medicine, or highly structured standardized patient examinations of history-taking or communication skills.

Educational *ability* or *aptitude* is another example of a familiar construct – a construct that may be even more intangible and abstract than *achievement* because there is less agreement about its meaning among educators and psychologists.¹⁶ Tests that purport to measure educational ability, such as the Medical College Admissions Test (MCAT), which is relied on heavily

in North America for selecting prospective students for medical school admission, must present scientifically sound evidence, from multiple sources, to support the reasonableness of using MCAT test scores as one important selection criterion for admitting students to medical school. An important source of validity evidence for an examination such as the MCAT is likely to be the predictive relationship between test scores and medical school achievement.

Validity requires an evidentiary chain which clearly links the interpretation of the assessment scores or data to a network of theory, hypotheses and logic which are presented to support or refute the reasonableness of the desired interpretations. Validity is never assumed and is an ongoing process of hypothesis generation, data collection and testing, critical evaluation and logical inference. The validity argument^{11,12} relates theory, predicted relationships and empirical evidence in ways to suggest which particular interpretative meanings are reasonable and which are not reasonable for a specific assessment use or application.

In order to meaningfully interpret scores, some assessments, such as achievement tests of cognitive knowledge, may require fairly straightforward content-related evidence of the adequacy of the content tested (in relationship to instructional objectives), statistical evidence of score reproducibility and item statistical quality and evidence to support the defensibility of passing scores or grades. Other types of assessments, such as complex performance examinations, may require both evidence related to content and considerable empirical data demonstrating the statistical relationship between the performance examination and other measures of medical ability, the generalizability of the sampled cases to the population of skills, the reproducibility of the score scales, the adequacy of the standardized patient training and so on.

Some typical sources of validity evidence, depending on the purpose of the assessment and the desired interpretation are: evidence of the content representativeness of the test materials, the reproducibility and generalizability of the scores, the statistical characteristics of the assessment questions or performance prompts, the statistical relationship between and among other measures of the same (or different but related) constructs or traits, evidence of the impact of assessment scores on students and the consistency of pass-fail decisions made from the assessment scores.

The higher the stakes associated with assessments, the greater the requirement for validity evidence from multiple sources, collected on an ongoing basis and continually re-evaluated.¹⁷ The ongoing documentation of validity evidence for a very high-stakes testing

programme, such as a licensure or medical specialty certification examination, may require the allocation of many resources and the contributions of many different professionals with a variety of skills – content specialists, psychometricians and statisticians, test editors and administrators.

In the next section, five major sources of validity evidence are discussed in the contexts of example assessments in medical education.

Sources of evidence for construct validity

According to the *Standards*: ‘Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests’¹ (p. 9). The current *Standards*¹ fully embrace this unitary view of validity, following closely on Messick’s work^{8,9} that considers all validity as construct validity, which is defined as an investigative process through which constructs are carefully defined, data and evidence are gathered and assembled to form an argument either supporting or refuting some very specific interpretation of assessment scores.^{11,12} Historically, the methods of validation and the types of evidence associated with construct validity have their foundations on much earlier work by Cronbach,³⁻⁵ Cronbach and Meehl⁶ and Messick.⁷ The earliest unitary conceptualization of validity as construct validity dates to 1957 in a paper by Loevinger.¹⁸ Kane¹¹⁻¹³ places validity into the context of an interpretive argument, which must be established for each assessment; Kane’s work has provided a useful framework for validity and validation research.

The Standards

The *Standards*¹ discuss five distinct sources of validity evidence (Table 1): content, responses, internal structure, relationship to other variables and consequences. Each source of validity evidence (Table 1) is associated with some examples of the types of data that might be collected to support or refute specific assessment interpretations (validity). Some types of assessment demand a stronger emphasis on one or more sources of evidence as opposed to other sources and not all sources of data or evidence are required for all assessments. For example, a written, objectively scored test covering several weeks of instruction in microbiology, might emphasize content-related evidence, together with some evidence of response quality, internal structure and consequences, but very likely would not seek much or any evidence concerning relationship to other variables. On the other hand, a high-stakes

Table 1 Some sources of validity evidence for proposed score interpretations and examples of some types of evidence

Content	Response process	Internal structure	Relationship to other variables	Consequences
<ul style="list-style-type: none"> • Examination blueprint • Representativeness of test blueprint to achievement domain • Test specifications • Match of item content to test specifications • Representativeness of items to domain • Logical/empirical relationship of content tested to achievement domain • Quality of test questions • Item writer qualifications • Sensitivity review 	<ul style="list-style-type: none"> • Student format familiarity • Quality control of electronic scanning/scoring • Key validation of preliminary scores • Accuracy in combining different formats scores • Quality control/accuracy of final scores/marks/grades • Subscore/subscale analyses: • Accuracy of applying pass-fail decision rules to scores • Quality control of score reporting to students/faculty • Understandable/accurate descriptions/interpretations of scores for students 	<ul style="list-style-type: none"> • Item analysis data: 1. Item difficulty/discrimination (ICCs/TCCs) 2. Inter-item correlations 3. Item-total correlations 4. Item-total correlations • Score scale reliability • Standard errors of measurement (SEM) • Generalizability • Dimensionality • Item factor analysis • Differential Item Functioning (DIF) • Psychometric model 	<ul style="list-style-type: none"> • Correlation with other relevant variables • Convergent correlations - internal/external: 1. Similar tests • Divergent correlations-internal/external 1. Dissimilar measures • Test-criterion correlations • Generalizability of evidence 	<ul style="list-style-type: none"> • Impact of test scores/results on students/society • Consequences on learners/future learning • Positive consequences outweigh unintended negative consequences? • Reasonableness of method of establishing pass-fail (cut) score • Pass-fail consequences: 1. P/F Decision reliability- Classification accuracy 2. Conditional standard error of measurement at pass score (CSEM) • False positives/negatives • Instructional/learner consequences

summative Objective Structured Clinical Examination (OSCE), using standardized patients to portray and rate student performance on an examination that must be passed in order to proceed in the curriculum, might require all of these sources of evidence and many of the data examples noted in Table 1, to support or refute the proposed interpretation of the scores.

Sources of validity evidence for example assessments

Each of the five sources of validity evidence will now be considered, in the context of a written assessment of cognitive knowledge or achievement and a performance examination in medical education. Both example assessments are high-stakes, in that the consequences of passing or failing are very important to students, faculty and, ultimately, patients. The written assessment is a summative comprehensive examination in the basic sciences – a test consisting of 250 multiple-choice questions (MCQs) covering all the pre-clinical instruction in the basic sciences – and a test that must be passed in order to proceed into clinical training. The performance examination is a standardized patient (SP) examination, administered to medical students toward the end of their clinical training, after having completed all of their required clerkship rotations. The purpose of the SP examination is to comprehensively assess graduating medical students' ability to take a history and do a focused physical examination in an ambulatory primary care setting. The SP examination consists of 10 20-minute SP cases, presented by a lay, trained standardized patient who simulates the patient's presenting problem and rates the student's performance at the conclusion of the examination. The SP examination must be passed in order to graduate medical school.

Documentation of these five sources of validity evidence consists of the systematic collection and presentation of information and data to present a convincing argument that it is reasonable and defensible to interpret the assessment scores in accordance with the purpose of the measurement. The scores have little or no intrinsic meaning; thus the evidence presented must convince the skeptic that the assessment scores can reasonably be interpreted in the proposed manner.

Content evidence

For the written assessment, documentation of validity evidence related to the content tested is the most essential. The outline and plan for the test, described by a detailed test blueprint or test specifications, clearly

relates the content tested by the 250 MCQs to the domain of the basic sciences as described by the course learning objectives. The test blueprint is sufficiently detailed to describe subcategories and subclassifications of content and specifies precisely the proportion of test questions in each category and the cognitive level of those questions. The blueprint documentation shows a direct linkage of the questions on the test to the instructional objectives. Independent content experts can evaluate the reasonableness of the test blueprint with respect to the course objectives and the cognitive levels tested. The logical relationship between the content tested by the 250 MCQs and the major instructional objectives and teaching/learning activities of the course should be obvious and demonstrable, especially with respect to the proportionate weighting of test content to the actual emphasis of the basic science courses taught. Further, if most learning objectives were at the application or problem-solving level, most test questions should also be directed to these cognitive levels.

The quality of the test questions is a source of content-related validity evidence. Do the MCQs adhere to the best evidence-based principles of effective item-writing?¹⁹ Are the item-writers qualified as content experts in the disciplines? Are there sufficient numbers of questions to adequately sample the large content domain? Have the test questions been edited for clarity, removing all ambiguities and other common item flaws? Have the test questions been reviewed for cultural sensitivity?

For the SP performance examination, some of the same content issues must be documented and presented as validity evidence. For example, each of the 10 SP cases fits into a detailed content blueprint of ambulatory primary care history and physical examination skills. There is evidence of faculty content-expert agreement that these specific 10 cases are representative of primary care ambulatory cases. Ideally, the content of the 10 clinical cases is related to population demographic data and population data on disease incidence in primary care ambulatory settings. Evidence is documented that expert clinical faculty have created, reviewed and revised the SP cases together with the checklists and ratings scales used by the SPs, while other expert clinicians have reviewed and critically critiqued the SP cases. Exacting specifications detail all the essential clinical information to be portrayed by the SP. Evidence that SP cases have been competently edited and that detailed SP training guidelines and criteria have been prepared, reviewed by faculty experts and implemented by experienced SP trainers are all important sources of content-related validity evidence.

There is documentation that during the time of SP administration, the SP portrayals are monitored closely to ensure that all students experience nearly the same case. Data are presented to show that a different SP, trained on the same case, rates student case performance about the same. Many basic quality-control issues concerning performance examinations contribute to the content-related validity evidence for the assessment.²⁰

Response process

As a source of validity evidence, response process may seem a bit strange or inappropriate. *Response process* is defined here as evidence of data integrity such that all sources of error associated with the test administration are controlled or eliminated to the maximum extent possible. Response process has to do with aspects of assessment such as ensuring the accuracy of all responses to assessment prompts, the quality control of all data flowing from assessments, the appropriateness of the methods used to combine various types of assessment scores into one composite score and the usefulness and the accuracy of the score reports provided to examinees. (Assessment data quality-control issues could also be discussed as content evidence.)

For evidence of response process for the written comprehensive examination, documentation of all practice materials and written information about the test and instructions to students is important. Documentation of all quality-control procedures used to ensure the absolute accuracy of test scores is also an important source of evidence: the final key validation after a preliminary scoring – to ensure the accuracy of the scoring key and eliminate from final scoring any poorly performing test items; a rationale for any combining rules, such as the combining into one final composite score of MCQ, multiple true-false and short-essay question scores.

Other sources of evidence may include documentation and the rationale for the type of scores reported, the method chosen to report scores and the explanations and interpretive materials provided to explain fully the score report and its meaning, together with any materials discussing the proper use and any common misuses of the assessment score data.

For the SP performance examination, many of the same response process sources may be presented as validity evidence. For a performance examination, documentation demonstrating the accuracy of the SP rating is needed and the results of an SP accuracy study is a particularly important source of response process evidence. Basic quality control of the large amounts of data from an SP performance examination is important

to document, together with information on score calculation and reporting methods, their rationale and, particularly, the explanatory materials discussing an appropriate interpretation of the performance-assessment scores (and their limitations).

Documentation of the rationale for using global versus checklist rating scores, for example, may be an important source of response evidence for the SP examination. Or, the empirical evidence and logical rationale for combining a global rating-scale score with checklist item scores to form a composite score may be one very important source of response evidence.

Internal structure

Internal structure, as a source of validity evidence, relates to the statistical or psychometric characteristics of the examination questions or performance prompts, the scale properties – such as reproducibility and generalizability, and the psychometric model used to score and scale the assessment. For instance, scores on test items or sets of items intended to measure the same variable, construct, or content area should be more highly correlated than scores on items intended to measure a different variable, construct, or content area.

Many of the statistical analyses needed to support or refute evidence of the test's internal structure are often carried out as routine quality-control procedures. Analyses such as item analyses – which computes the difficulty (or easiness) of each test question (or performance prompt), the discrimination of each question (a statistical index indicating how well the question separates the high scoring from the low scoring examinees) and a detailed count of the number or proportion of examinees who responded to each option of the test question, are completed. Summary statistics are usually computed, showing the overall difficulty (or easiness) of the total test scale, the average discrimination and the internal consistency reliability of the test.

Reliability is an important aspect of an assessment's validity evidence. Reliability refers to the reproducibility of the scores on the assessment; high score reliability indicates that if the test were to be repeated over time, examinees would receive about the same scores on retesting as they received the first time. Unless assessment scores are reliable and reproducible (as in an experiment) it is nearly impossible to interpret the meaning of those scores – thus, validity evidence is lacking.

There are many different types of reliability, appropriate to various uses of assessment scores. In both example assessments described above, in which the

stakes are high and a passing score has been established, the reproducibility of the pass-fail decision is a very important source of validity evidence. That is, analogous to score reliability, if the ultimate outcome of the assessment (passing or failing) can not be reproduced at some high level of certainty, the meaningful interpretation of the test scores is questionable and validity evidence is compromised.

For performance examinations, such as the SP example, a very specialized type of reliability, derived from generalizability theory (GT)^{21,22} is an essential component of the internal structure aspect of validity evidence. GT is concerned with how well the specific samples of behaviour (SP cases) can be generalized to the population or universe of behaviours. GT is also a useful tool for estimating the various sources of contributed error in the SP exam, such as error due to the SP raters, error due to the cases (case specificity), and error associated with examinees. As rater error and case specificity are major threats to meaningful interpretation of SP scores, GT analyses are important sources of validity evidence for most performance assessments such as OSCEs, SP exams and clinical performance examinations.

For some assessment applications, in which sophisticated statistical measurement models like Item Response Theory (IRT) models^{23,24} the measurement model itself is evidence of the internal structure aspect of construct validity. In IRT applications, which might be used for tests such as the comprehensive written examination example, the factor structure, item-inter-correlation structure and other internal structural characteristics all contribute to validity evidence.

Issues of bias and fairness also pertain to internal test structure and are important sources of validity evidence. All assessments, presented to heterogeneous groups of examinees, have the potential of validity threats from statistical bias. Bias analyses, such as differential item functioning (DIF)^{25,26} analyses and the sensitivity review of item and performance prompts are sources of internal structure validity evidence. Documentation of the absence of statistical test bias permits the desired score interpretation and therefore adds to the validity evidence of the assessment.

Relationship to other variables

This familiar source of validity evidence is statistical and correlational. The correlation or relationship of assessment scores to a criterion measure's scores is a typical design for a 'validity study', in which some newer (or simpler or shorter) measure is 'validated'

against an existing, older measure with well known characteristics.

This source of validity evidence embodies all the richness and complexity of the contemporary theory of validity in that the relationship to other variables aspect seeks both confirmatory and counter-confirmatory evidence. For example, it may be important to collect correlational validity evidence which shows a strong positive correlation with some other measure of the same achievement or ability and evidence indicating no correlation (or a strong negative correlation) with some other assessment that is hypothesized to be a measure of some completely different achievement or ability.

The concept of convergence and divergence of validity evidence is best exemplified in the classic research design first described by Campbell and Fiske.²⁷ In this 'multitrait multimethod' design, different measures of the same trait (achievement, ability, performance) are correlated with different measures of the same trait. The resulting pattern of correlation coefficients may show the convergence and divergence of the different assessment methods on measures of the same and different abilities or proficiencies.

In the written comprehensive examination example, it may be important to document the correlation of total and subscale scores with achievement examinations administered during the basic science courses. One could hypothesize that a subscale score for biochemistry on the comprehensive examination would correlate more highly with biochemistry course test scores than with behavioural science course scores. Additionally, the correlation of the written examination scores with the SP final examination may show a low (or no) correlation, indicating that these assessment methods measure some unique achievement, while the correlation of the SP scores with other performance examination scores during the students' clinical training may be high and positive.

As with all research, issues of the generalizability of the results of these studies and the limitations of data interpretation pertain. Interpretation of correlation coefficients, as validity coefficients, may be limited due to the design of the study, systematic bias introduced by missing data from either the test or the criterion or both and statistical issues such as restriction of the range of scores (lack of variance).

Consequences

This aspect of validity evidence may be the most controversial, although it is solidly embodied in the current *Standards*.¹ The consequential aspect of validity refers to the impact on examinees from the assessment

scores, decisions and outcomes, and the impact of assessments on teaching and learning. The consequences of assessments on examinees, faculty, patients and society can be great and these consequences can be positive or negative, intended or unintended.

High-stakes examinations abound in North America, especially in medicine and medical education. Extremely high-stakes assessments are often mandated as the final, summative hurdle in professional education. For example, the United States Medical Licensure Examination (USMLE) sequence, sponsored by the National Board of Medical Examiners (NBME), consists of three separate examinations (Steps 1, 2 and 3) which must be passed in order to be licensed as a physician. The consequences of failing any of these examinations is enormous, in that medical education is interrupted in a costly manner or the examinee is not permitted to enter graduate medical education or practice medicine. Likewise, most medical specialty boards in the USA mandate passing a high-stakes certification examination in the specialty or subspecialty, after meeting all eligibility requirements of postgraduate training. The consequences of passing or failing these types of examinations are great, as false positives (passing candidates who should fail) may do harm to patients through the lack of a physician's specialized knowledge or skill and false negatives (failing candidates who should pass) may unjustly harm individual candidates who have invested a great deal of time and resources in graduate medical education.

Thus, consequential validity is one very important aspect of the construct validity argument. Evidence related to consequences of testing and its outcomes is presented to suggest that no harm comes directly from the assessment or, at the very least, more good than harm arises from the assessment. Much of this evidence is more subjective than other sources.

In both example assessments, sources of consequential validity may relate to issues such as passing rates (the proportion who pass), the subjectively judged appropriateness of these passing rates, data comparing the passing rates of each of these examinations to other comprehensive examinations such as the USMLE Step 1 and so on. Evaluations of false positive and false negative outcomes relate to the consequences of these two high-stakes examinations.

The passing score (or grade levels) and the process used to determine the cut scores, the statistical properties of the passing scores, and so on all relate to the consequential aspects of validity.²⁸ Documentation of the method used to establish a pass-fail score is key consequential evidence, as is the rationale for the

selection of a particular passing score method. The psychometric characteristics of the passing score judgments and the qualification and number of expert judges – all may be important to document and present as evidence of consequential validity.

Other psychometric quality indicators concerning the passing score and its consequences (for both example assessments) include a formal, statistical estimation of the pass-fail decision reliability or classification accuracy²⁹ and some estimation of the standard error of measurement at the cut score.³⁰

Equally important consequences of assessment methods on instruction and learning have been discussed by Newble and Jaeger.³¹ The methods and strategies selected to evaluate students can have a profound impact on what is taught, how and exactly what students learn, how this learning is used and retained (or not) and how students view and value the educational process.

Threats to validity

The next essay in this series will discuss the many threats to the meaningful interpretation of assessment scores and suggest methods to control these validity threats.

Conclusion

This paper has reviewed the contemporary meaning of validity, a unitary concept with multiple facets, which considers construct validity as the whole of validity. Validity evidence refers to the data and information collected in order to assign meaningful interpretation to assessment scores or outcomes, which were designed for a specific purpose and at one specific point in time. Validity always refers to score interpretations and never to the assessment itself. The process of validation is closely aligned with the scientific method of theory development, hypothesis generation, data collection for the purpose of hypothesis testing and forming conclusions concerning the accuracy of the desired score interpretations. Validity refers to the impartial, scientific collection of data, from multiple sources, to provide more or less support for the validity hypothesis and relates to logical arguments, based on theory and data, which are formed to assign meaningful interpretations to assessment data.

This paper discussed five typical sources of validity evidence – content, response process, internal structure, relationship to other variables and consequences – in the context of two example assessments in medical education.

Acknowledgements

The author wishes to thank Michael T Kane, PhD, for his critical review of this manuscript.

Funding

There was no external funding for this project.

References

- American Educational Research Association, American Psychological Association, National Council on Measurement in Education. *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association 1999.
- Crossley J, Humphris G, Jolly B. Assessing health professionals. *Med Educ* 2002;**36**:800-4.
- Cronbach LJ. Test validation. In: *Educational Measurement*, 2nd edn. Ed: Thorndike RL. Washington, DC: American Council on Education 1971:443-507.
- Cronbach LJ. Five perspectives on validity argument. In: *Test Validity*. Eds: Wainer H, Braun H. Hillsdale, NJ: Lawrence Erlbaum 1988:3-17.
- Cronbach LJ. Construct validation after 30 years. In: *Intelligence: Measurement, Theory, and Public Policy*. Ed: Linn RE. Urbana, IL: University of Illinois Press 1989:147-71.
- Cronbach LJ, Meehl PE. Construct validity in psychological tests. *Psychol Bull* 1955;**52**:281-302.
- Messick S. The psychology of educational measurement. *J Educ Measure* 1984;**21**:215-37.
- Messick S. Validity. In: *Educational Measurement*, 3rd edn. Ed: Linn RL. New York: American Council on Education and Macmillan 1989:13-104.
- Messick S. Validity of psychological assessment: validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *Am Psychologist* 1995;**50**:741-9.
- Messick S. Standards of validity and the validity of standards in performance assessment. *Educ Measure Issues Prac* 1995;**14**:5-8.
- Kane MT. An argument-based approach to validation. *Psychol Bull* 1992;**112**:527-35.
- Kane MT. Validating interpretive arguments for licensure and certification examinations. *Evaluation Health Professions* 1994;**17**:133-59.
- Kane MT. Current concerns in validity theory. *J Educ Measure* 2001;**38**:319-42.
- Kane MT, Crooks TJ, Cohen AS. Validating measures of performance. *Educ Measure Issues Prac* 1999;**18**:5-17.
- Cureton EE. Validity. In: *Educational Measurement*. Ed: Lingquist EF. Washington, DC: American Council on Education 1951:621-94.
- Lohman DF. Teaching and testing to develop fluid abilities. *Educational Reser* 1993;**22**:12-23.
- Linn RL. Validation of the uses and interpretations of results of state assessment and accountability systems. In: *Large-Scale Assessment Programs for All Students: Development, Implementation, and Analysis*. Eds: Tindal G, Haladyna T. Mahwah, NJ: Lawrence Erlbaum 2002.
- Loevinger J. Objective tests as instruments of psychological theory. *Psychol Reports, Monograph* 1957;**3** (Suppl.) 635-94.
- Haladyna TM, Downing SM, Rodriguez MC. A review of multiple-choice item-writing guidelines for classroom assessment. *Appl Measure Educ* 2002;**15**:309-34.
- Boulet JR, McKinley DW, Whelan GP, Hambleton RK. Quality assurance methods for performance-based assessments. *Adv Health Sci Educ* 2003;**8**:27-47.
- Brennan RL. *Generalizability Theory*. New York: Springer-Verlag 2001.
- Crossley J, Davies H, Humphris G, Jolly B. Generalisability; a key to unlock professional assessment. *Med Educ* 2002;**36**:972-8.
- Van der Linden WJ, Hambleton RK. Item response theory. Brief history, common models, and extensions. In: *Handbook of Modern Item Response Theory*. Eds: van der Linden WJ, Hambleton RK. New York: Springer-Verlag 1997:1-28.
- Downing SM. Item response theory: Applications of modern test theory in medical education. *Med Educ* 2003;**37**:1-7.
- Holland PW, Wainer H, eds. *Differential Item Functioning*. Mahwah, NJ: Lawrence Erlbaum 1993.
- Penfield RD, Lam RCM. Assessing differential item functioning in performance assessment: review and recommendations. *Educ Measure Issues Prac* 2000;**19**:5-15.
- Campbell DT, Fiske DW. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psych Bull* 1959;**56**:81-105.
- Norcini JJ. Setting standards on educational tests. *Med Educ* 2003;**37**:464-9.
- Subkoviak MJ. A practitioner's guide to computation and interpretation of reliability indices for mastery tests. *J Educ Measure* 1988;**25**:47-55.
- Angoff WH. Scales, norms, and equivalent scores. In: *Educational Measurement*, 2nd edn. Ed: Thorndike RL. Washington, DC: American Council on Education 1971:508-600.
- Newble DI, Jaeger K. The effects of assessment and examinations on the learning of medical students. *Med Educ* 1983;**17**:165-71.

Received 29 May 2003; accepted for publication 3 June 2003

รศ.ดร. นพ.เชิดศักดิ์ ไอรณรัตน์

หัวข้อ : Reliability

Reliability

เชิดศักดิ์ ไอรณรัตน์

ภาควิชาศัลยศาสตร์ คณะแพทยศาสตร์ศิริราชพยาบาล

มหาวิทยาลัย มหิดล

Reliability

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

Outline

- Classical test theory
 - Reliability of standard written exam
 - Reliability of mastery tests
 - Reliability of performance assessment
- Generalizability theory

Classical Test Theory

$$T = O + e$$

T = True score

O = Observe score

e = Error

Error

- Systematic error
- Random error

Random Error

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

Reliability of Test Scores

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

Reliability of Written Tests

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

Internal Consistency Reliability

- Split-half method

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

r = Reliability for half test

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

KR-20

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

n = number of items

Var = Variance of the whole test

p = Proportion of people passing the item

q = Proportion of people failing the item

How Much is Enough?

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

Improving Reliability

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

11

12

Spearman-Brown Formula

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

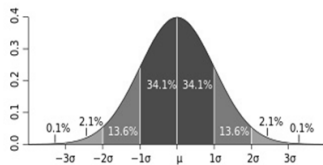
- r_k = Reliability of a test "k" times long
- r_1 = Reliability of the original test
- k = factor by which test length is changed

Example

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

True Score Theory

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



SEM

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

SD = standard deviation
r = internal consistency reliability

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

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

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

Reliability of Mastery Tests

- Consistency of decisions on two test forms

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

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

Performance Assessment

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

Generalizability Theory (GT)

- Multi-faceted assessment
 - Assessment of performance of residents in multiple rotations
 - Sources of error
 - Residents
 - Rotations
 - Items
 - Interaction of these sources

Generalizability

24 residents x 13 rotations (raters) x 11 items

$$X_{PRI} = \mu + v_P + v_R + v_I + v_{PI} + v_{PR} + v_{RI} + v_{PRI}$$

X_{PRI} = An observed score of a resident P given by rater R on item I

μ = Grand mean of the population

v_P, v_R, v_I = Effect of a resident P , rater R , item I

v_{PI} = Effect of resident P crossed with item I

v_{PR} = Effect of resident P crossed with rater R

v_{RI} = Effect of rater R crossed with item I

v_{PRI} = Effect of resident P crossed with rater R and crossed with item I

21

Generalizability

- The nature of decision: Absolute
- Absolute error variance: σ^2_{Δ}

$$\sigma^2_{\Delta} = \sigma^2_R + \sigma^2_I + \sigma^2_{PI} + \sigma^2_{PR} + \sigma^2_{RI} + \sigma^2_{PRI}$$

- Variance component study revealed:

$\sigma^2_R = 0.024$	$\sigma^2_I = 0.028$
$\sigma^2_{PI} = 0.000$	$\sigma^2_{PR} = 0.134$
$\sigma^2_{RI} = 0.021$	$\sigma^2_{PRI} = 0.091$

22

Summary

- Classical test theory
 - Reliability of standard written exam
 - Reliability of mastery tests
 - Reliability of performance assessment
- Generalizability theory

ผศ. พญ.กชญา รักขมณี

หัวข้อ : Standard setting

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

วิธีการตั้งเกณฑ์สอบผ่าน (passing standard) (ตอนที่ 1)

เชิดศักดิ์ ไอรอมณีรัตน์

เกณฑ์สอบผ่าน (passing standard) คือคะแนนสอบที่น้อยที่สุดที่คณาจารย์ยินยอมให้นักเรียนสามารถสอบผ่าน นักเรียนที่สอบได้คะแนนน้อยกว่าเกณฑ์สอบผ่านจะถูกตัดสินว่าสอบตก การตั้งเกณฑ์สอบผ่านจัดเป็นขั้นตอนที่มีความสำคัญมาก ในการจัดสอบ แต่กลับไม่ได้รับความสนใจเท่าที่ควรในการวัดผลทางแพทยศาสตรศึกษาจำนวนมาก ในบทความนี้ผมขอเสนอ เกร็ดความรู้เกี่ยวกับวิธีการตั้งเกณฑ์สอบผ่าน ผมหวังว่าอาจารย์ผู้อ่านจะสามารถนำเกร็ดความรู้นี้ไปใช้พัฒนาคุณภาพของการตั้ง เกณฑ์สอบผ่านได้ไม่มากก็น้อยครับ

เกณฑ์สอบผ่านในทางแพทยศาสตรศึกษาจัดว่ามีความสำคัญมากเนื่องจากเกณฑ์สอบผ่านเป็นการแสดงออกถึง มาตรฐานของวิชาชีพที่อาจารย์ยอมรับ เกณฑ์สอบผ่านที่ดีต้องได้รับการตั้งขึ้นโดยใช้ดุลยพินิจของคณาจารย์ผู้เชี่ยวชาญใน สาขาวิชานั้นๆ เพื่อรักษามาตรฐานการประกอบวิชาชีพเพื่อให้สังคมได้รับบริการทางการแพทย์ที่มีคุณภาพ ในขณะที่เดียวกันกับให้ ความเป็นธรรมกับนักเรียนผู้สอบ เนื่องจากเกณฑ์สอบผ่านเป็นการแสดงออกถึง "ความยอมรับได้" ในดุลยพินิจของคณาจารย์ ผู้เชี่ยวชาญ จึงไม่มีวิธีการทางวิทยาศาสตร์ใดที่จะตัดสินว่าเกณฑ์ที่ตั้งขึ้นนั้นถูกหรือผิด สิ่งที่สำคัญที่สุดในการตั้งเกณฑ์สอบผ่าน หาใช่ "ตัวเลข" คะแนนที่จะใช้ตัดสินได้ตก หากแต่เป็น "กระบวนการ" ให้ได้มาซึ่งเกณฑ์ดังกล่าว เกณฑ์สอบผ่านที่ตั้งขึ้นโดยใช้ อาจารย์ 1 ท่านเลือกตัวเลข 1 ตัวเลขขึ้นมาโดยไม่ได้พิจารณาถึงข้อสอบหรือนักเรียนผู้สอบ เป็นวิธีการตั้งเกณฑ์ที่ล่อแหลมต่อการ ถูกวิจารณ์ (และประท้วง) โดยผู้ที่ไม่พอใจในผลสอบ วิธีการตั้งเกณฑ์สอบผ่านที่ดีนั้นต้องมีหลักการและเหตุผลประกอบ และผ่าน ดุลยพินิจของคณาจารย์ จำนวนของอาจารย์ผู้เชี่ยวชาญที่ต้องใช้ในการตั้งเกณฑ์นั้นขึ้นกับความสำคัญของการสอบนั้นๆ ในการ สอบที่มีความสำคัญสูงเช่นการสอบวุฒิบัตรแพทย์ผู้เชี่ยวชาญ แนะนำให้ใช้คณาจารย์อย่างน้อย 6 – 8 ท่าน ในการตั้งเกณฑ์ แต่ หากเป็นการสอบเล็กๆ เช่น การทดสอบหลังการสอนกลุ่มย่อย อาจใช้อาจารย์เพียง 1 ท่านก็ได้

การตั้งเกณฑ์สอบผ่านมี 2 ชนิดคือ การตัดสินแบบอิงเกณฑ์ (criterion-referenced standard, absolute standard) และการตัดสินแบบอิงกลุ่ม (norm-referenced standard, relative standard) การตัดสินแบบอิงเกณฑ์ เป็นการตั้งว่า คะแนนเท่าไร จึงจัดว่าผ่านการสอบ ในทางตรงข้าม การตัดสินแบบอิงกลุ่ม เป็นการตั้งว่า จะให้ นักเรียนจำนวนเท่าไร ผ่านการสอบ การตัดสินแบบอิงเกณฑ์นั้นเหมาะกับการสอบเพื่อวัดว่าผู้สอบมีความรู้ความสามารถในด้านใดด้านหนึ่งเพียงพอหรือไม่ ส่วนการสอบแบบอิงกลุ่มนั้นเหมาะสำหรับการสอบแข่งขันเพื่อเข้าศึกษาต่อ หรือ ทำงาน ในสถาบันที่มีตำแหน่งที่จะรับได้จำกัด เช่น การสอบเข้าโรงเรียนแพทย์ หรือ การสอบคัดเลือกแพทย์ประจำบ้าน การสอบส่วนใหญ่ในทางแพทยศาสตรศึกษานั้นเหมาะกับการตัดสินแบบอิงเกณฑ์ หากผู้สอบทุกคนมีความสามารถเพียงพอก็ไม่จำเป็นต้องมีผู้สอบตก การใช้การตัดสินแบบอิงกลุ่มเพื่อวัดความรู้ ความสามารถในสถานการณ์อื่นนอกจากการสอบคัดเลือกนั้นเป็นการส่งเสริมให้นักเรียนเกิดความแข่งขันกัน (แทนที่จะช่วยกัน เรียน) โดยไม่จำเป็น

เนื่องจากการสอบทางแพทยศาสตรศึกษาแทบทั้งหมดเหมาะกับการตั้งเกณฑ์สอบผ่านแบบอิงเกณฑ์ ผมจะขอขยาย ความวิธีการตั้งเกณฑ์สอบผ่านแบบอิงเกณฑ์ที่สำคัญและใช้บ่อย 2 วิธีใหญ่ๆ คือ 1. การตั้งเกณฑ์โดยพิจารณาข้อสอบ และ 2. การตั้งเกณฑ์โดยพิจารณาจากผู้สอบ ในบทความตอนต่อไปครับ

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

วิธีการตั้งเกณฑ์สอบผ่าน (passing standard) (ตอนที่ 2)

เชิดศักดิ์ ไชยมณีรัตน์

ในบทความนี้ผมจะขอแนะนำวิธีการตั้งเกณฑ์สอบผ่านโดยพิจารณาตัวข้อสอบที่ใช้สอบ วิธีการตั้งเกณฑ์ผ่านแบบนี้เหมาะสำหรับการสอบ multiple-choice questions ซึ่งอาจารย์ผู้ตั้งเกณฑ์ผ่านสามารถประเมินความน่าจะเป็นของการตอบข้อสอบแต่ละข้อถูกต้อง การตั้งเกณฑ์ผ่านแบบนี้ประกอบด้วย 3 ขั้นตอนหลักคือ

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

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

3. ทำการคิดเกณฑ์สอบผ่านสำหรับข้อสอบนั้น
 - 3.1. Angoff's method เกณฑ์ผ่านคือผลรวมของความน่าจะเป็นของการตอบข้อสอบแต่ละข้อถูก

Item	1	2	3	4	5	Passing score
Probability	0.95	0.85	0.30	0.40	0.70	3.20
 - 3.2. Ebel's method เกณฑ์ผ่านคือผลรวมของ (จำนวนข้อสอบในแต่ละกลุ่ม x ความน่าจะเป็นของการตอบข้อสอบถูกสำหรับข้อสอบในกลุ่มนั้น) จากข้อสอบทั้ง 12 กลุ่ม

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

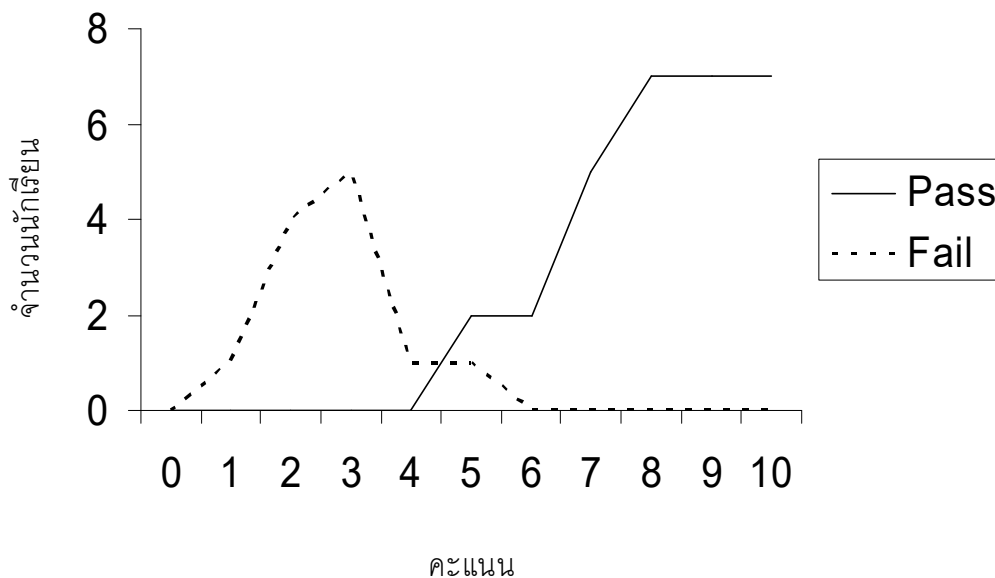
Iramaneerat C. Passing standard: Part III [Thai]. Medical Education Pamphlet 2006; 2(3): 1.

วิธีการตั้งเกณฑ์สอบผ่าน (passing standard) (ตอนที่ 3)

เชิดศักดิ์ ไชรมณีรัตน์

ในบทความนี้ผมจะขอแนะนำวิธีการตั้งเกณฑ์สอบผ่านโดยพิจารณาจากผู้สอบ วิธีการตั้งเกณฑ์ผ่านแบบนี้เหมาะสำหรับการสอบวัดทักษะ การสอบสัมภาษณ์ หรือการประเมินการปฏิบัติงาน ซึ่งมักตัดสินการสอบผ่านโดยดูจากความสามารถของผู้สอบโดยรวมได้ง่ายกว่าดูจากคะแนนที่ได้ในหัวข้อประเมินแต่ละข้อ วิธีการตั้งเกณฑ์ผ่านลักษณะนี้ที่ใช้อยู่มีด้วยกัน 2 วิธีคือ

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



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

หัวข้อ : Grading

GRADING

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

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

Thomas Guskey

1

2

Objectives

- เมื่อสิ้นสุดการบรรยายแล้ว ผู้เข้าอบรมสามารถ
 - อธิบายถึงข้อดี ข้อดีของการตัดสินผลการเรียนแบบอิงเกณฑ์ และอิงกลุ่มได้
 - เลือกใช้วิธีการตัดเกรดที่เหมาะสมกับบริบทของสถาบันในการตัดสินผลการศึกษานักศึกษา
 - บอกถึงแนวทางที่จะพัฒนาคุณภาพการตัดสินผลการศึกษานักศึกษาในสถาบันและหน่วยงานของตนได้อย่างเหมาะสม

3

Outline

- What is grading?
- Why do we grade our students?
- How can we grade our students?
- How should we combine test scores?
- What does research tell us about grading?
- An example of grading criteria set up

4

What is grading?

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

5

Why do we grade our students?

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

6

How can we grade our students?

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

7

Who should receive an A?

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

8

Absolute Grading

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

9

Relative Grading

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

10

Standardization of Scores

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

Z = standard score

X = raw score

M = mean

SD = standard deviation

11

What does research tell us about grading?

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

12

Guidelines for Fair Grading

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

13

Summary

- What is grading?
- Why do we grade our students?
- How can we grade our students?
- How should we combine test scores?
- What does research tell us about grading?
- An example of grading criteria set up

14

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

John F. Kennedy

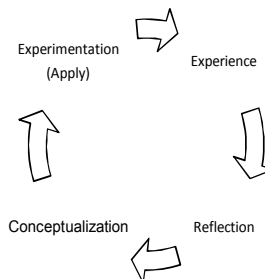
15

หัวข้อ : Summary

Summary

นพ. เชิดศักดิ์ ไอรมณีรัตน์
 ภาควิชาศัลยศาสตร์ คณะแพทยศาสตร์ศิริราชพยาบาล
 มหาวิทยาลัยมหิดล

Experiential Learning Theory



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

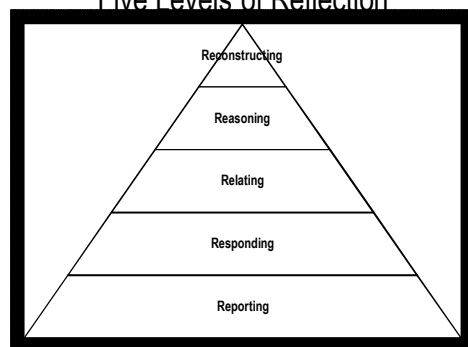
A complex and deliberate process of thinking about and interpreting experience in order to learn from it.

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

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

Atkins and Murphy (1995)

Five Levels of Reflection



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

Summary of the Workshop

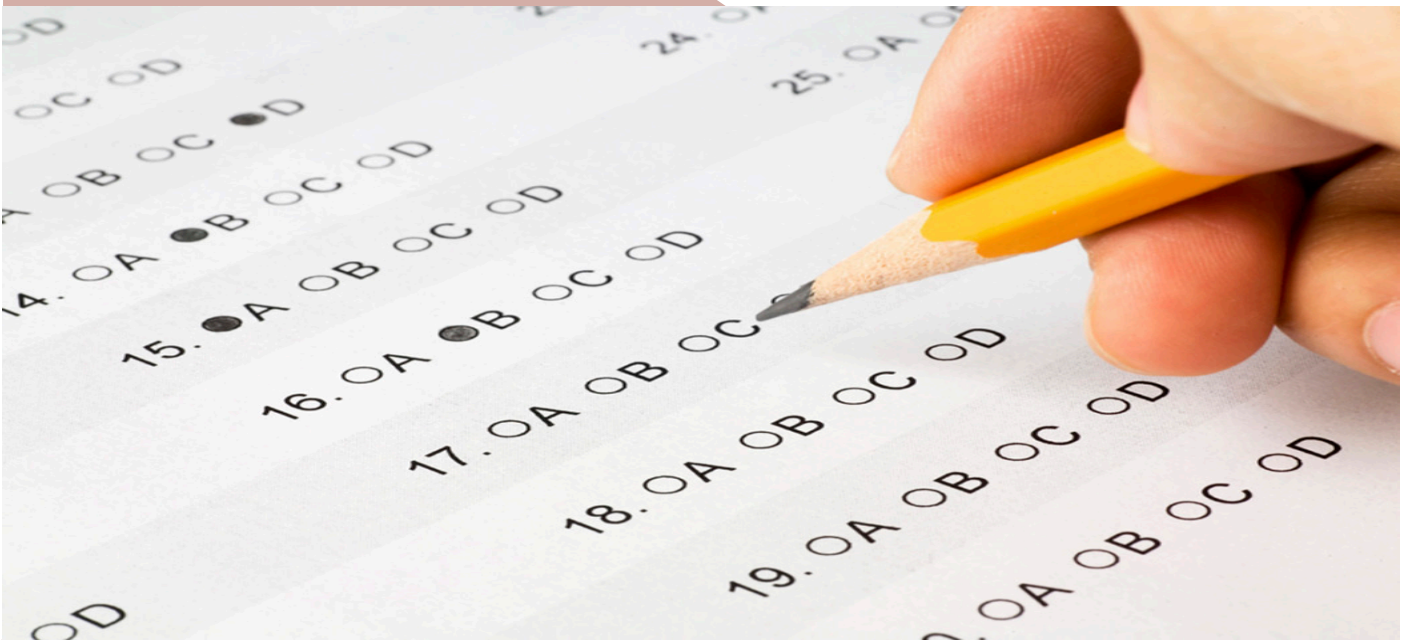
- Morning
 - What is good assessment?
 - How to choose assessment methods?
 - Validity
 - Reliability
- Afternoon
 - Standard setting
 - Grading

5



Shee.si.mahidol.ac.th

เอกสารประกอบการอบรม



15 March 2018

Part 2 : Multiple-choice questions and constructed response items

ศ. พญ.บุญมี สกาศิตยวงศ์

หัวข้อ : Multiple-choice questions item development

MAHIDOL UNIVERSITY
 MAHIDOL UNIVERSITY
 Wisdom of the Land

ASSESSMENT IN MEDICAL EDUCATION : MCQ

Boonmee Sathapatayavongs, MD
 Prof. Channiwat Kasemsunt
 Faculty of Medicine Ramathibodi Hospital
 March 15, 2018

Assessment Methods According to Miller's Competency Pyramid

Does
 Portfolio, structured report, log book
 Rating scales, checklists, scoring rubrics
 • 360° global rating

Shows how
 • OSCE, simulation & models, long case
 • SOE, case discussion
 • Chart audit

Knows how
 • MEQ, computer-based exam
 • EMI, MCQ, CRQ

Knows
 • MCQ

ACGME 2002
 Magery Davis 2003

IS MCQ A GOOD TEST ?

IS MCQ A GOOD TEST ?

- Validity
- Reliability
- Objectivity
- Feasibility / practicability
- Educational effect
- Catalytic effect
- Acceptability

Learning Objectives and MCQs

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

Learning Objectives Predicts student performance } MCQ Assesses student performance

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

1. a wide pulse pressure
2. increased cardiac output
3. dilatation of the left ventricle
4. pulmonary hypertension

-----> **Stem**

-----> **Correct Answer**

-----> **Distracter**

Multiple True- False (Simple, X-Type)

Serum electrolytes include ... Stem

Options		Answers	
		Y	N
A.	sodium	<input checked="" type="checkbox"/>	<input type="checkbox"/>
B.	potassium	<input checked="" type="checkbox"/>	<input type="checkbox"/>
C.	albumin	<input type="checkbox"/>	<input checked="" type="checkbox"/>
D.	chloride	<input checked="" type="checkbox"/>	<input type="checkbox"/>
E.	globulin	<input type="checkbox"/>	<input checked="" type="checkbox"/>

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:

- Hemolytic-uremic syndrome
- Minimal change nephrotic syndrome
- Henoch-Schoenlein purpura with nephritis
- Acute poststreptococcal glomerulonephritis
- Focal and segmental glomerulosclerosis

Matching (B type)

Set 6-8

(A) Captopril	-----> Stem or Header Composed of 5 individual options
(B) Chlorthiazide	
(C) Clonidine	
(D) Guanethidine	
(E) Propranolol	

Adverse effects :

...6... Postural hypotension	-----> An introductory phrase
...7... Bradycardia	
...8... Hypokalemia	

-----> **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 B₁₂ (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/mm³ (80% lymphocytes with many lymphocytes exhibiting typical 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)

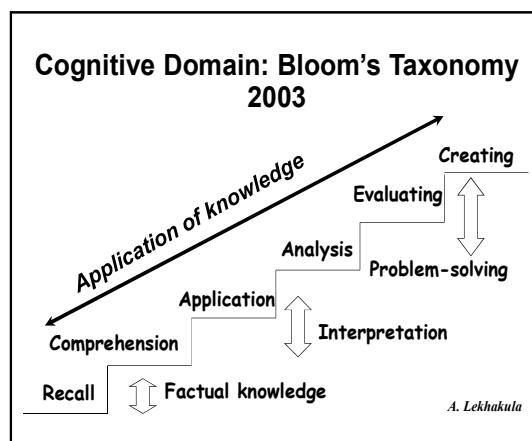
Table of Specification (Blueprint)

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)

Example: Table of Specifications

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



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

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

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

Food	Ate a particular food			Did not eat a particular 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



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



Writing Good MCQ

NONSENSE EXAMINATION

- An exercise to demonstrate how students who know nothing can still get good marks if we do not write good MCQ items (Ref: Professor DE Benor, Ben-Gurion University of the Negev, Beer Sheva, Israel)
- Please try your best in 10 min.

EXPERIENTIAL LEARNING

Construct each item to assess single written objective



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

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 ?

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 แพทย์ต้องการให้เภสัชกรเตรียมยาละลายชื้อในปริมาณ 30 มล. แต่ห้องยามี 25 % NaHCO_3 แอมพูลละ 10 มล. เภสัชกรต้องเตรียมยา ดังกล่าวอย่างไร

- 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 แพทย์ต้องการให้เภสัชกรเตรียมยาละลายชื้อในปริมาณ 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 ซ้ำมา และได้รับการรักษาไปด้วย clindamycin และ gentamicin ต่อมาผู้ป่วยชายเข้ามารับกัมกับที่น้บกรุงทพฯ ผู้ป่วยเล่าว่า 1 สัปดาห์หลังจากเริ่มมี septicemia ผู้ป่วยมีอาการท้องเสีย และท้องบวม ถ่ายเป็นน้ำวันละ 8 ครั้ง ไม่มีเลือด ไม่มีมูกปน ผู้ป่วยมีประวัติที่อุ้มลูกมาหลายปี

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

ท่านคิดว่าอาการท้องเสียของผู้ป่วยน่าจะเกิดจากโรคอะไรมากที่สุด

- A. Faces 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

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)	

การสร้างข้อสอบปรนัย

ผู้ช่วยศาสตราจารย์ นายแพทย์เชิดศักดิ์ โอรมนิรัตน์

ภาควิชาศัลยศาสตร์, คณะแพทยศาสตร์ศิริราชพยาบาล, มหาวิทยาลัยมหิดล, กรุงเทพมหานคร ๑๐๗๐๐.

ข้อสอบปรนัย (multiple-choice question) เป็นรูปแบบการประเมินผลที่นิยมใช้กันอย่างแพร่หลาย ในวงการแพทยศาสตรศึกษาเนื่องด้วยคุณสมบัติที่ดีหลายประการด้วยกัน ได้แก่ ประสิทธิภาพในการประเมินความรู้ปริมาณมากในเวลาอันสั้น ผลการประเมินที่ไม่มีผลกระทบจากความรู้สึกส่วนตัวของผู้ตรวจให้คะแนน คะแนนที่มีความเที่ยงสูง รวมถึงผลการวิจัยจำนวนมากที่สนับสนุนความถูกต้องของผลการประเมินด้วยข้อสอบปรนัย^{๑-๖} ข้อสอบปรนัยที่พัฒนาขึ้นอย่างดีนั้นสามารถวัดความรู้ได้ทั้งระดับการจดจำ การทำความเข้าใจ และการประยุกต์ความรู้ไปใช้ในการดูแลคนไข้^{๓-๔} อย่างไรก็ตาม การประยุกต์ความรู้ไปใช้ในการดูแลคนไข้^{๓-๔} อย่างไรก็ดี ผลการศึกษาวิจัยเกี่ยวกับคุณภาพของข้อสอบปรนัยที่พัฒนาขึ้นใช้ในโรงเรียนแพทย์หลายแห่งพบว่าข้อสอบจำนวนไม่น้อยมีลักษณะที่ไม่เหมาะสม^{๕-๖} ข้อสอบปรนัยที่ถูกพัฒนาขึ้นอย่างไม่ถูกหลักการนั้นส่งผลเสียหลายอย่าง เช่น ทำให้ข้อสอบยากขึ้นโดยไม่จำเป็น ทำให้ผู้สอบเกิดความสับสน ทำให้ผู้สอบบางกลุ่มเสียเปรียบผู้สอบคนอื่น ทำให้การตัดสินใจผลสอบผิดพลาด เป็นต้น^{๖-๗} ดังนั้นการออกข้อสอบปรนัยที่ดี วางอยู่บนหลักการที่ถูกต้องจึงมีความสำคัญมากในการควบคุมคุณภาพการศึกษาในโรงเรียนแพทย์ บทความนี้จะจึงถูกเขียนขึ้นเพื่อเป็นการรวบรวมหลักการพื้นฐานในการออกข้อสอบปรนัยที่ได้รับการยอมรับกันทั่วไปในวงการวัดและประเมินผล ผู้นิพนธ์หวังว่าข้อแนะนำต่าง ๆ ที่ได้นำเสนอในบทความนี้จะเป็แนวทางที่เป็นประโยชน์ในการพัฒนาข้อสอบปรนัยที่มีคุณภาพให้ผู้อ่านไม่มากก็น้อย

รูปแบบพื้นฐานของข้อสอบปรนัย

ข้อสอบปรนัยคือข้อสอบชนิดที่มีคำถามแล้วมีตัวเลือกให้ผู้สอบเลือกตัวเลือกที่เหมาะสมเพื่อตอบคำถามดังกล่าว ข้อสอบปรนัยสามารถแบ่งออกได้เป็น ๒ รูปแบบ^๘ ได้แก่

๑. ข้อสอบถูกผิด (True/false item)

ในข้อสอบประเภทนี้จะมีข้อความให้ผู้สอบพิจารณาว่าถูกหรือผิด ในยุคแรกข้อสอบเหล่านี้แต่ละข้อจะแยกเป็นอิสระจากกัน ผู้สอบตัดสินใจว่าข้อความแต่ละข้อถูกหรือผิดโดยไม่เกี่ยวข้องกับข้อความในข้ออื่น ต่อมาเมื่อผู้พัฒนาข้อสอบเป็นชุดของข้อความ (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 แต่ข้อสอบชนิดนี้มีข้อจำกัดที่สำคัญคือสามารถใช้ได้เฉพาะกับเนื้อหาที่มีความถูกต้องชัดเจนเท่านั้น ซึ่งการตัดสินใจทางการแพทย์ส่วนมากไม่เป็นเช่นนั้น การตัดสินใจในการวินิจฉัย การตรวจค้นเพิ่มเติม หรือการรักษาผู้ป่วยส่วนใหญ่นั้นแพทย์ตัดสินใจเลือกกระหว่างทางเลือกที่แตกต่างกันสามสี่อย่างซึ่งทุกทางเลือกมีความเป็นไปได้ มีส่วนถูก หรือมีความเหมาะสมในบางด้าน

แต่ก็มีความไม่เหมาะสมในด้านอื่นด้วย เช่นการเลือกใช้ยาในผู้ป่วยที่มีการติดเชื้อ นักศึกษาแพทย์มักคิดว่าควรใช้ยาปฏิชีวนะ ซึ่งยาปฏิชีวนะหลายชนิดก็รักษาการติดเชื้อชนิดนั้นๆ ได้ แต่นักศึกษาต้องเลือกระหว่างยาที่ล้วนใช้ได้ในการรักษานั้นว่ายาใดที่มีประสิทธิภาพสูงสุด เหมาะสมที่สุดกับชนิดของเชื้อก่อโรคที่พบบ่อยในการติดเชื้อนั้นมีผลข้างเคียงน้อยที่สุด และราคาเหมาะสมด้วย ซึ่งในสถานการณ์นี้ข้อสอบชนิดถูกผิดจะนำมาใช้ได้ยาก ด้วยเหตุนี้ทำให้ข้อสอบชนิดถูกผิดไม่เป็นที่นิยมกันมากนักในปัจจุบัน

๒. ข้อสอบเลือกคำตอบที่ถูกที่สุด (one best response item)

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

เนื่องจากข้อสอบชนิดที่มีใช้กันแพร่หลายในวงการแพทยศาสตรศึกษาในประเทศไทยในปัจจุบันคือข้อสอบประเภทที่มีตัวเลือก ๔-๕ ตัวเลือก (A-type) ผู้นิพนธ์จะขอเน้นหลักการสำหรับการออกข้อสอบประเภทนี้เป็นสำคัญ

องค์ประกอบของข้อสอบปรนัยชนิดเลือกคำตอบที่ถูกที่สุด

ข้อสอบปรนัยแต่ละข้อมีส่วนประกอบสำคัญ ๒ ส่วนด้วยกันคือ

๑. โจทย์ (stem) เป็นข้อมูลของโรค หรือภาวะ หรือผู้ป่วยตามด้วยคำถาม หรือเว้นช่องว่างสำหรับเติมคำ หรือข้อความที่เหมาะสมลงไป

๒. ตัวเลือก (options) คือคำ หรือข้อความที่

ผู้ออกข้อสอบนำเสนอตามหลังจากโจทย์เพื่อให้ผู้สอบเลือกไปใช้ตอบคำถาม หรือเติมลงในช่องว่างในโจทย์

๒.๑ ตัวเลือกที่ถูกต้อง (correct option) เป็นคำตอบที่ถูกต้องมีเพียงตัวเลือกเดียวต่อข้อสอบข้อหนึ่ง

๒.๒ ตัวลวง (distractors) เป็นคำตอบที่ผิดหรือไม่เหมาะสม มีไว้ลวงให้ผู้สอบที่ไม่มีความรู้ หรือมีความเข้าใจไม่ถูกต้องในเรื่องที่นำมาออกข้อสอบเลือกตอบ ตัวลวงไม่จำเป็นต้องเป็นคำตอบที่ผิดชัดเจนเสมอไป ตัวลวงที่ดีมักมีส่วนถูกบ้าง แต่มีระดับของความถูกต้องเหมาะสมน้อยกว่าคำตอบที่ถูก

ข้อแนะนำพื้นฐานของการเขียนข้อสอบปรนัย

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

๑. เนื้อหาข้อสอบ

๑.๑ ข้อสอบหนึ่งข้อควรมุ่งเน้นประเมินความรู้เพียงเรื่องเดียว

ก่อนเริ่มเขียนข้อสอบอาจารย์ผู้ออกข้อสอบควรตั้งวัตถุประสงค์ให้ชัดเจนว่าต้องการประเมินความรู้ของผู้สอบในเรื่องใด และเขียนโจทย์เพื่อตอบสนองวัตถุประสงค์ดังกล่าวเท่านั้น เนื่องจากเนื้อหาวิชาทางการแพทย์มีมาก อาจารย์แต่ละท่านเมื่อทำการสอนไปแล้วจึงอยากจะทดสอบความรู้ในหลายเรื่องที่ได้สอนไป แต่กลับมีโควตาจำกัดในการออกข้อสอบ ทำให้อาจารย์จำนวนไม่น้อยเขียนข้อสอบหนึ่งข้อถามทั้งเรื่องการวินิจฉัยโรค การตรวจค้นเพิ่มเติม การรักษาโรค และภาวะแทรกซ้อนของโรคไปพร้อมกัน ลักษณะข้อสอบเช่นนี้ไม่ควรใช้ เพราะมักซับซ้อนเกินไป เมื่อผู้สอบตอบข้อสอบผิด ก็ไม่สามารถวินิจฉัยได้ว่าผู้สอบขาดความรู้ ความเข้าใจในเรื่องใด

๑.๒ หลีกเลี่ยงการถามความรู้ในรายละเอียดปลีกย่อยที่ไม่มีที่ใช้ทางคลินิก (trivial content)

องค์ความรู้ทางการแพทย์นั้นมีปริมาณมาก ไม่มีผู้ใดที่จดจำเนื้อหาที่มีในตำรา หรือวารสารทางการแพทย์ได้ทั้งหมด แม้ว่าองค์ความรู้หลายเรื่องมีความน่าสนใจ แต่มีประโยชน์ในการประยุกต์ใช้ทางคลินิกค่อนข้างน้อย องค์ความรู้ดังกล่าวจัดเป็นรายละเอียดปลีกย่อย (trivial content) ซึ่งไม่แนะนำให้ทำการทดสอบ สิ่งที่เราควรทำการประเมินคือความสามารถในการประยุกต์ใช้ความรู้ในทางคลินิก (application of knowledge) ไม่แนะนำการทดสอบวัดความสามารถในการจดจำเป็นหลัก อย่างไรก็ตามการที่แนะนำให้ออกข้อสอบที่เน้นการประยุกต์ใช้ความรู้ ไม่ได้หมายความว่า การแก้ปัญหาผู้ป่วยนั้นไม่ต้องใช้ความจำเลย ตรงกันข้ามการจดจำเนื้อหาเป็นพื้นฐานที่สำคัญในการแก้ปัญหาทางคลินิก ผู้สอบย่อมต้องจำเนื้อหาได้บ้าง จึงจะประยุกต์องค์ความรู้ดังกล่าวไปแก้โจทย์ปัญหาที่นำเสนอได้

๑.๓ หลีกเลี่ยงการถามความรู้ในเรื่องที่ยังมีความขัดแย้งกันในแนวทางปฏิบัติ (controversy)

ความรู้ทางการแพทย์ในหลายหัวข้อยังเป็นเรื่องที่ยังผู้เชี่ยวชาญยังมีความเห็นแตกต่างกัน ผู้ป่วยรายเดียวกันไปพบแพทย์สองคนอาจได้รับการรักษาที่แตกต่างกันซึ่งวิธีการรักษาทั้งสองวิธีก็มีความวิจัยสนับสนุนด้วยกันทั้งคู่ อย่างไรก็ตามยังคงมีความขัดแย้ง (controversy) ในเรื่องดังกล่าวอยู่ เนื้อหาในลักษณะนี้ไม่ควรนำมาออกสอบด้วยข้อสอบปรนัย เนื่องจากในขณะที่ทำข้อสอบอยู่นั้น ผู้สอบไม่มีทางรู้ได้เลยว่าอาจารย์ผู้ออกข้อสอบอ้างอิงจากตำราหรือบทความวิชาการใด เนื้อหาที่ยังมีความขัดแย้ง ที่ผู้เชี่ยวชาญจากต่างสถาบันมีแนวทางในการปฏิบัติที่ต่างกันอย่างนี้แนะนำให้ใช้ข้อสอบในรูปแบบอื่นในการทดสอบเช่นข้อสอบอัตนัย เป็นต้น

๑.๔ หลีกเลี่ยงการลอกประโยคหรือข้อความจากตำราโดยตรง

ดังได้กล่าวแล้วว่าข้อสอบที่ดีควรมุ่งเน้นการประเมินความเข้าใจ หรือ การประยุกต์ใช้ความรู้ ไม่ควรออกข้อสอบที่ประเมินความสามารถในการจำรายละเอียดปลีกย่อย การออกข้อสอบโดยวิธีการเปิดตำราแล้วคัดลอกประโยคจากตำราโดยตรงมักจะลงเอยด้วยข้อสอบที่ทดสอบความจำว่าผู้สอบท่องเนื้อหาในตำราตรงส่วนนั้นได้หรือไม่

ข้อสอบที่ดีควรได้จากการดูผู้ป่วย โจทย์ที่ดีควรเป็นปัญหาของผู้ป่วยที่พบในการทำงานนั่นเอง ตัวเลือกก็ได้จากข้อผิดพลาดที่นักศึกษาหรือแพทย์ประจำบ้านมักปฏิบัติกับผู้ป่วยแล้วทำให้ผลการรักษาไม่ดีนั่นเอง

๑.๕ หลีกเลี่ยงการนำเสนอข้อสอบที่ประเมินความรู้ในเรื่องเดียวกันสองข้อในข้อสอบชุดเดียวกัน

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

๒. การจัดรูปแบบข้อสอบ

๒.๑ เลือกใช้คำศัพท์หรือรูปประโยคที่ง่ายต่อการทำความเข้าใจ

อาจารย์ผู้ออกข้อสอบต้องระลึกไว้เสมอว่าข้อสอบที่อาจารย์ออกเพื่อใช้ในการประเมินผลนักศึกษาแพทย์หรือแพทย์ประจำบ้านนั้นมีวัตถุประสงค์เพื่อทดสอบความรู้ทางการแพทย์เป็นสำคัญ มิใช่การประเมินความรู้ทางภาษาศาสตร์ ดังนั้นการเขียนข้อสอบของอาจารย์ควรเลือกใช้รูปแบบประโยคที่ง่ายต่อการทำความเข้าใจ อย่าเขียนประโยคซับซ้อนที่มีความยาวประโยคหลายบรรทัด มุ่งเน้นให้ภาษาเป็นสื่อในการนำเสนอความคิดของอาจารย์ผู้ออกข้อสอบไปยังผู้สอบ อย่าให้

เขตนศกศรศรศ

บทความหัวบ

ผู้สอบต้องอ่านข้อสอบย้อนไปมาหลายรอบกว่าจะเข้าใจจุดประสงค์ของข้อสอบ แล้วจึงตัดสินใจเลือกคำตอบโดยทั่วไปแนะนำให้อาจารย์นำเสนอรายละเอียดต่าง ๆ ไว้ในตัวโจทย์ให้มากที่สุด ส่วนตัวเลือกเขียนเป็นคำหรือข้อความสั้น ๆ

๓.๓ หลีกเลียงการเขียนโจทย์ที่มีรูปประโยคเป็นเชิงปฏิเสธ

โจทย์ที่ดีไม่ควรอยู่ในประโยคเชิงปฏิเสธ เช่นถามถึงสิ่งที่เป็นข้อยกเว้น สิ่งที่ไม่ควรปฏิบัติ สิ่งที่มีน้อยที่สุด หรือสิ่งที่ไม่น่าถึงเป็นต้น งานวิจัยส่วนใหญ่พบว่าข้อสอบที่มีโจทย์ในรูปแบบปฏิเสธเหล่านี้มีระดับความยากง่ายไม่ต่างจากข้อสอบอื่น ๆ แต่งานวิจัยบางชิ้นพบว่าข้อสอบที่มีโจทย์ในรูปแบบปฏิเสธมีความยากมากกว่าข้อสอบอื่นชัดเจนโดยเฉพาะในข้อสอบวัดความรู้ระดับสูง^{๑๑-๑๒} แต่ผู้เชี่ยวชาญในการประเมินผลส่วนใหญ่มีความเห็นพ้องกันว่าข้อสอบประเภทนี้สามารถสร้างความสับสนให้กับผู้สอบได้ จึงไม่แนะนำให้ใช้ แต่หากอาจารย์ผู้ออกข้อสอบมีความจำเป็นต้องใช้ข้อสอบที่มีการใช้คำปฏิเสธในโจทย์ แนะนำให้พิมพ์คำปฏิเสธให้เด่นชัด โดยใช้ตัวหนาและขีดเส้นใต้เพื่อให้ผู้สอบเห็นชัด^{๑๑}

๔. การเขียนตัวเลือก

๔.๑ เขียนตัวเลือกที่มีประสิทธิภาพให้มีจำนวนมากที่สุดเท่าที่เหมาะสมกับบริบท

เรื่องจำนวนตัวเลือกที่เหมาะสมนี้เป็นเรื่องผู้เชี่ยวชาญด้านการประเมินผลจำนวนมากสนใจ มีงานวิจัยเกี่ยวกับเรื่องจำนวนตัวเลือกที่เหมาะสมในข้อสอบปรนัยอยู่มากมาย^{๑๓} อาจารย์ผู้ออกข้อสอบส่วนมากจะคุ้นเคยกับข้อสอบปรนัยชนิดที่มีห้าตัวเลือก บ่อยครั้งที่อาจารย์ออกข้อสอบแล้วนึกตัวเลือกได้เพียงสามหรือสี่ตัว จึงเกิดคำถามว่าจำเป็นต้องมีตัวเลือกครบห้าตัวเลือกหรือไม่ งานวิจัยบางชิ้นพบว่าการลดจำนวนตัวเลือกลงทำให้ข้อสอบง่ายขึ้น^{๑๓-๑๔} แต่งานวิจัยบางชิ้นพบว่าการลดจำนวนตัวเลือกลงทำให้ได้ข้อสอบยากขึ้น^{๑๕-๑๖} ผู้เชี่ยวชาญในการประเมินผลเสนอว่าข้อสอบปรนัยที่มีตัวเลือกเพียงสามตัวเลือกก็สามารถทดสอบความรู้ได้อย่างมีประสิทธิภาพ^{๑๖-๑๗, ๑๗} แต่มีอาจารย์จำนวนไม่น้อยที่ไม่สบายใจที่มีตัวเลือกในข้อสอบแต่ละข้อน้อยกว่าห้าตัว

เลือกด้วยกังวลว่าจะทำให้มีโอกาสสูงที่ผู้สอบที่ไม่มีความรู้จะเดาสุ่มได้คำตอบที่ถูกต้อง แต่จากข้อมูลที่ปรากฏในปัจจุบันพบว่าผู้สอบในการสอบในระดับสูงนั้นพฤติกรรมเดาสุ่มโดยที่ผู้สอบปราศจากความรู้นั้นน่าจะมีบทบาทน้อยมาก ผู้สอบส่วนใหญ่มักพอมีความรู้บ้างและสามารถตัดตัวเลือกที่ไม่สมเหตุผลอย่างชัดเจนได้^{๑๖} ในการศึกษาข้อสอบปรนัยส่วนใหญ่พบตัวเลือกที่ไม่ทำงานเป็นจำนวนไม่น้อย^{๑๘} ข้อมูลที่ได้จากการวิเคราะห์ข้อสอบปรนัยที่ใช้ในทางแพทยศาสตรศึกษาในประเทศไทยหลายครั้งก็สอดคล้องกับงานวิจัยในต่างประเทศที่พบว่าข้อสอบส่วนใหญ่มักมีตัวเลือกที่ทำงานจริงราวสามหรือสี่ตัวเลือก มีข้อสอบน้อยข้อมากที่ตัวเลือกทั้งห้าตัวเลือกทำงานอย่างมีประสิทธิภาพ

ด้วยข้อมูลจากการศึกษาต่าง ๆ ข้อแนะนำในการออกข้อสอบปรนัยในปัจจุบันคือให้อาจารย์เขียนจำนวนตัวเลือกมากที่สุดที่มีความเหมาะสมกับเนื้อหาโจทย์ ไม่จำเป็นต้องเขียนตัวเลือก ๕ ตัวเลือกเสมอไป เนื่องจากตัวเลือกที่ห้าที่เขียนขึ้นเพื่อเติมเต็มโดยไม่สมเหตุผลนั้นมักไม่ค่อยมีคนเลือก หากเนื้อหาที่อาจารย์นำมาสอบมีตัวเลือกที่เหมาะสมเพียงสามหรือสี่ตัวเลือกก็เขียนจำนวนตัวเลือกเพียงสามหรือสี่ตัวเลือก^{๑๑} แต่อย่างไรก็ตามให้อาจารย์ศึกษาข้อกำหนดของแต่ละการสอบที่อาจารย์เกี่ยวข้องด้วย เนื่องจากนโยบายของแต่ละการสอบแตกต่างกันไป องค์กรที่จัดสอบทางแพทยศาสตรศึกษาจำนวนไม่น้อยยังคงตั้งข้อกำหนดให้ใช้ข้อสอบ ๕ ตัวเลือกเสมอ ซึ่งหากอาจารย์ไม่ทำตามข้อกำหนดดังกล่าวข้อสอบที่ออกไปอาจไม่ได้รับการพิจารณาได้

๔.๒ จัดให้ตัวเลือกที่ถูกต้องมีการกระจายตำแหน่งไปให้มีจำนวนพอ ๆ กันในทุกตัวเลือก

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

๔.๓ เขียนตัวเลือกแต่ละข้อให้เป็นอิสระ ไม่ขึ้นต่อกัน

๓๓

มกราคม-มิถุนายน ๒๕๕๕, ปีที่ ๕, ฉบับที่ ๑

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

๔.๔ เขียนตัวเลือกให้ทุกตัวเลือกมีความเป็นเนื้อเดียวกัน (homogeneous)

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

ในบางบริบทของการดูแลรักษาผู้ป่วย สิ่งแพทย์ต้องตัดสินใจเลือกอาจมีทั้งการเลือกที่จะให้การรักษาเลยหรือจะส่งตรวจค้นเพิ่มเติมก่อน ในกรณีนี้อาจารย์สามารถเขียนตัวเลือกที่มีการรักษาและการตรวจเพิ่มเติมปะปนกันได้ แต่การเขียนรูปประโยคคำถามต้องไม่เป็นการบอกใบ้ว่าจะไปทิศทางใด แต่ต้องเลือกใช้คำถามที่เป็นกลาง เช่น ท่านจะปฏิบัติต่อผู้ป่วยอย่างไร, ท่านจะดำเนินการอย่างไรต่อไป เป็นต้น

๔.๕ เขียนตัวเลือกแต่ละข้อให้มีความยาวพอกัน

จากการสังเกตข้อสอบปรนัยจำนวนมากจะพบว่าตัวเลือกที่ถูกต้องมักมีความยาวมากกว่าตัวเลือกอื่น ซึ่งข้อสังเกตนี้ผู้สอบจำนวนไม่น้อยก็ทราบดี และผู้สอบส่วนมากเมื่อไม่ทราบคำตอบก็มักเลือกตัวเลือกที่มีความยาวมากที่สุด ดังนั้นอาจารย์ผู้ออกข้อสอบควรระมัดระวังไม่ให้ตัวเลือกตัวใดตัวหนึ่งมีความยาวแตกต่างไปจากตัวเลือกอื่นชัดเจน เพราะจะทำให้ผู้สอบเดาคำตอบที่ถูกต้องได้ง่าย

๔.๖ หลีกเลี่ยงการใช้ตัวเลือก “ถูกทุกข้อ” หรือ “ไม่มีข้อใดถูก”

ตัวเลือก “ถูกทุกข้อ” เป็นตัวเลือกที่ผู้เชี่ยวชาญในการประเมินผลส่วนใหญ่เห็นสอดคล้องกันว่าไม่ควรใช้เนื่องจากมักช่วยใบ้ตัวเลือกที่ถูกต้องให้กับผู้สอบ ทำให้ผู้สอบส่วนหนึ่งตอบถูกโดยไม่ต้องอาศัยองค์ความรู้ที่สมบูรณ์ในเรื่องที่ทดสอบ งานวิจัยพบว่าข้อสอบที่มีตัวเลือกชนิดนี้จะมีผลให้ค่าความเที่ยงของคะแนนสอบลดลง^{๑๑} จึงแนะนำให้หลีกเลี่ยงการใช้

ตัวเลือก “ไม่มีข้อใดถูก” เป็นประเด็นที่ผู้เชี่ยวชาญในการประเมินผลยังคงถกเถียงกันอยู่บ้าง ผู้เชี่ยวชาญบางส่วนเห็นว่าไม่ควรใช้ตัวเลือกประเภทนี้ แต่ผู้เชี่ยวชาญบางส่วนให้ความเห็นว่าสามารถใช้ได้ในบางกรณี^{๑๒} เหตุผลที่ตัวเลือกชนิดนี้เป็นปัญหาคือการใช้ตัวเลือกนี้มักสร้างความลำบากใจให้กับผู้สอบในการเลือกคำตอบที่ถูกในกรณีที่ตัวเลือกแต่ละตัวเลือกไม่ถูกหรือผิดชัดเจน เพราะผู้สอบจะต้องทำการเปรียบเทียบตัวเลือกที่น่าเสนอในข้อสอบกับทางเลือกอื่น ๆ ที่เขานึกได้^{๑๓} หากโจทย์ถามว่า ยาใดที่ควรให้แก่ผู้ป่วย แล้วมีชื่อยาสี่ชนิด และมีตัวเลือก “ไม่มีข้อใดถูก” นอกจากที่ผู้สอบต้องนึกว่าในบรรดา ยาที่ปรากฏในตัวเลือกนั้นเหมาะสมหรือไม่แล้วเขายังนึกต่อไปอีกว่ามียาอื่นใดที่สามารถให้ผู้ป่วยรายนี้ได้อีก หากเขานึกออกว่ามียาอื่นที่น่าจะเหมาะสมกับผู้ป่วยมากกว่ายาในตัวเลือก (ด้วยเหตุผลที่อาจแตกต่างไปจากที่อาจารย์ผู้ออกข้อสอบคิด) เขาก็จะเลือก “ไม่มีข้อใดถูก”

การใช้ตัวเลือก “ไม่มีข้อใดถูก” จะยังเป็นปัญหามากขึ้นในข้อสอบที่ถามถึงสิ่งที่ไม่ควรทำ เช่นยาใดไม่ควรใช้ในผู้ป่วย ซึ่งนอกจากยาที่น่าเสนอในตัวเลือกแล้วย่อมมียาชนิดอื่นอีกมากมายในบัญชียาที่ไม่เหมาะสม ซึ่งไม่มี

ทางที่ใครจะรู้ว่าการที่ผู้สอบเลือกตอบ “ไม่มีข้อใดถูก” นั้นเขาคิดถึงยาใด และยานั้นไม่เหมาะสมมากไปกว่ายาที่มีอยู่ในตัวเลือกหรือไม่ งานวิจัยทั้งหมดที่ศึกษาถึงตัวเลือกชนิดนี้ได้ข้อสรุปที่ตรงกันว่าข้อสอบที่ใช้ตัวเลือกประเภทนี้เพิ่มระดับความยากให้ข้อสอบ^{๑๖} โดยทั่วไปแล้วจึงไม่แนะนำให้ใช้ตัวเลือกประเภทนี้ในการสอบทางแพทยศาสตรศึกษาซึ่งทางเลือกสำหรับสถานการณ์ที่น่าเสนาหามีได้มากและการตัดสินใจเลือกคำตอบต้องอาศัยการเปรียบเทียบข้อดีข้อเสียของแต่ละตัวเลือก

สรุป

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

มีข้อควรพิจารณาในการประยุกต์ใช้ข้อแนะนำเหล่านี้ในการพัฒนาข้อสอบที่ผู้นิพนธ์ขอล่าถึงประการหนึ่งคือ แม้ว่าข้อแนะนำที่กล่าวถึงเหล่านี้หลายข้อมีการศึกษาวิจัยสนับสนุนที่ชัดเจน แต่สิ่งเหล่านี้ก็เป็นเพียงข้อแนะนำว่าผู้ออกข้อสอบควรปฏิบัติ ไม่ใช่กฎเกณฑ์ตายตัว การเขียนข้อสอบปรนัยนั้นเป็นงานที่ต้องอาศัยทั้งศาสตร์และศิลปะผสมผสานกันอย่างเหมาะสม

หาใช้สูตรคณิตศาสตร์ที่ไม่มีข้อยกเว้น ผู้นิพนธ์ไม่คาดหวังให้อาจารย์ผู้พัฒนาข้อสอบยึดข้อแนะนำเหล่านี้เสมือนกฎเกณฑ์ตายตัวที่ต้องทำตามในทุกกรณี หากแต่ต้องการให้อาจารย์ใช้เป็นแนวทางในการสร้างข้อสอบ ในบางบริบทผู้ออกข้อสอบอาจเลือกที่จะไม่ปฏิบัติตามข้อแนะนำบางประการได้บ้าง แต่การที่จะไม่ปฏิบัติตามข้อแนะนำเหล่านี้จำเป็นต้องมีเหตุผลที่เหมาะสม และควรทำไม่บ่อยนัก ยกตัวอย่างเช่นข้อแนะนำว่า โจทย์ไม่ควรเขียนถามข้อยกเว้น จะพบได้ว่ามีบางบริบทที่การรู้ข้อยกเว้น หรือข้อห้ามปฏิบัติก็เป็นองค์ความรู้ที่สำคัญในการดูแลรักษาผู้ป่วย ดังนั้นในบริบทที่เหมาะสมผู้นิพนธ์เองก็เห็นด้วยว่าอาจเขียนโจทย์ที่ถามข้อยกเว้นได้ แต่อย่างไรก็ตามการจะไม่ปฏิบัติตามข้อแนะนำนี้ต้องไม่ทำบ่อยจนเกินจำเป็น หากออกข้อสอบ ๑๐๐ ข้อ จะมีข้อสอบที่ถามข้อยกเว้น ประมาณ ๒-๓ ข้อ ย่อมเป็นสิ่งที่ยอมรับได้ แต่หากในชุดข้อสอบมีข้อสอบถึงร้อยละ ๒๐ - ๓๐ ที่โจทย์เขียนในรูปประโยคปฏิเสธ ถามสิ่งที่ไม่ควรปฏิบัติ หรือสิ่งที่ไม่ถูกต้อง อย่างนี้ย่อมจัดว่าละเลยแนวทางในการพัฒนาข้อสอบอย่างไม่เหมาะสม ซึ่งย่อมส่งผลให้คุณภาพของข้อสอบด้อยลงอย่างชัดเจน

เอกสารอ้างอิง

1. Downing SM. Assessment of knowledge with written test forms. In: Norman GR, van der Vleuten C, Newble DI, editors. International handbook of research in medical education. Dordrecht: Kluwer Academic Publishers, 2002:647 - 72.
2. Haladyna TM, Downing SM. A taxonomy of multiple-choice item-writing rules. Appl Meas Educ 1989;2:37-50.
3. Haladyna TM. Developing and validating multiple-choice test items, 3rd ed. Mahwah, NJ: Lawrence Erlbaum Associates, 2004.
4. Maatsch JL, Huang RR, Downing SM, Munger BS. The predictive validity of test formats and a psychometric theory of clinical competence. The 23rd Conference on Research in Medical Education. Washington, DC: Association of American Medical Colleges, 1984.
5. Jozefowicz RF, 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.

ข้อผิดพลาดที่ควรระวังในการสร้าง ข้อสอบปรนัย

ผู้ช่วยศาสตราจารย์ นายแพทย์เชิดศักดิ์ โอรมนิรัตน์

ภาควิชาศัลยศาสตร์, คณะแพทยศาสตร์ศิริราชพยาบาล, มหาวิทยาลัยมหิดล, กรุงเทพมหานคร ๑๐๓๐๐.

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

การสร้างข้อสอบปรนัยที่มีคุณภาพดีนั้นควรเริ่มต้นจากการมีองค์ความรู้พื้นฐานในการสร้างข้อสอบแล้ว เกิดการฝึกฝนทักษะ สังเกตประสบการณ์ในการออกข้อสอบ จนเกิดความชำนาญ ปัญหาที่พบบ่อยในโรงเรียนแพทย์หลายแห่งคือมีอาจารย์จำนวนไม่น้อยที่ได้รับมอบหมายให้ออกข้อสอบปรนัย โดยไม่ได้มีการพัฒนาองค์ความรู้พื้นฐานที่เหมาะสมก่อน ซึ่งเป็นเหตุให้มีข้อสอบปรนัยที่มีลักษณะไม่เหมาะสมตามหลักการออกข้อสอบปะปนมาในข้อสอบที่ให้นักศึกษาแพทย์และแพทย์ประจำบ้านทำอยู่บ้าง ผู้นิพนธ์จึงเห็นความสำคัญของการเผยแพร่องค์ความรู้พื้นฐานของการออกข้อสอบปรนัย องค์ความรู้พื้นฐานในการสร้างข้อสอบปรนัยนั้นมีสองส่วน ส่วนแรกเป็นหลัก การของการสร้างข้อสอบทั่วไปซึ่งได้มีผู้รวบรวมเป็นข้อแนะนำตีพิมพ์ในตำราและวารสารทางวิชาการอยู่บ้าง^{๖,๕-๗} ส่วนที่สองเป็นข้อผิดพลาดในการสร้างข้อสอบที่อาจารย์ผู้ออกข้อสอบพึงหลีกเลี่ยง ในบทความนี้ผู้นิพนธ์จะมุ่งเน้นในส่วนที่สองนี้ โดยจะรวบรวมข้อผิดพลาดในการสร้างข้อสอบปรนัย ที่อาจเป็นตัวบดบังให้ผู้สอบที่ไม่มีความรู้ในเรื่องที่ทำการทดสอบสามารถเลือกคำตอบที่ถูกต้องได้ ดังนั้นการที่อาจารย์ผู้ออกข้อสอบทราบถึงสิ่งเหล่านี้และหลีกเลี่ยงเสียจะส่งผลให้ข้อสอบปรนัยที่สร้างขึ้นสามารถใช้วัดองค์ความรู้ทางการแพทย์ได้จริง โดยปราศจากปัจจัยรบกวนจากการสังเกตพบสิ่งบดบังคำตอบ

๓/๓๗

กรกฎาคม-ธันวาคม ๒๕๕๕, ปีที่ ๕, ฉบับที่ ๒

เวบบทีกิธีราช

บทความทั่วไป

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

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

๑. ข้อผิดพลาดในไวยากรณ์

ตัวเลือกทุกตัวต้องสามารถตอบโจทย์ได้อย่างถูกต้องตามหลักไวยากรณ์ บ่อยครั้งอาจารย์ผู้ออกข้อสอบมุ่งความสนใจไปที่คำตอบที่ถูก และให้ความสนใจกับตัวลวงน้อยไปจนทำให้ตัวลวงผิดหลักไวยากรณ์ โดยมักพบบ่อยในข้อสอบที่เป็นภาษาอังกฤษ ข้อผิดพลาดที่พบได้บ่อยเช่น ความไม่เข้ากันของ article (A, An, The) กับคำนามที่ตามหลัง, คำนามกับกริยาที่ไม่เข้ากันในเชิงเอกพจน์หรือพหูพจน์, การเติมคำในประโยคที่เว้นว่างไว้สำหรับเติมคำนามแต่ตัวลวงเป็นกริยาหรือเป็นคำนามในลักษณะที่ไม่เข้ากับรูปประโยค เป็นต้น

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

- A. lumbar puncture
- B. computerized scan of the brain
- C. glucose with Thiamine
- D. Sodium bicarbonate

ในตัวอย่างที่ ๑ นี้โจทย์ให้ผู้สอบเลือกตัวเลือกไปเติมในช่องว่าง ซึ่งสิ่งที่เติมลงในช่องว่างได้นั้นต้องเป็นยาที่สามารถให้ทางหลอดเลือดดำได้ ผู้สอบที่มีทักษะการทำข้อสอบดีจะตัดตัวเลือก A และ B ได้โดยไม่ต้องใช้ความรู้ทางการแพทย์

ตัวอย่างที่ ๒. Which organism is the cause of syphilis?

- A. *Neisseria gonorrhoeae*
- B. *Chlamydia trachomatis* and *Giardia lamblia*
- C. *Treponema pallidum*
- D. *Ureaplasma urealyticum* and *Mycoplasma genitalium*

ในตัวอย่างที่ ๒ นี้โจทย์ถามหาเชื้อก่อโรค โดยใช้รูปประโยคถามหาคำตอบที่เป็นเอกพจน์ ดังนั้นคำตอบที่ถูกต้องย่อมมีเชื้อก่อโรคตัวเดียว ผู้สอบที่มีทักษะการทำข้อสอบดีจะตัดตัวเลือก B และ D ได้โดยไม่ต้องใช้ความรู้ทางการแพทย์

๒. การไปคำตอบด้วยหลักตรรกะ

ในการเขียนตัวเลือก อาจารย์ผู้ออกข้อสอบต้องระมัดระวังไม่ให้ผู้สอบสามารถตัดตัวเลือกได้ด้วยหลักตรรกศาสตร์ เนื่องจากผู้สอบที่มีทักษะการทำข้อสอบดีจะสามารถพิจารณาความเป็นไปได้ของตัวเลือกต่าง ๆ และตัดตัวลวงที่ไม่มีทางเป็นไปได้ตามหลักของเหตุและผลออกไปได้โดยไม่ต้องอาศัยความรู้เรื่องที่ว่าอาจารย์ตั้งเป้าหมายว่าจะทดสอบ

ตัวอย่างที่ ๓. ภาวะไส้เลื่อนบริเวณขาหนีบ (inguinal hernia)

- A. พบในผู้ชายบ่อยกว่าผู้หญิง
- B. พบในผู้หญิงบ่อยกว่าผู้ชาย
- C. พบเกิดขึ้นในผู้หญิงและผู้ชายในอัตราเท่ากัน
- D. พบบ่อยในผู้ที่มีเศรษฐกิจฐานะยากจน
- E. พบในผู้ที่มีภูมิละเนาในทวีปเอเชีย มากกว่าผู้ที่มีภูมิละเนาในทวีปยุโรป

ในตัวอย่างที่ ๓ นี้อาจารย์ผู้ออกข้อสอบต้องการวัดความรู้เรื่องอุบัติการณ์ของไส้เลื่อนขาหนีบ แต่หาก

เวบบิ้นทีกีรราช

บทความทั่วไป

พิจารณาดตามหลักตรรกศาสตร์แล้ว ตัวเลือก A, B, และ C เพียงสามตัวเลือกก็ครอบคลุมสิ่งที่เป็นไปได้ทั้งหมดแล้ว (เนื่องจากมนุษย์มีสองเพศ ภาวะได้เลื่อนนี้หากไม่มีอัตราการเกิดเท่ากันในสองเพศแล้วก็ต้องมีเพศใดเป็นมากกว่าอีกเพศหนึ่ง) ดังนั้นผู้สอบที่มีทักษะการทำข้อสอบดีสามารถตัดตัวเลือก D และ E ได้โดยไม่ต้องมีความรู้เรื่องได้เลื่อนเลย

๓. การใช้คำคุณศัพท์บอกระดับของความแน่ชัด

อาจารย์ผู้ออกข้อสอบพึงระมัดระวังการใช้คำคุณศัพท์ที่บ่งบอกถึงความแน่ชัดของข้อความ ซึ่งจะมีหลายระดับ โดยทั่วไปแล้วคำคุณศัพท์ที่แสดงความแน่ชัดมาก แสดงความมั่นใจมาก (เช่น always, never) มักไม่ถูกต้อง เนื่องจากในทางการแพทย์นั้นมีความไม่แน่นอนเกิดขึ้นเป็นประจำ ข้อความที่บอกเล่าถึงสิ่งที่จะเป็นไปได้โดยไม่ชี้ชัดลงไปว่าต้องเกิดขึ้นแน่นอน (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 นี้ผู้สอบต้องเลือกข้อความเกี่ยวกับได้เลื่อนขาหนีบที่ถูกต้องหนึ่งข้อความ หากสังเกตดูทั้งสี่ข้อความมีการใช้คำคุณศัพท์บอกความแน่ชัดของข้อความ ได้แก่ may (ตัวเลือก A), always (ตัวเลือก B), the only (ตัวเลือก C), inevitably (ตัวเลือก D) ซึ่งจะเห็นว่าตัวเลือก B, C, และ D เป็นข้อความที่แสดงความแน่ชัดว่าต้องเป็นแน่ ต้องใช่แน่นอน ไม่มีทางเลี่ยงได้ ข้อความทำนองนี้มีโอกาสสูงที่จะผิด ในทางตรงข้ามตัวเลือก A เป็นข้อความบอกว่ามีโอกาสเป็นไปได้โดยไม่ชี้ชัดว่าต้องเกิด

ผู้สอบที่มีทักษะการทำข้อสอบดีจะตัดตัวเลือก B, C, และ D ได้โดยไม่ต้องอาศัยความรู้ทางการแพทย์เลย

๔. ความยาวของตัวเลือก

มีการตั้งข้อสังเกตว่าอาจารย์แพทย์มักชอบสอนและอธิบายแม้กระทั่งในการสอบอาจารย์แพทย์หลายท่านก็ติดนิสัยรักการสอนนี้มาด้วย ทำให้อาจารย์มักเขียนตัวเลือกที่ถูกต้องที่มีคำอธิบายประกอบอย่างครบถ้วนทำให้ตัวเลือกที่ถูกมักมีความยาวมากกว่าตัวลวง^๕ นักศึกษาผู้เข้าสอบจำนวนไม่น้อยรู้ถึงความจริงข้อนี้และมักเลือกตัวเลือกที่มีความยาวมากที่สุด หากเขาไม่สามารถหาคำตอบได้ด้วยความรู้ทางการแพทย์ที่เขา

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

A. ให้เปลี่ยนไปใช้การใส่ห่วงอนามัย

B. ให้ใช้ถุงยางอนามัย

C. ให้กินยาคุมกำเนิดต่อได้เนื่องจากมีการศึกษาแล้วว่ายาคุมกำเนิดชนิดกินไม่ส่งผลให้เกิดการเพิ่มขึ้นของน้ำหนักตัว

D. ให้รับประทานยาลดความอ้วน

ในตัวอย่างที่ ๕ นี้จะสังเกตเห็นว่าตัวเลือก C มีการอธิบายเหตุผลประกอบส่งผลให้มีความยาวมากกว่าตัวเลือกอื่นชัดเจน ลักษณะเช่นนี้จะเป็นการบอกใบ้ให้นักศึกษาเลือกตัวเลือกนี้

๕. การใช้คำซ้ำในโจทย์และตัวเลือก

การใช้คำเดียวกัน หรือคำที่มีความหมายเหมือนกันในโจทย์และตัวเลือก มักเป็นการบอกใบ้ว่าตัวเลือกดังกล่าวเป็นตัวเลือกที่ถูกต้อง^๖

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

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

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

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 sacular theory ซึ่งหากแปลความหมายก็น่าจะเป็นเรื่องที่เกี่ยวข้องกับถุง (sac) ผู้สอบที่มีทักษะการทำข้อสอบดีจะหาตัวเลือกที่มีคำที่มีความหมายเกี่ยวกับถุง แล้วเลือกตัวเลือกดังกล่าวทันที ซึ่งในที่นี้จะพบคำว่า diverticulum ซึ่งมีความหมายว่าถุงในข้อ B การที่มีคำที่มีความหมายซ้ำกันเช่นนี้เป็นตัวบอกใบ้คำตอบที่อาจารย์ผู้ออกข้อสอบต้องตรวจตราให้ดีก่อนนำข้อสอบไปใช้

๖. การเข้าพวของคำ หรือข้อความที่ปรากฏในตัวเลือก

ข้อสอบจำนวนไม่น้อยนำเสนอรายการของหลายอย่างในตัวเลือก (เช่น ชื่อการตรวจค้นเพิ่มเติม ชื่อโรค ชื่อยา ฯลฯ) มีผู้เชี่ยวชาญในการประเมินผลตั้งข้อสังเกตว่าในข้อสอบเหล่านี้ตัวเลือกที่ถูกต้องมักมีลักษณะเข้าพวกับตัวเลือกอื่นมากที่สุด หากเป็นรายการของตัวเลือกที่ถูกก็คือข้อที่มีจำนวนรายการซ้ำกับตัวเลือกอื่นมากที่สุด ดังนั้นในการนำเสนอตัวเลือกอาจารย์ผู้ออกข้อสอบพึงระมัดระวังอย่าให้ตัวเลือกที่ถูกต้องมีลักษณะที่เข้าพวได้อย่างชัดเจน พยายามทำตัวลวงอื่นให้มีลักษณะเข้าพวให้ใกล้เคียงกับตัวเลือกที่ถูกต้อง

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 ดังนั้นตัวเลือกที่มีพวมากที่สุดคือตัวเลือก A ซึ่งเป็นคำตอบที่ถูกต้อง

การเข้าพวของตัวเลือกที่ถูกนั้นไม่จำเป็นต้องเป็นลักษณะของการมีจำนวนหรือความถี่ของคำมากที่สุดเพียงเท่านั้น อาจหมายถึงรวมถึงการมีรูปร่างลักษณะหรือความหมายคล้ายคลึงกันได้ด้วย

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

A. Amoxicillin PO

B. Ceftazidime IV + Amikacin IV

C. Amphotericin B IV + Ceftazidime IV

D. Cloxacillin IV + Metronidazole IV

ในตัวอย่างที่ ๘ นี้โจทย์ถามถึงยาที่ควรให้กับผู้ป่วย ในตัวเลือกสี่ตัวเลือกนี้มียาเกินเพียงข้อเดียว (A) ที่เหลือเป็นยาชนิดสองขนานควบกัน ดังนั้นตัวเลือกข้อ A ไม่เข้าพว จะถูกตัดทิ้งได้โดยง่าย ในบรรดา ยาชนิดจะเห็นว่ามียาต้านเชื้อราที่ไม่เข้าพว (ตัวเลือก C) ดังนั้นจะเหลือตัวเลือกที่นักศึกษาต้องคิดเลือกจริง ๆ เพียงตัวเลือก B กับ D ซึ่งหากดูกลุ่มยา ก็จะพบว่ายากลุ่ม Cephalosporin เข้าพวมากที่สุด ทำให้ผู้สอบที่มีทักษะการทำข้อสอบดีสามารถเลือกคำตอบที่ถูกต้อง (ตัวเลือก B) ได้โดยไม่ต้องมีความรู้เรื่องการรักษาผู้ป่วย febrile neutropenia

๗. การบอกใบ้คำตอบโดยโจทย์ข้ออื่น

ข้อผิดพลาดนี้เป็นข้อผิดพลาดที่ตัวผู้เขียนข้อสอบไม่ค่อยรู้ แต่ผู้ที่จะตรวจพบข้อผิดพลาดนี้คืออาจารย์ผู้เลือกข้อสอบไปใช้ เนื่องจากในการสอบแต่ละครั้งใช้ข้อสอบจำนวนมาก หากเลือกข้อสอบโดยไม่ระมัดระวังอาจมีข้อสอบสองข้อที่ถามเกี่ยวกับโรคหรือกลุ่มอาการเดียวกัน ซึ่งข้อมูลจากโจทย์ในข้อหนึ่งอาจเป็นตัวบอกใบ้คำตอบของข้อสอบอีกข้อได้ ดังนั้นเมื่อทำการเลือกข้อสอบเสร็จแล้วจัดหน้ากระดาษเข้ารูปเล่มข้อสอบแล้วอาจารย์ควรอ่านข้อสอบฉบับสมบูรณ์นี้อีกหนึ่งหรือสองรอบก่อนส่ง

เวบบททศกรรช

บทความหัวโ

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 ซึ่งผู้สอบที่มีทักษะการทำข้อสอบดีสามารถย้อนกลับไปอ่านโจทย์ในข้อก่อนหน้านี้ได้ ข้อมูลว่าผู้ป่วยที่นำเสนอว่าเป็น febrile neutropenia มีไข้ขึ้น ๑๔ วันหลังได้ยาเคมีบำบัด ก็สามารถตอบข้อสอบข้อนี้ถูกต้องได้ง่าย

สรุป

ผู้นิพนธ์ได้รวบรวมข้อผิดพลาดในการสร้างข้อสอบปรนัยที่ผู้สอบอาจใช้เป็นแนวทางในการเลือกคำตอบที่ถูกต้องโดยไม่ต้องอาศัยความรู้ทางการแพทย์ที่อาจารย์ต้องการประเมินผล โดยเรียงเรียงเป็นเจ็ดกลุ่มข้อผิดพลาดด้วยกัน ผู้อ่านทุกท่านพึงตระหนักว่าสิ่งเหล่านี้ไม่ใช่หลักการทางวิทยาศาสตร์ที่ชัดเจนดังกฎทางคณิตศาสตร์หรือฟิสิกส์ หากแต่เป็นการรวบรวมข้อสังเกต

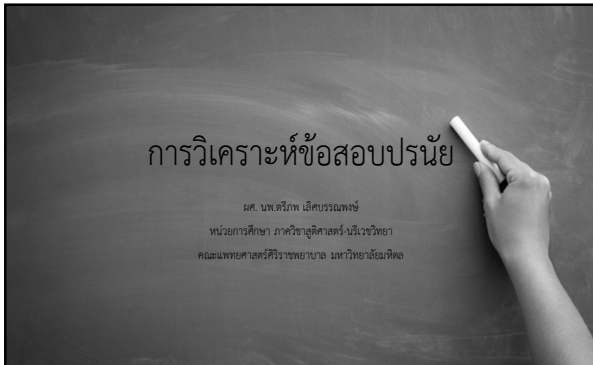
และคำแนะนำของผู้เชี่ยวชาญทางการวัดและประเมินผล จึงเป็นเพียงแนวทางเบื้องต้นในการพิจารณาตรวจสอบเนื้อหาของข้อสอบเท่านั้น การประยุกต์ใช้องค์ความรู้นี้คงต้องอาศัยศิลปะพอสมควรเพื่อที่จะได้ข้อสอบที่ดีสามารถวัดองค์ความรู้ทางการแพทย์ของนักศึกษาหรือแพทย์ประจำบ้านที่เข้าสอบได้ตามวัตถุประสงค์ของการสอบ

เอกสารอ้างอิง

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

ผศ. นพ.ตรีภพ เลิศบรรณพงษ์

หัวข้อ : Multiple-choice questions item analysis



เป้าหมายการเรียนรู้

1. วิเคราะห์ข้อสอบปรนัยรายข้อได้
2. วิเคราะห์ชุดข้อสอบปรนัยได้
3. อธิบายการประยุกต์ใช้ (application) ผลการวิเคราะห์ข้อสอบปรนัยได้
4. อธิบายข้อจำกัด (limitation) ของผลการวิเคราะห์ข้อสอบปรนัยได้

Classical test theory

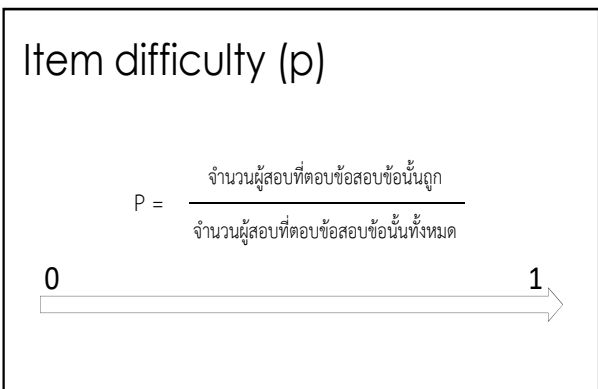
- 1.Item analysis : การวิเคราะห์ข้อสอบรายข้อ
- 2.Test analysis : การวิเคราะห์ชุดข้อสอบ



Item analysis

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



Item difficulty (p)

$$P = \frac{\text{จำนวนผู้สอบที่ตอบข้อสอบข้อนั้นถูก}}{\text{จำนวนผู้สอบที่ตอบข้อสอบข้อนั้นทั้งหมด}}$$

Activity 2 (5 min)

ผลการวิเคราะห์ข้อสอบที่ท่านได้ไป
มีข้อสอบที่ติ่มากก็ข้อ ยากมากก็ข้อ และง่ายมากก็ข้อ

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}	%	r _{pbi}	%		
-0.24	21.31	-0.10	13.52	0.37	54.92	-0.16	6.15	-0.07	4.10				

ข้อสอบที่ตีพิมพ์ จำนวน ข้อ
(P = 0.45-0.75)

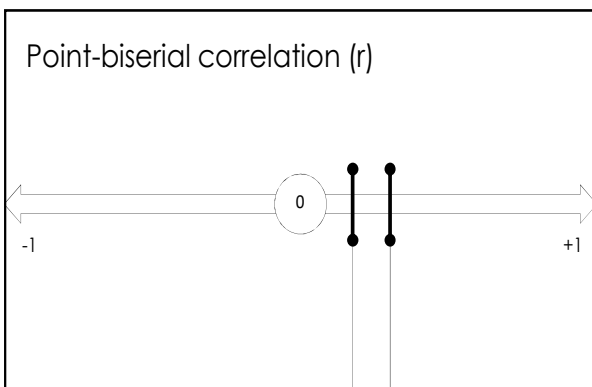
ข้อสอบง่ายมาก ได้แก่
(P ≥ 0.92)

ข้อสอบยากมาก ได้แก่
(P ≤ 0.24)

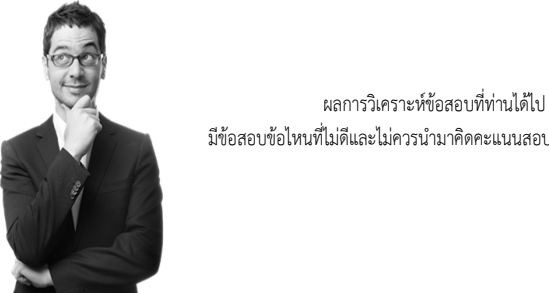
Item discrimination (r)

Point-biserial correlation (r) = $\frac{M_p - M_q}{SD} \sqrt{pq}$

M_p = คะแนนรวมเฉลี่ยของผู้สอบที่ตอบคำตอบถูก
 M_q = คะแนนรวมเฉลี่ยของผู้สอบที่ตอบคำตอบผิด
 SD = ค่าเบี่ยงเบนมาตรฐานของคะแนนสอบ
 p = สัดส่วนของผู้สอบที่ตอบข้อสอบถูกต้องผู้สอบทั้งหมด
 q = สัดส่วนของผู้สอบที่ตอบข้อสอบผิดต่อผู้สอบทั้งหมด



Activity 3 (3 min)




ผลการวิเคราะห์ข้อสอบที่ท่านได้ไป
มีข้อสอบข้อไหนที่ไม่ดีและไม่ควรนำมาคิดคะแนนสอบ

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. : 1 p Value : 0.55 r_{pbi} : 0.37

ข้อสอบที่ไม่ดี ได้แก่ (r < 0)

<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>



Distractor functionality

ตัวลงที่มีประสิทธิภาพดี มีคุณสมบัติ 2 ประการ

1. ล่อผู้สอบมาคิดกับได้ไม่น้อยกว่า ร้อยละ 5 ของผู้เข้าสอบทั้งหมด
2. ค่า point-biserial ของตัวลงนั้นเป็น ลบ

Distractor functionality

หากค่า point-biserial ของตัวลงเป็น.....

ให้สงสัยว่าข้อสอบข้อนั้นหรือมีคำตอบที่

ตัวลงที่ผู้สอบเลือกตอบน้อย หรือ ลงให้ผู้มีความรู้ดีมาเลือกเยอะ แสดงว่า.....

แนะนำให้ หรือ

Activity 4



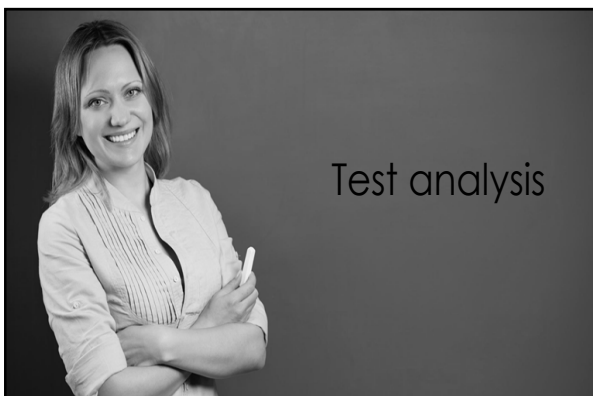
ประสิทธิภาพของตัวลงในข้อสอบต่อไปนี้เป็นอย่างไรร

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. : 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. : 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. : 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



TEST analysis

ข้อสอบทั้งชุดที่ใช้สอบเป็นอย่างไร

- น่าเชื่อถือหรือไม่
- การกระจายตัวของคะแนนเป็นอย่างไร
- ยากหรือง่ายเกินไป



TEST analysis

Internal consistency reliability
ความเที่ยงตรงของคะแนนสอบในสภาวะการหนึ่ง ๆ

Standard deviation and mean score
การกระจายตัวของคะแนนและคะแนนเฉลี่ย

Average difficulty
ความยากง่ายของชุดข้อสอบ

Average discrimination
ความสามารถในการแยกแยะผู้สอบ

Internal consistency reliability

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

เมื่อ α = สัมประสิทธิ์ อัลฟา (Coefficient Alpha)

n = จำนวนชุดย่อยของข้อสอบที่ทำการแบ่งออกเพื่อหาความเที่ยง

σ_x^2 = การกระจายตัว (variance) ของคะแนนรวม

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

Internal consistency reliability

Coefficient Alpha (α)

Standard deviation & Mean score

สะท้อนประสิทธิภาพของการเรียนการสอน


Average difficulty

Average discrimination

\bar{X} (point-biserial)



Activity 5



ผลการวิเคราะห์ข้อสอบที่ท่านได้ไป เป็นอย่างไร

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))

Application

1. ปรับแก้คะแนนสอบ
X
2. ปรับปรุงคุณภาพข้อสอบ
X
X
3. บริหารจัดการคลังข้อสอบ
4. พัฒนาคุณภาพการจัดการเรียนการสอน
X



Limitations



1. จำนวนของผู้สอบ
2. การกระจายระดับความสามารถของกลุ่มผู้สอบ
X
X
X
3. การวิเคราะห์ข้อสอบไม่ได้เป็นการตัดสินผลการสอบ

"Students can escape **BAD teaching**
but... They cannot escape **BAD assessment**"



David Boud
University of Technology, Sydney

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 AnalysisFaculty 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



โปรแกรมวิเคราะห์ข้อสอบ

รุ่น 2.0

การสอบ : SIID 521 (Basic Sciences)

วันที่ : 22 ธันวาคม 2555

จำนวนข้อสอบ = 120

จำนวนผู้เข้าสอบ = 244

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

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

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

SCORE STATISTICS

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

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

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

DIFFICULTY INDEX (p value)

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

DISCRIMINATION INDEX (D or r value)

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

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

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

การวิเคราะห์ข้อสอบปรนัย

อาจารย์ นายแพทย์เชิดศักดิ์ โสมณรัตน์

ภาควิชาศัลยศาสตร์, คณะแพทยศาสตร์ศิริราชพยาบาล, มหาวิทยาลัยมหิดล, กรุงเทพมหานคร ๑๐๓๑๐.

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

เกี่ยวข้อง และดำเนินการปรับปรุงคุณภาพของข้อสอบได้อย่างเหมาะสม

ความรู้พื้นฐานเกี่ยวกับข้อสอบปรนัย

ก่อนที่จะกล่าวถึงรายละเอียดในการวิเคราะห์ข้อสอบ ผู้นิพนธ์ก็จะขอทบทวนความรู้พื้นฐานเกี่ยวกับข้อสอบปรนัยก่อน โดยทั่วไปข้อสอบปรนัยแต่ละข้อมีส่วนประกอบสำคัญ ๒ ส่วนด้วยกันคือ

๑. โจทย์ (stem) เป็นข้อมูลของโรค หรือภาวะหรือผู้ป่วยตามด้วยคำถาม หรือเว้นช่องว่างสำหรับเติมคำหรือข้อความที่เหมาะสมลงไป

๒. ตัวเลือก (options) คือคำ 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

เวชบัณฑิตศิริราช

บทความทั่วไป

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

เนื่องจากการวิเคราะห์ข้อสอบในวงการแพทยศาสตร์ศึกษาในประเทศไทยทั้งหมดในปัจจุบันยังใช้เทคนิคต่าง ๆ ตามทฤษฎีการสอบแบบดั้งเดิมอยู่ ดังนั้นผู้นิพนธ์จะขอกล่าวถึงเทคนิคการวิเคราะห์ข้อสอบตามทฤษฎีการสอบแบบดั้งเดิมเท่านั้น เพราะจะเป็นสิ่งที่อาจารย์แพทย์ทุกท่านจะได้พบและใช้งานเป็นประจำ

การวิเคราะห์ข้อสอบตามทฤษฎีการสอบแบบดั้งเดิม

การวิเคราะห์ข้อสอบตามทฤษฎีการสอบแบบดั้งเดิมนี้อาศัยประกอบไปด้วย ๒ ส่วนใหญ่ ๆ คือ (๑) การ

วิเคราะห์ข้อสอบรายข้อ (item analysis) และ (๒) การวิเคราะห์ข้อสอบโดยรวม (test analysis)

๑. การวิเคราะห์ข้อสอบรายข้อ (item analysis)

การวิเคราะห์ข้อสอบแต่ละข้อให้อาจารย์พิจารณา ๓ ปัจจัย คือ

๑.๑ ความยากง่ายของข้อสอบ (item difficulty, p)

ความยากง่ายของข้อสอบวัดโดยใช้ค่า p ซึ่งย่อมาจาก proportion of examinees answering items correctly (สัดส่วนของผู้สอบที่ตอบข้อสอบข้อนั้นถูก) ซึ่งหาได้จากการนำจำนวนผู้สอบที่ตอบข้อสอบข้อนั้นถูกต้องหารด้วยจำนวนผู้สอบที่ตอบข้อสอบข้อนั้นทั้งหมด หากข้อสอบข้อนั้นเป็นข้อสอบที่ง่ายผู้สอบทุกคนตอบถูกค่า p ก็จะเป็น ๑ หากไม่มีผู้สอบคนใดตอบถูกเลยข้อสอบข้อนั้นก็จะมีค่า p เป็น ๐ หากมีคนตอบถูก ๗๐% ข้อสอบข้อนั้นก็จะมีค่า p เท่ากับ ๐.๗ ข้อสอบที่ดีมากจะมีค่า p อยู่ในช่วง ๐.๔๕ - ๐.๗๕, ข้อสอบที่ดีจะมีค่า p อยู่ในช่วง ๐.๗๖ - ๐.๙๑, ข้อสอบที่พอใช้ได้มีค่า p อยู่ในช่วง ๐.๒๕ - ๐.๔๔, ข้อสอบที่มีค่า p ต่ำกว่า ๐.๒๕ เป็นข้อสอบที่ยากเกินไป และข้อสอบที่มีค่า p สูงกว่า ๐.๙๑ เป็นข้อสอบที่ง่ายเกินไป^{๔-๖}

๑.๒ ความสามารถในการจำแนกผู้สอบตามระดับความสามารถ (item discrimination, r)

ความสามารถในการจำแนกผู้สอบ หมายถึงความสามารถของข้อสอบข้อหนึ่ง ๆ ในการแยกผู้สอบที่ทำคะแนนได้ดี ออกจากผู้สอบที่ทำคะแนนได้ไม่ดี ข้อสอบที่มีความสามารถในการแยกแยะได้ดีนั้นผู้สอบที่ตอบข้อสอบข้อนั้นถูกมักจะได้คะแนนสูง และผู้สอบที่ตอบข้อสอบข้อนั้นผิดมักจะได้คะแนนต่ำ ดัชนีที่ใช้วัดความสามารถในการจำแนกผู้สอบที่ใช้กันมากที่สุดในปัจจุบันคือค่า point-biserial correlation ซึ่งนิยมใช้อักษรย่อเป็น $r^{๑๔}$ ซึ่งสามารถคำนวณได้จากสูตรต่อไปนี้^๗

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

เวชบัณฑิตศิริราช

บทความทั่วไป

- เมื่อ Mp = คะแนนรวมเฉลี่ยของผู้สอบที่ตอบข้อสอบถูก
- Mq = คะแนนรวมเฉลี่ยของผู้สอบที่ตอบข้อสอบผิด
- SD = ค่าเบี่ยงเบนมาตรฐาน (standard deviation) ของคะแนนสอบ
- p = สัดส่วนของผู้สอบที่ตอบข้อสอบถูกต้องต่อผู้สอบทั้งหมด
- 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^{๕๖} (โดยทั่วไปแล้วข้อสอบที่มีค่า point-biserial ติดลบ ให้สงสัยว่าจะเฉลยผิด)

๑.๓ ประสิทธิภาพของตัวลวง (distractor functionality)

ตัวลวงที่มีประสิทธิภาพนั้นมีคุณสมบัติ ๒ ประการคือ^๖

(๑) มีผู้สอบเลือกตัวลวงนั้นไม่ต่ำกว่าร้อยละ ๕ ของจำนวนผู้สอบทั้งหมด

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

ตัวลวงใดที่มีผู้สอบเลือกน้อย หรือลวงให้ผู้ที่มี

ความรู้ดีมาเลือกจัดเป็นตัวลวงที่ไม่ดี สมควรพิจารณาตัดทิ้งหรือปรับเปลี่ยน

๒. การวิเคราะห์ข้อสอบโดยรวม (test analysis)

การวิเคราะห์ข้อสอบโดยรวมเป็นการพิจารณาว่าเมื่อข้อสอบทั้งชุดทำงานร่วมกันแล้วผลสอบที่ได้ออกมาเป็นอย่างไร มีระดับความยากง่ายเป็นอย่างไร มีการกระจายตัวของคะแนนเป็นอย่างไร มีความน่าเชื่อถือของคะแนนสอบมากน้อยเพียงใด ดัชนีต่าง ๆ ที่ต้องพิจารณาได้แก่

๒.๑ ความเที่ยงตรงของคะแนนสอบ (internal consistency reliability)

การประเมินความเที่ยงตรงของคะแนนสอบเป็นการตรวจสอบว่าคะแนนที่ได้ออกมานั้นมีความน่าเชื่อถือเพียงใด เป็นการตอบคำถามว่าหากนำผู้สอบมาสอบใหม่ในสภาวะการณ์เดิม ด้วยข้อสอบที่มีระดับความยากง่ายเท่าเดิม และผู้สอบมีความรู้เท่าเดิมไม่ได้ไปศึกษาหาความรู้เพิ่มเติม จะได้คะแนนสอบเท่าเดิมหรือไม่^{๖๖}

ดัชนีชี้วัดความเที่ยงตรงของคะแนนสอบที่นิยมใช้ในการรายงานผลสอบด้วยข้อสอบปรนัยคือค่าสัมประสิทธิ์ อัลฟา (Coefficient Alpha) ซึ่งสามารถคำนวณได้จากสูตร^{๖๖}

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

- เมื่อ α = สัมประสิทธิ์ อัลฟา (Coefficient Alpha)
- n = จำนวนชุดย่อยของข้อสอบที่ทำการแบ่งออกเพื่อหาความเที่ยง
- σ_x^2 = การกระจายตัว (variance) ของคะแนนรวม
- $\sigma_{x_i}^2$ = การกระจายตัว (variance) ของคะแนนข้อสอบย่อยชุดที่ i

ค่าสัมประสิทธิ์อัลฟานี้มีค่าอยู่ในช่วง ๐ - ๑ ค่าต่ำแสดงว่าคะแนนที่ได้มีความเชื่อถือได้น้อย ไม่แตกต่างไปจากการเดาสุ่ม ค่าสูงแสดงว่าคะแนนที่ได้นั้นมีความน่าเชื่อถือมาก หากทำการทดสอบซ้ำคะแนนที่ได้ก็จะใกล้เคียงเดิม โดยทั่วไประดับของความเที่ยงตรง

เวชบัณฑิตศิริราช

บทความทั่วไป

ของคะแนนสอบที่ยอมรับได้นั้นขึ้นกับว่าต้องการนำเอาคะแนนสอบไปใช้ทำอะไร หากการตัดสินผลสอบนั้นมีความสำคัญมาก (high-stakes examination) เช่น การตัดสินผลสอบขอรับใบประกอบวิชาชีพเวชกรรม หรือประกาศนียบัตรแพทย์ผู้เชี่ยวชาญเฉพาะสาขา มักต้องการคะแนนสอบที่มีค่าสัมประสิทธิ์ อัลฟา ไม่ต่ำกว่า ๐.๗ หากการตัดสินผลสอบนั้นมีความสำคัญปานกลาง (medium-stakes examination) เช่นการสอบลงกอนการสอบเลื่อนชั้นเรียน มักต้องการคะแนนสอบที่มีค่าสัมประสิทธิ์ อัลฟา อยู่ในช่วง ๐.๘ - ๐.๘๗ หากการตัดสินผลสอบนั้นมีความสำคัญน้อย (low-stakes examination) เช่นการสอบย่อยในชั้นเรียน การสอบแบบ formative assessment มักต้องการคะแนนสอบที่มีค่าสัมประสิทธิ์ อัลฟา อยู่ในช่วง ๐.๗ - ๐.๗๗^{๑๒}

ประเด็นสำคัญที่ต้องพิจารณาคือเมื่อได้คะแนนสอบที่มีค่าสัมประสิทธิ์ อัลฟาต่ำ จะต้องดำเนินการอย่างไรเพื่อพัฒนาให้การสอบครั้งต่อไปไม่ประสบปัญหาเรื่องความไม่น่าเชื่อถือของคะแนนสอบอีก ปัจจัยหลักที่จะช่วยเพิ่มความเที่ยงตรงของคะแนนสอบปรนัยมี ๓ ปัจจัยด้วยกัน^{๑๓} คือ

(๑) เพิ่มจำนวนข้อสอบให้มากขึ้น ยังมีข้อสอบมากข้อคะแนนที่ได้ก็จะมีค่าสัมประสิทธิ์เพิ่มขึ้น

(๒) ปรับให้ข้อสอบมีการคละกันของข้อสอบที่ยากและง่ายอย่างเหมาะสม เพื่อปรับให้คะแนนมีการกระจายตัวมากขึ้น หากข้อสอบทั้งหมดประกอบด้วยข้อสอบที่ง่ายหมด ผู้สอบเกือบทั้งหมดได้คะแนนสูงมาก จะทำให้มีความแตกต่างของคะแนนน้อย โอกาสที่จะแยกแยะผู้สอบที่มีความรู้ดีออกจากผู้ที่มีความรู้ปานกลาง หรือไม่ผู้ดีได้อย่างมั่นใจก็เป็นไปได้น้อย ดังนั้นหากอาจารย์ปรับให้มีการคละกันของข้อสอบยากและง่ายอย่างเหมาะสม ก็จะทำให้ผู้สอบมีระดับคะแนนแตกต่างกันมาก ค่าสัมประสิทธิ์อัลฟาก็จะสูงขึ้นด้วย

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

๒.๒ การกระจายตัวของคะแนน และคะแนน

เฉลี่ย (standard deviation and mean score)

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

๒.๓ ค่าความยากง่ายเฉลี่ยของข้อสอบ (average difficulty)

จากการวิเคราะห์ข้อสอบรายข้อ เราได้ค่าความยากง่ายของข้อสอบแต่ละข้อ (p) เมื่อนำค่า p ของข้อสอบทุกข้อมาหาค่าเฉลี่ย เราก็จะได้ค่าความยากง่ายของข้อสอบทั้งหมด ค่าที่ได้มานี้ใช้เป็นดัชนีชี้วัดว่าข้อสอบทั้งหมดโดยรวมแล้วมีระดับความยากง่ายเป็นอย่างไร หากผู้สอบเป็นนักศึกษาในกลุ่มใหญ่พอที่เราจะตั้งสมมติฐานว่าระดับความสามารถมีการกระจายตัวอย่างเหมาะสมและไม่ต่างจากระดับความสามารถเฉลี่ยของกลุ่มผู้สอบปีก่อน ๆ เราก็สามารถนำค่าความยากง่ายของข้อสอบทั้งหมดนี้มาเทียบได้ว่าข้อสอบที่นำมาใช้ในปีนี้อาจง่ายกว่าข้อสอบปีก่อน ๆ ซึ่งอาจารย์อาจนำข้อมูลนี้มาใช้พิจารณาปรับเกณฑ์การตัดเกรดด้วยว่าต้องมีการปรับระดับคะแนนที่ได้เกรดต่าง ๆ หรือไม่ อย่างไร

๒.๔ ค่าความสามารถในการแยกแยะผู้สอบเฉลี่ย (average discrimination)

การนำค่า point-biserial correlation ของข้อสอบทั้งหมดมาหาค่าเฉลี่ย เป็นการบอกคร่าว ๆ ว่าโดยรวมแล้วข้อสอบชุดนี้มีความสามารถในการแยกแยะผู้สอบตามระดับความสามารถเพียงใด ยิ่งได้ค่าสูงก็ยิ่งดี แต่มีข้อควรระวังในการแปลผลในกรณีที่การเรียนการสอนเป็นไปได้ดี และผู้สอบทั้งหมด หรือเกือบทั้งหมดทำคะแนนได้สูง ค่า point-biserial correlation เฉลี่ยของข้อสอบทั้งหมดจะไม่สูงแต่ไม่ได้แปลว่าข้อสอบที่ใช้มีคุณภาพไม่ดี^{๑๕}

เวชบันทึทศิธิราช

บทความทั่วไป

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 หรือเพิ่มเติมข้อมูลบางประการเข้าไปเพื่อให้การวินิจฉัย

ชัดเจนขึ้น ข้อสอบที่พบว่าง่ายเกินไปอาจพิจารณาปรับให้ยากขึ้นโดยการแก้ไขโจทย์หรือตัวเลือก ข้อสอบที่มีค่า point-biserial ต่ำมักเกิดจากโจทย์ที่คลุมเครือ สร้างความสับสนให้ผู้สอบ สมควรได้รับการปรับโจทย์คำถามใหม่

นอกจากนี้อาจารย์ยังต้องพิจารณาถึงการทำงานของตัวเลือกด้วย ปัญหาที่พบบ่อยมากในการวิเคราะห์ข้อสอบปรนัยคือมีตัวลวงจำนวนมากที่ไม่ทำงาน (มีผู้สอบเลือกน้อยมาก หรือลวงเฉพาะผู้ที่มีความรู้ดีให้มาเลือก) จากการศึกษาวิจัยข้อสอบปรนัยจำนวนมากพบว่าข้อสอบส่วนใหญ่มักมีตัวเลือกที่ทำงานจริงเพียง ๓ ตัวเลือกเท่านั้น^๔ ตัวเลือกที่เหลือเป็นตัวเลือกที่ไม่มีประโยชน์ พิมพ์ลงมาในข้อสอบก็เป็นการเปลืองเนื้อที่หน้ากระดาษ และเสียเวลาอ่านโดยใช้เหตุอาจารย์ควรพิจารณาตัดตัวลวงที่ไม่ทำงานออกเสียหรือเปลี่ยนเป็นตัวลวงอื่นที่น่าจะมีประสิทธิภาพมากขึ้น

๓. ใช้เป็นประโยชน์ในการบริหารคลังข้อสอบ

ข้อสอบแต่ละข้อนั้นได้มาด้วยความยากลำบาก อาจารย์แต่ละท่านต้องใช้เวลาและความคิดอย่างมากเพื่อพัฒนาข้อสอบที่ดีขึ้นมาใช้ ดังนั้นเมื่อนำข้อสอบมาใช้แล้วผลการวิเคราะห์ข้อสอบแสดงว่าข้อสอบข้อใดเป็นข้อสอบที่ดี มีระดับความยากง่ายเหมาะสม มีความสามารถในการจำแนกผู้สอบที่ดีก็ควรพิจารณาเลือกเก็บข้อสอบดังกล่าวไว้ในคลังข้อสอบเพื่อที่จะได้นำกลับมาใช้ใหม่ในอนาคต ในการเก็บข้อสอบเข้าในคลังข้อสอบก็ต้องมีการแนบข้อมูลเกี่ยวกับประวัติการใช้งานและผลการวิเคราะห์ข้อสอบในแต่ละครั้งไว้คู่กันด้วย เพื่อที่จะได้เป็นประโยชน์ในการเลือกข้อสอบมาใช้งาน หากอาจารย์ต้องการข้อสอบที่มีระดับความยากง่าย หรือความสามารถในการจำแนกผู้สอบมากนักเพียงใดจะได้ดึงเอาข้อสอบที่มีคุณลักษณะตามต้องการออกมาใช้ได้ตามต้องการ

๔. ใช้เป็นประโยชน์ในการพัฒนาคุณภาพการสอน

การพิจารณาผลการวิเคราะห์ข้อสอบโดยละเอียดในหัวข้อที่อาจารย์ท่านใดท่านหนึ่งรับผิดชอบ

เวชบัณฑิตศิริราช

บทความทั่วไป

ในการสอนนักเรียนหรือแพทย์ประจำบ้านอยู่นั้นจะทำให้ได้ข้อมูลที่เป็นประโยชน์ในการพัฒนาการเรียนการสอนได้ กล่าวคืออาจารย์สามารถตรวจสอบดูได้ว่านักเรียนหรือแพทย์ประจำบ้านมีความเข้าใจที่ถูกต้องในเรื่องดังกล่าวหรือไม่ ประเด็นใดที่มีผู้เข้าใจผิดอยู่มากก็สมควรที่อาจารย์จะทำการเน้นย้ำในบรรดานักเรียนหรือแพทย์ประจำบ้านในการสอนครั้งต่อไป เพื่อแก้ไขความเข้าใจผิดดังกล่าว ประเด็นใดที่นักเรียนหรือแพทย์ประจำบ้านมีความเข้าใจดีมากอยู่แล้ว อาจารย์อาจไม่ต้องใช้เวลามากนักในการสอนเรื่องดังกล่าว แต่เอาเวลาไปใช้สอนในเรื่องที่นักเรียนหรือแพทย์ประจำบ้านยังไม่ค่อยเข้าใจให้มากขึ้นได้

ข้อจำกัดของการวิเคราะห์ข้อสอบ

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

พื้นฐานสำคัญที่เป็นข้อจำกัดของผลการวิเคราะห์ข้อสอบด้วยทฤษฎีการสอบแบบดั้งเดิมคือค่าต่าง ๆ ที่ได้มาจากการวิเคราะห์นั้นขึ้นอยู่กับกลุ่มตัวอย่างที่ใช้ในการเก็บข้อมูล^{๑๑,๑๒} หากได้ข้อมูลมาจากกลุ่มตัวอย่างที่มีขนาดใหญ่พอและมีการกระจายตัวของระดับความสามารถของผู้สอบที่เหมาะสม ค่าต่าง ๆ ที่ได้ (p , r , coefficient alpha) จะค่อนข้างเที่ยงตรง ปัญหาที่สำคัญในการวิเคราะห์ข้อสอบในโรงเรียนแพทย์คือการสอบจำนวนมากจัดในนักศึกษาในกลุ่มเล็ก และ

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

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

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

ค่าสัมประสิทธิ์อัลฟา เป็นค่าที่มีความเฉพาะเจาะจงกับการสอบของนักเรียนกลุ่มใดกลุ่มหนึ่งเท่านั้น หากใช้เป็นคุณสมบัติติดตัวข้อสอบแต่ละข้อไม่ หากข้อสอบชุดหนึ่งทำการสอบกับนักเรียนกลุ่มหนึ่งแล้วพบว่าคะแนนสอบที่ได้มานั้นมีค่าสัมประสิทธิ์อัลฟาสูงในระดับที่ต้องการก็ไม่ได้เป็นตัวรับประกันว่าหากนำข้อสอบชุดเดิมนั้นไปทำการสอบกับนักเรียนกลุ่มอื่นจะได้ค่าสัมประสิทธิ์อัลฟาที่สูงเช่นเดียวกัน นอกจากนี้ค่าสัมประสิทธิ์อัลฟาที่สูงไม่ได้เป็นตัวบอกถึงคุณภาพของข้อสอบรายข้อแต่อย่างใด

ค่าสัมประสิทธิ์อัลฟาที่สูงช่วยบอกแค่เพียงว่า

เวชบัณฑิตศิริราช

บทความทั่วไป

คะแนนสอบในข้อสอบข้อหนึ่งมีความผันแปรไปในทิศทางเดียวกันกับคะแนนสอบในข้อสอบข้ออื่นในการสอบชุดเดียวกัน นั่นคือในข้อสอบชุดที่มีค่าสัมประสิทธิ์อัลฟาสูงก็อาจประกอบไปด้วยข้อสอบที่ดี และข้อสอบที่ไม่ดีรวมกันอยู่ ต้องไปตรวจสอบดัชนีชี้วัดคุณภาพของข้อสอบตัวอื่น ๆ ในแต่ละข้ออีกครั้ง

ข้อควรจำในการวิเคราะห์ข้อสอบที่ผู้นิพนธ์ข้อย้าในตอนท้ายของบทความนี้ก็คือค่าดัชนีชี้วัดคุณภาพต่าง ๆ ของข้อสอบที่กล่าวมาทั้งหมดนี้เป็นเพียงตัวช่วยให้อาจารย์เข้าใจข้อสอบดีขึ้นและช่วยแนะแนวทางในการพัฒนาปรับปรุงข้อสอบให้ดีขึ้น ดัชนีเหล่านี้ไม่ใช่ค่าตัดสินหรือตัวชี้ชะตาของข้อสอบ ไม่มีดัชนีใดที่ได้จากการวิเคราะห์ข้อสอบจะมาทดแทนดุลยพินิจของอาจารย์ไปได้ ดัชนีคุณภาพของข้อสอบไม่ว่าจะคำนวณมาด้วยวิธีการที่ถูกต้องแล้วก็ตามก็เป็นเพียงตัวเลขที่สามารถเกิดความผิดพลาดในการแปลผลได้ดังเช่นการแปลผลการวิเคราะห์ทางสถิติต่าง ๆ บทบาทของอาจารย์ในการวิเคราะห์ข้อสอบคงไม่ใช่การยึดถือตัวเลขดัชนีต่าง ๆ เป็นกฎตายตัว หากแต่ใช้ดัชนีเหล่านี้ช่วยเป็นแนวทางในการพิจารณาข้อสอบ หากดัชนีตัวใดระบุว่าข้อสอบอาจมีปัญหา อาจารย์ก็นำข้อสอบนั้นมาพิจารณากันโดยคณะกรรมการข้อสอบ หากหลังจากการพิจารณาโดยที่ถ้วนแล้วอาจารย์คิดว่าข้อสอบข้อนั้นเหมาะสมแล้ว ไม่ควรทำการปรับแก้เนื้อหา อาจารย์ก็ยืนยันไปว่าไม่แก้ไข อาจารย์คงไม่ตัดสินการรักษาสภาพผู้ป่วยโดยใช้ผลเลือดตัวใดตัวหนึ่งเป็นเกณฑ์โดยไม่พิจารณาอาการและอาการแสดงของผู้ป่วยร่วมด้วย ฉันทัดก็ฉันทัน อาจารย์

ไม่ควรตัดสินชะตากรรมของข้อสอบโดยใช้เพียงค่า p หรือ r โดยไม่พิจารณาความเหมาะสมของเนื้อหาโจทย์และตัวเลือกต่าง ๆ ในข้อสอบข้อนั้น

เอกสารอ้างอิง

- Livingston SA. Item analysis. In: Downing SM, Haladyna TM, eds. Handbook of test development. Mahwah, NJ: Lawrence Erlbaum Associates; 2006:421-41.
- Brown W, Thomson GH. The essentials of mental measurement, 2nd ed. Cambridge, England: University Press; 1921.
- Yen WM, Fitzpatrick AR. Item response theory. In: Brennan RL, ed. Educational measurement, 4th ed. Westport, CT: Praeger Publishers; 2006:111-53.
- Haladyna TM. Writing test items to evaluate higher order thinking. Boston, MA: Allyn and Bacon; 1997.
- Haladyna TM. Writing multiple choice items. Chicago, IL: CAT Inc.; 2003.
- Haladyna TM. Developing and validating multiple-choice test items, 3rd ed. Mahwah, NJ: Lawrence Erlbaum Associates; 2004.
- Aleamoni LM, Spencer RE. A comparison of biserial discrimination, point biserial discrimination, and difficulty indices in item analysis data. Educ Psychol Meas 1969;29:353-8.
- Haladyna TM, Downing SM. How many options is enough for a multiple-choice test item? Educ Psychol Meas 1993;53:999-1010.
- Gronlund NE. Assessment of student achievement, 7th ed. Boston: Allyn & Bacon, 2003.
- Linn RL, Miller MD. Measurement and assessment in teaching, 9th ed. Upper Saddle River, NJ: Prentice Hall, 2004.
- Haertel EH. Reliability. In: Brennan RL, editor. Educational measurement, 4th ed. Westport, CT: Praeger Publishers; 2006:65-110.
- Downing SM. Reliability: On the reproducibility of assessment data. Med Educ 2004;38:1006-12.
- Embretson SE, Reise SP. Item response theory for psychologists. Mahwah, NJ: Lawrence Erlbaum Associates, 2000.
- Smith EV. Evidence for the reliability of measures and validity of measure interpretation: A Rasch measurement perspective. In: Smith EV, Smith RM, eds. Introduction to Rasch measurement: Theory, models, and applications. Maple Grove, MN: JAM Press, 2004:93-112

ผศ. นพ.สุประพัฒน์ สนใจพานิชย์

หัวข้อ : Constructed response item development

Constructed Response Items

Suprath Sonjaipanich MD.

Department of Pediatrics

Faculty of Medicine Siriraj Hospital

Mahidol University

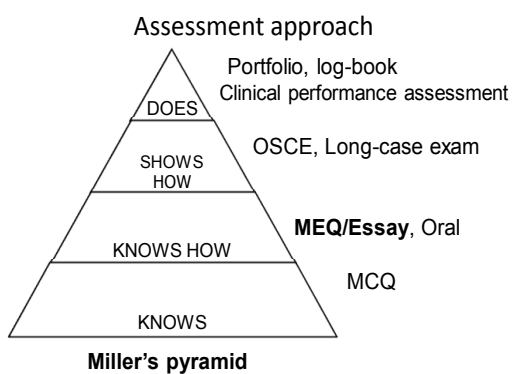
1

Written Tests

Two major types of written test forms

1. Selected Response items
2. Constructed response items

2



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.

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. International handbook of research in medical Education, part 2. Dordrecht: Kluwer, 2002, p. 647-672.

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 สำหรับนักศึกษาในระดับคลินิก

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

Modified Essay Question

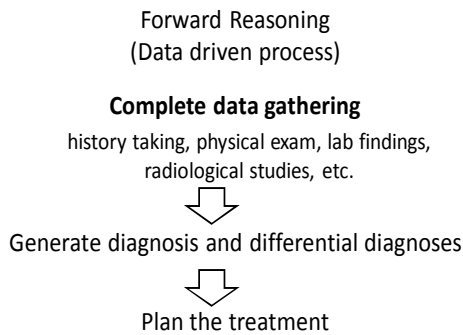
- การประยุกต์ให้เสมือนการแก้ปัญหาผู้ป่วยในชีวิตจริง
- การแก้ปัญหาของผู้ป่วยรายหนึ่ง ๆ ประกอบด้วยหลายขั้นตอน
 - จะไม่มีข้อมูลทั้งหมดตั้งแต่เริ่มเห็นผู้ป่วย
 - ต้องค่อยๆสืบค้นหาข้อมูลเพิ่มเติมและวิเคราะห์ ตัดสินใจแก้ปัญหาไปทีละขั้นตอน
 - เมื่อทำแต่ละขั้นตอนแล้ว ไม่สามารถย้อนกลับไปแก้ไขสิ่งที่ได้ทำไปก่อนหน้านี้ได้

13

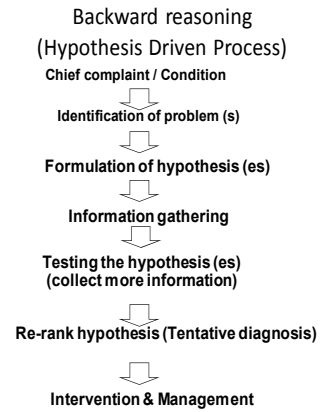
Clinical Problem Solving Methods

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

14

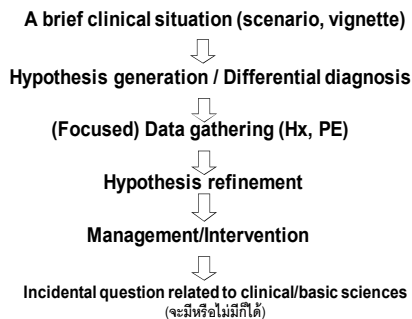


15



16

MEQ Process



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

รัชช อนุบาลศิริกุล การประเมินความรู้ในการแก้ปัญหาผู้ป่วยทางคลินิก สารศึกษา 2534, 43(2): 123 - 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)**
 - ควรตั้งคำถามให้ชัดเจน 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 decision-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; 70: 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 decision-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. ปัญหาที่นักศึกษาหรือแพทย์ประจำบ้านผิดพลาดบ่อย
 4. ปัญหาที่เกี่ยวข้องกับหลายระบบ เช่น ผู้ป่วยมีปัญหาระบบ GI 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 ไม่จำเป็นต้องจำกัดเฉพาะ biomedical บางสถานการณ์อาจเป็นเรื่อง ethical, medicolegal, prevention
 - บริหารหรือในกลุ่มผู้เชี่ยวชาญ 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 จุกเงิน 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 ให้บอกชื่อโรคที่ชื่อ

- เช่น
- จงบอกชื่อโรคที่ผู้ป่วยรายนี้น่าจะเป็นมากที่สุด 1 โรค
 - จงบอกสิ่งตรวจพบจากการตรวจร่างกายที่สำคัญที่จะช่วยในการยืนยันการวินิจฉัยโรค มา 3 ประการ
 - จงเขียนคำสั่งการรักษาสำหรับผู้ป่วยรายนี้ในคำสั่งการรักษาที่จัดให้

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 (0)
 - Harmful treatment
 - negative score (but not cross items)

40

Time

- อาจารย์ผู้ออกข้อสอบ ควรทดลองตอบคำถามด้วยตนเอง และจับเวลา หรือ ให้เพื่อนอาจารย์ทดลองทำข้อสอบ
- เวลาที่นักศึกษาใช้ในการตอบ จะมากกว่าเวลาที่อาจารย์ใช้ในการตอบคำถามนั้นๆ 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.....

สถาบัน..... จำนวนข้อสอบทั้งหมด.....ข้อ เวลาสอบรวม นาที

ข้อที่	เรื่องที่ออกข้อสอบ	จำนวนข้อสอบ	Physician tasks / Competencies													
			Problem Identification	Hypothesis generation	Data Gathering	Data Interpretation	Clinical Reasoning	Patient Management	Patient Education	Ethical analysis	Evidence-based	Basic Knowledge	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 <input checked="" type="checkbox"/>	คะแนนเต็ม
<input type="checkbox"/> Health promotion and maintenance	
<input type="checkbox"/> Mechanism of diseases	
<input type="checkbox"/> Data Gathering (Hx & PE)	
<input type="checkbox"/> Data Gathering (Investigation)	
<input type="checkbox"/> Hypothesis Generation (Differential diagnosis)	
<input type="checkbox"/> Hypothesis Refinement (Diagnosis)	
<input type="checkbox"/> Emergency management	
<input type="checkbox"/> Acute management	
<input type="checkbox"/> Long term management	
<input type="checkbox"/> Counseling education	
<input type="checkbox"/> 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

คำแนะนำในการออกข้อสอบ Modified Essay Question (MEQ)

1. เรื่องที่ออกข้อสอบควรเป็น Problem-oriented มากกว่า Disease-oriented โดยเน้นปัญหาที่พบบ่อย ปัญหาที่เกี่ยวข้องกับหลายระบบ ปัญหาที่นักศึกษาฝึกฝนได้บ่อย เป็นต้น
2. ข้อสอบควรเป็นลักษณะที่เน้นให้ผู้สอบได้มีการคิดวิเคราะห์ การใช้เหตุผล เพื่อประเมินทักษะการแก้ปัญหาและการตัดสินใจ
3. ควรมีการบูรณาการของสาขาวิชาในการออกข้อสอบบ้างตามสมควร
4. ความยากง่ายควรเหมาะสมกับระดับ พ.บ. (เป็นกลุ่มโรคที่ต้องรู้ตามเกณฑ์แพทยสภามากกว่ากลุ่มโรคที่ควรรู้หรือน่ารู้)
5. ข้อสอบที่ใช้จะมีจำนวน 10 ข้อใหญ่ ซึ่งในแต่ละข้อใหญ่จะมีข้อย่อยประมาณ 3-5 ข้อ
6. เวลาที่ใช้ในการสอบแต่ละข้อใหญ่ประมาณ 15 นาที
7. ข้อสอบแต่ละข้อ ควรเป็นไปตามลำดับขั้นตอนที่คล้ายคลึงกับสถานการณ์จริงในการดูแลผู้ป่วย เช่น บอกสถานที่ปฏิบัติงาน ในกรณีที่จำเป็น มีการให้ข้อมูลเพิ่มเติมก่อนถามคำถามในข้อต่อไป ตลอดจนมีการสรุปปัญหาหรือการวินิจฉัยเบื้องต้นก่อนการส่งตรวจ lab และการรักษา
8. ข้อมูลเริ่มต้นไม่ควรน้อยเกินไป ควรมีเพียงพอที่จะสามารถให้ผู้ตอบคิดตั้งปัญหาและวินิจฉัยแยกโรคได้พอควร เพื่อให้คิดข้อมูลที่ต้องการถามเพิ่มเติมได้โดยใช้ความคิดมากกว่าการท่องจำ และในการให้ข้อมูลการตรวจร่างกาย ไม่ควรขยักบางส่วนไว้ เพราะผู้ตอบอาจเข้าใจผิดว่าปกติ เช่น ให้ข้อมูลการตรวจร่างกายว่า chest: lungs clear แต่ต้องการให้ตอบเพิ่มเติมว่า decreased breath sound หรือให้ข้อมูลว่า abdomen: soft, normal bowel sound แต่ให้ตอบเพิ่มเติมว่า hepatosplenomegaly
9. คำถามย่อยให้ออกตาม Competency ที่เกี่ยวข้อง โดยเลือกประเด็นสำคัญ (key features) มาออกข้อสอบ เช่น ประวัติเพิ่มเติมที่สำคัญ การตรวจร่างกายที่สำคัญที่ต้องมองหา หรือตรวจเพิ่มเติม วินิจฉัยโรค หรือ การวินิจฉัยแยกโรค การสืบค้นเพิ่มเติมเพื่อ confirm หรือ exclude การวินิจฉัย การรักษาที่เฉพาะเจาะจงกับโรค รวมถึงประเด็นที่เกี่ยวข้องกับ ethical, medicolegal ไม่จำเป็นต้องออกข้อสอบต้องเริ่มต้นด้วยการถามประวัติหากต้องการประเมินความรู้ความสามารถด้านอื่นๆ
10. คำตอบควรเป็นคำตอบสั้นๆ และชัดเจน มีการกำหนดการให้คะแนนแต่ละคำตอบชัดเจนพอที่ผู้ที่ไม่ได้เป็นผู้ออกข้อสอบสามารถตรวจให้คะแนนได้ และคะแนนรวมของคำตอบแต่ละข้อไม่ควรเกินคะแนนรวมที่กำหนดไว้ในแต่ละคำถาม (รวม 100 คะแนน)
11. สามารถใช้รูปภาพประกอบได้ เช่น x-ray, EKG เป็นต้น โดยที่เป็น File (.jpg) ที่มีความละเอียดชัดเจน เนื่องจากการสอบโดยใช้คอมพิวเตอร์

อ้างอิงจากแนวทางการออกข้อสอบของ (ศ.ร.ว.)

(หนังสือเลขที่ ศ.ร.ว. ๖.139/2557 วันที่ 15 ต.ค.57)

the metric of medical education

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

ELIZABETH A FARMER¹ & GORDON PAGE²

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

¹Royal Australian College of General Practitioners, Melbourne, Victoria, Australia

²Department 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 Page² 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.⁵ 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 low³ 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 Bordage⁶ 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 Bordage⁶ 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 <i>or</i> 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

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 Devitt²Address: ¹Centre for Learning and Professional Development, University of Adelaide, Adelaide, Australia and ²Dept 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

Received: 11 April 2007

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

Accepted: 28 November 2007

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

© 2007 Palmer and Devitt; licensee BioMed Central Ltd.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/2.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.**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 curric-

ula 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

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 attach-

ment 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.

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.

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-

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

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 at 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.

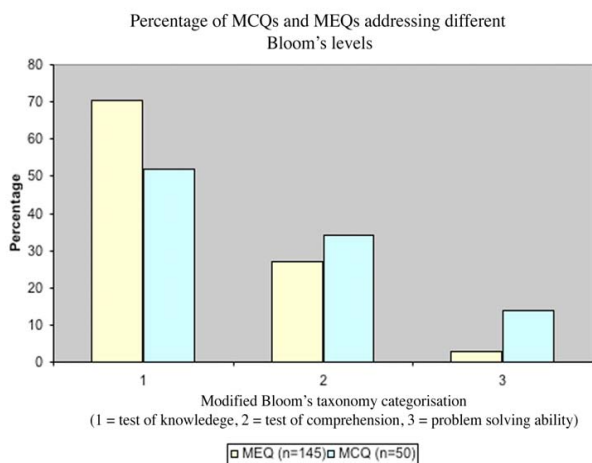


Figure 1
Percentage of MCQs and MEQs addressing different Bloom's levels of cognitive skills.

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.

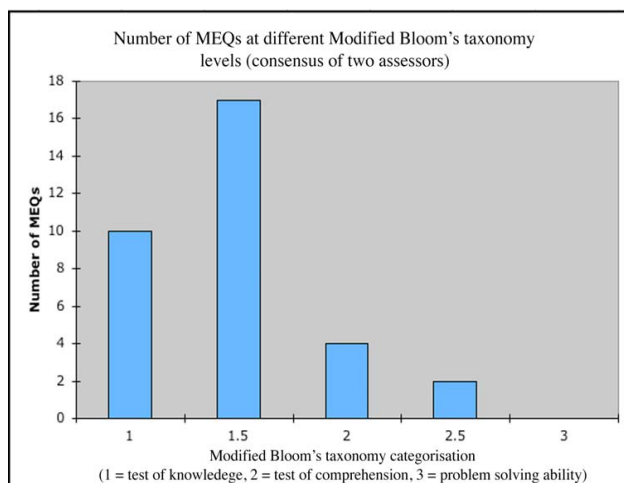


Figure 2
Number of MEQs at different Modified Bloom's taxonomy levels (consensus of two assessors).

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 HJLM: **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 medical 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>

Publish with **BioMed Central** and every scientist can read your work free of charge

"BioMed Central will be the most significant development for disseminating the results of biomedical research in our lifetime."

Sir Paul Nurse, Cancer Research UK

Your research papers will be:

- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:

http://www.biomedcentral.com/info/publishing_adv.asp



การสร้างข้อสอบอัตนัยประยุกต์

ผู้ช่วยศาสตราจารย์ นายแพทย์เชิดศักดิ์ ไธรมณีรัตน์ พ.บ., ป.ชั้นสูง (ศึกษาศาสตร์), ว.จ. 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) ซึ่งผู้ออกข้อสอบคาดหวังว่าการสอบในลักษณะที่ผู้สอบไม่มี

ตัวเลือก แต่ต้องคิดคำตอบด้วยตนเองนี้จะสามารถวัดความรู้ขั้นสูงในระดับการวิเคราะห์ สังเคราะห์ หรือประเมินคุณค่าได้^{1,2}

อย่างไรก็ตามข้อสอบในรูปแบบอัตนัยแบบดั้งเดิมนั้นประสบปัญหาในการใช้ประเมินความรู้ทางการแพทย์อยู่หลายประการ ทั้งความยากในการตรวจให้คะแนน ความจำกัดในปริมาณเนื้อหาที่สามารถสอบได้ในเวลาที่มี ความเห็นที่แตกต่างกันของผู้ตรวจให้คะแนน ความไม่เที่ยงของคะแนนสอบ เป็นต้น^{1,2} ปัญหาที่สำคัญยิ่งที่ทำให้การสอบอัตนัยแบบดั้งเดิมไม่ได้รับความนิยมในการประเมินความรู้ในระดับคลินิกคือ การที่ข้อสอบอัตนัยแบบดั้งเดิมนั้นมักวัดความรู้ในระดับการท่องจำ หรือความเข้าใจพื้นฐานเท่านั้น และรูปแบบการคิดวิเคราะห์เพื่อตอบโจทย์ข้อสอบอัตนัยแบบดั้งเดิมนั้นมีลักษณะแตกต่างไปจากกระบวนการแก้ปัญหาในระดับคลินิกที่แพทย์ปฏิบัติจริง

ข้อสอบอัตนัยแบบดั้งเดิมที่ดีนั้นผู้ออกข้อสอบสามารถประเมินทักษะการคิดวิเคราะห์ขั้นสูงได้ แต่อุปสรรคสำคัญที่ทำให้ไม่สามารถบรรลุวัตถุประสงค์ดังกล่าวได้คือการสร้างข้อสอบที่ผู้สอบตั้งใจให้ตรวจให้คะแนนได้ง่ายเป็นสำคัญ ทำให้ข้อสอบอัตนัยแบบดั้งเดิมส่วนใหญ่ทำการประเมินเพียงความรู้ระดับความจำหรือความเข้าใจพื้นฐานเท่านั้น

สมมติฐานพื้นฐานในการตอบข้อสอบอัตนัยแบบดั้งเดิมคือการวิเคราะห์และหาแนวทางแก้ปัญหาเป็นกระบวนการที่ทำในขั้นตอนเดียว ดังนั้นข้อสอบจึง

นำเสนอข้อมูลทั้งหมดในขั้นตอนเดียวแล้วให้ผู้เข้าสอบ แสดงการวิเคราะห์และแก้ปัญหา ซึ่งเป็นกระบวนการแก้ปัญหาทางคลินิกที่แพทย์ใช้ในกรณีเจอผู้ป่วยที่ไม่ซับซ้อนที่ไม่ต้องการกระบวนการคิดวิเคราะห์ที่ซับซ้อนมากนัก อย่างไรก็ตามปัญหาผู้ป่วยที่มีความซับซ้อนและต้องการวิเคราะห์มากมักต้องการกระบวนการแก้ปัญหาหลายขั้นตอน แพทย์จะต้องทำการประเมินข้อมูลพื้นฐานที่ได้จากผู้ป่วย แล้วซักประวัติ หรือตรวจร่างกายเพื่อเก็บข้อมูลเพิ่มเติมอย่างเหมาะสม เมื่อได้ข้อมูลพื้นฐานมาแล้ว แพทย์ต้องทำการตั้งสมมติฐานถึงโรคที่ผู้ป่วยน่าจะเป็น แล้วทำการสืบค้นเพิ่มเติมด้วยการตรวจทางห้องปฏิบัติการ หรือใช้ภาพถ่ายรังสี ในบางกรณีแพทย์จำเป็นต้องให้การรักษารักษาเบื้องต้นก่อน พร้อมกับทำการสืบค้นเพิ่มเติม ซึ่งเมื่อเวลาผ่านไปแพทย์จะได้รับข้อมูลของผู้ป่วยมากขึ้นเรื่อยๆ จากผลตรวจทางห้องปฏิบัติการ หรือการตอบสนองต่อการรักษาที่ให้ เมื่อได้ข้อมูลมากขึ้นแพทย์จะต้องทำการประเมินสถานการณ์ใหม่ ข้อมูลที่เพิ่มขึ้นอาจทำให้แพทย์สามารถให้การวินิจฉัยที่แน่ชัด และวางแผนการรักษาที่เหมาะสมได้ จะเห็นได้ว่ากระบวนการแก้ปัญหาของแพทย์มักทำเป็นหลายขั้นตอนหลายตอน แต่ละขั้นตอนจะได้ข้อมูลเพิ่มเติมขึ้นเรื่อยๆ การตัดสินใจในแต่ละขั้นเมื่อได้เลือกที่จะตรวจหรือให้การรักษาใดแก่ผู้ป่วยแล้ว ไม่สามารถย้อนเวลากลับไปแก้ไขการตัดสินใจที่ทำผิดพลาดไปก่อนหน้านี้ได้

จากข้อจำกัดของข้อสอบอัตนัยแบบดั้งเดิมที่กล่าวมาข้างต้น ทำให้มีการพัฒนารูปแบบการสอบเป็นข้อสอบอัตนัยประยุกต์ (modified essay question, MEQ) ซึ่งเป็นข้อสอบที่เริ่มจากการให้สถานการณ์ของผู้ป่วย แล้วมีโจทย์ถามให้ผู้สอบตอบคำถามที่เกี่ยวกับการแก้ปัญหาผู้ป่วยในสถานการณ์นั้นโดยไม่มีตัวเลือกให้เมื่อผู้สอบตอบคำถามแล้วจะมีการเปิดเผยข้อมูลเพิ่มเติมเกี่ยวกับผู้ป่วยมากขึ้นทีละน้อย และมีโจทย์ถามคำถามเพิ่มเติมเป็นลำดับ โดยที่ผู้สอบไม่มีโอกาสย้อนกลับไปแก้ไขคำตอบของตนเองที่ได้ตอบไปในขั้นตอนก่อนหน้านี้^{1,3} รูปแบบของข้อสอบอัตนัยประยุกต์ที่นิยมใช้กันมากในยุคแรกๆ มีลักษณะเป็นการสอบถามกระบวนการดูแลผู้ป่วยตั้งแต่ต้นจนจบในรูปแบบที่เรียกว่าการจัดการ

ปัญหาของผู้ป่วย (Patient management problem, PMP)^{1,4,5}

เนื่องจากข้อสอบอัตนัยประยุกต์ที่ใช้ในการแพทย์มักมุ่งเน้นการประเมินทักษะการวินิจฉัยโรค 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 และตรวจร่างกายอย่างมีจุดหมายเพื่อทดสอบสมมติฐานที่ตั้งขึ้นจนค่อยๆ ตัดโรคที่ไม่สอดคล้องกับข้อมูลที่ได้รับออกไปเรื่อยๆ โดยทั่วไปแล้ววิธีอุปนัยเป็นวิธีที่มีประสิทธิภาพน้อยกว่าวิธีนิรนัย เนื่องจากการเก็บข้อมูลเป็นไปอย่างขาดจุดหมายทำให้เสียเวลาและอาจพลาดการเก็บข้อมูลที่สำคัญไป⁶

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

เวบบ์ทีกีธีรธา

บทความทัวโอ

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

พัฒนาการของข้อสอบอัตนัยประยุกต์

หลังจากที่มีรายงานการใช้ข้อสอบอัตนัยประยุกต์ในการประเมินผลทางแพทยศาสตรศึกษาตั้งแต่ปี พ.ศ. 2514 โดยราชวิทยาลัยแพทย์เวชปฏิบัติทั่วไปเพื่อประเมินทักษะการแก้ปัญหาทางคลินิกแล้ว^{3,7,8} ข้อสอบอัตนัยประยุกต์ก็ได้ถูกใช้ในการประเมินทางการแพทย์และสาธารณสุขในหลากหลายบริบท⁹⁻¹² โดยรูปแบบที่เป็นที่นิยมกันมากเป็นการสอบถามกระบวนการดูแลผู้ป่วยตั้งแต่ต้นจนจบในรูปแบบ การจัดการปัญหาของผู้ป่วย (Patient management problem, PMP) ซึ่งการแก้ปัญหาผู้ป่วยแต่ละรายมักใช้เวลานานมาก ทำให้การสอบแต่ละครั้งมักมีจำนวนสถานการณ์ผู้ป่วยที่นำมาสอบไม่มากนัก¹³

จากการใช้ข้อสอบอัตนัยประยุกต์ในรูปแบบการจัดการปัญหาของผู้ป่วย พบว่ามีข้อจำกัดบางประการ กล่าวคือ ข้อสอบส่วนใหญ่มุ่งเน้นวัดความครบถ้วนสมบูรณ์ของคำตอบมากกว่าการตัดสินใจแก้ปัญหา จำนวนสถานการณ์ผู้ป่วยที่มีจำนวนน้อยทำให้ไม่สามารถครอบคลุมองค์ความรู้ที่ต้องการประเมินได้ครบ และความ

เที่ยงของคะแนนสอบที่ต่ำ^{4,13,14} ปัญหาที่สำคัญยิ่งในการสอบด้วยสถานการณ์ผู้ป่วยจำนวนน้อยคือ ทักษะในการแก้ปัญหาทางคลินิกมีความจำเพาะต่อบริบทของผู้ป่วยแต่ละราย (case specificity)¹⁵⁻¹⁸ การที่ผู้เข้าสอบสามารถแก้ปัญหาผู้ป่วยที่มีอาการเจ็บหน้าอกได้นั้นไม่สามารถจะบอกได้ว่าผู้เข้าสอบคนดังกล่าวจะสามารถแก้ปัญหาผู้ป่วยที่มีอาการปวดศีรษะได้ดีด้วยหรือไม่ ดังนั้นหลักการที่สำคัญประการหนึ่งในการสร้างข้อสอบอัตนัยประยุกต์ก็คือการจัดทำข้อสอบให้มีหลากหลายสถานการณ์ เพื่อให้สามารถประเมินการแก้ปัญหาของผู้เข้าสอบได้ในหลากหลายบริบท ในหลายระบบอวัยวะ จากปัญหาในการใช้ข้อสอบอัตนัยประยุกต์ต่าง ๆ เหล่านี้ ทำให้นักการศึกษาได้มีการพัฒนารูปแบบข้อสอบอัตนัยประยุกต์ให้ต่างไปจากรูปแบบดั้งเดิม รูปแบบข้อสอบที่ผู้เชี่ยวชาญในการประเมินผลแนะนำในปัจจุบันคือ การแก้ปัญหาสำคัญ (key features problems, KFP)

ข้อสอบอัตนัยประยุกต์แบบแก้ปัญหาสำคัญนี้ได้รับการพัฒนาบนหลักการสำคัญคือในการแก้ปัญหาผู้ป่วยแต่ละรายมีประเด็นปัญหาที่เป็นหัวใจสำคัญเพียงไม่กี่ประเด็นเท่านั้น ซึ่งประเด็นปัญหาเหล่านี้เรียกว่า ปัญหาสำคัญ (key features)¹⁹ ซึ่งในผู้ป่วยแต่ละรายจะมีปัญหาสำคัญที่แพทย์ต้องให้ความสนใจต่างกันไป บางรายเป็นเรื่องการซักประวัติ บางรายเป็นการเลือกการส่งตรวจทางห้องปฏิบัติการ ในขณะที่บางรายเป็นการตัดสินใจเลือกวิธีการรักษาที่เหมาะสม เป็นต้น ในข้อสอบอัตนัยประยุกต์รูปแบบการแก้ปัญหาสำคัญจะมุ่งเน้นตั้งใจถามเฉพาะประเด็นปัญหาสำคัญเหล่านี้เท่านั้น ไม่จำเป็นต้องถามกระบวนการดูแลผู้ป่วยตั้งแต่ต้นจนจบในผู้ป่วยทุกราย การสร้างข้อสอบอัตนัยประยุกต์ในลักษณะนี้ทำให้ผู้สอบใช้เวลาในการแก้ปัญหาผู้ป่วยแต่ละรายไม่มากนัก และสามารถประเมินทักษะการแก้ปัญหาได้ในหลากหลายสถานการณ์ คะแนนสอบที่ได้จึงมีความเที่ยงสูง มีรายงานค่าความเที่ยงของคะแนนสอบถึง 0.8 ในการสอบอัตนัยประยุกต์แบบแก้ปัญหาสำคัญเป็นเวลาสี่ชั่วโมง¹⁴

เวบบันทึทศึรึรึรึ

บทความหัวโ

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 จึงเป็นไปได้ที่จะทำให้สถานการณ์ที่เป็นปัญหาทางคลินิกทุกอย่างจะมาปรากฏอยู่ในชุดข้อสอบ ดังนั้นการเลือกปัญหาทางคลินิกที่จะทำการสอบจึงต้องทำอย่างเป็นระบบ ควรมีการจัดทำตารางกำหนดลักษณะข้อสอบที่ชัดเจนว่าในการสอบครั้งหนึ่ง ๆ จะมีข้อสอบกี่ข้อ จะประเมินความรู้ในระบบอวัยวะใด และจัดสรรให้ข้อสอบไม่ซ้ำซ้อนกัน (ไม่ควรมีข้อสอบสองข้อถามความรู้ในระบบอวัยวะเดียวกัน ในขณะที่บางระบบอวัยวะไม่มีข้อสอบเลย)

ลักษณะปัญหาทางคลินิกที่ควรเลือกมาสอบด้วยข้อสอบอัตนัยประยุกต์ ได้แก่

- ปัญหาที่พบได้บ่อยในเวชปฏิบัติ
- ปัญหาที่แพทย์เกิดความผิดพลาดในการดูแลผู้ป่วยค่อนข้างบ่อย
- ปัญหาที่ยังไม่สามารถวินิจฉัยสาเหตุได้ชัดเจน
- ปัญหาที่มีความเกี่ยวข้องกับหลายระบบ

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

3. กำหนดปัญหาสำคัญ

เมื่อทีมคณาจารย์เลือกปัญหาทางคลินิกที่จะทำการสอบแล้ว คณาจารย์ต้องตั้งคำถามว่าขั้นตอนใดในการดูแลผู้ป่วยที่มีปัญหาดังกล่าวจัดเป็นขั้นตอนสำคัญที่สุดในการจัดการปัญหานั้น ซึ่งขั้นตอนดังกล่าวจะได้รับการกำหนดให้เป็น ปัญหาสำคัญของสถานการณ์ผู้ป่วยที่จะใช้สอบ ในบางกรณีที่มีทีมคณาจารย์ไม่สามารถเลือกขั้นตอนสำคัญในปัญหาทางคลินิกนั้น ๆ จากวิธีดังกล่าวได้

เขตนทกคทรธ

บทควมทวอ

อจจใช้ค้คำถมว่ข้ันตอนใดนการดูแลผู้ป่วยทม่ปัญหา ดงกล่วเป็นข้ันตอนท่นกคศึกษาแพทยหรือแพทยประจ้ บ้านทำผดพลดมกท่สุด⁴

ม่ข้อแ่นน้าสองประการส้หรับการกำหนด ปัญหาส้คัญนแต่ละสถานการณ้ได้แก่

- ส่ท่ต้องตดสนใจนผู้ป่วยม่เป็นส่ท่ถูกต้อง และควรปฏิบัติอจจม่ได้เป็นข้ันตอนส้คัญท่จะต้งน้ มาสอบเสมอไป การปฏิบัติตผู้ป่วยหลายอย่างท่ทำกัน เป็นปกติ โดยม่ต้องคดวิเคราะห์ เป็นข้ันตอนท่ม่ค้อย ทำผดพลด มกม่ใช่ปัญหาส้คัญนสถานการณ้นั้น

- ปัญหาส้คัญม่จ้ก้ค้อยเฉพาะประเด้นปัญหา ทาง ชีววิทยาการแพทย (biomedical) เท่านั้น นบาง สถานการณ้ปัญหาส้คัญอจจเป็นประเด้นทางจรรยาภม ฎหมาย หรือ การส่งเสริมสุขภาพและบ้องกันโรคก้ได้

4. เขยนจทย์ค้ถม

เม่อมีสถานการณ้ผู้ป่วยและข้ันตอนท่เป็นปัญหา ส้คัญนสถานการณ้นั้นแล้ว ท่มคณาจารย์ต้องเขยน จทย์ค้ถมท่ม่ความชัดเจน เพ่อประเมินว่ผู้เข้สอบม่ ความสมรถนการตดสนใจนการแก้ปัญหส้คัญน สถานการณ้ดงกล่วหรือม่ โดยท่วไปแล้วล้ขณะจทย์ ค้ถมท่ใช้บ้อยนข้อสอบอ้ตนัยประยุคต์ได้แก่

- จงสอบถามประวัติท่ม่ส้คัญเพิ่มเติม
- จงบอกการตรวจร่างกายท่ม่ส้คัญท่ต้องมอหา (หรือตรวจเพิ่มเติม) นผู้ป่วย
- จงให้การวจนจจย (หรือ การวจนจจยแยกโรค)
- จงส่งการตรวจค้้นเพิ่มเติมเพ่อให้การวจนจจยโรค
- จงส่งการรักษาท่ม่เหมาะสมม่ผู้ป่วย

โดยท่วไปแล้วสถานการณ้ผู้ป่วยหนึ่ง ๆ ควรมี ค้ถมราว 2 – 3 ข้อ แต่ละข้อประเมินความสามารถน การจ้จัดการกับปัญหาส้คัญ 1 ประเด้น^{4,21} นการเขยน จทย์ค้ถมแต่ละข้อนั้นแ่นน้ให้ม่มีการกำหนดจ้นวน ค้ตอบท่ม่สามารถตอบได้ไว้ด้วย เช่น

- จงบอกชื่อโรคท่ม่ผู้ป่วยรายน้่นน่าจะเป็นมกท่ม่สุด 1 โรค
- จงบอกผลการตรวจร่างกายท่ม่ส้คัญท่ม่จะช่วย ย่นย่นการวจนจจยโรคมา 3 ประการ

- จงระบุการตรวจเพิ่มเติมทางห้องปฏิบัติการท่ม่ จะช่วยนการวจนจจยโรค 1 การตรวจ

การกำหนดจ้นวนค้ตอบน้จะทำม่ให้ผู้เข้สอบ ต้องเลือกล่ท่ม่ถูกต้องเหมาะสมท่ม่สุดเท่านั้นมาเขยนตอบ หากผู้เข้สอบเขยนค้ตอบเกินจ้นวนท่ม่กำหนด อจจจย ผู้ตรวจข้อสอบจะม่อ่านค้ตอบท่ม่เกินมา การปฏิบัติเช่นน้ จะช่วยจ้จัดปัญหาการตรวจกระดาษค้ตอบท่ม่ผู้เข้สอบ เขยนค้ตอบแบบหว่านแห ให้ครอบคลุมทุกอย่โดยท่ม่ ผู้เข้สอบเองม่ม่ความรู้ ความเข้ใจว่จ้ล่ใดเป็นประเด้น ส้คัญนการดูแลผู้ป่วยนข้ันตอนนั้น ๆ

เม่อทำการเขยนจทย์ค้ถมและจ้นวนค้ตอบ ท่ต้องการแล้ว ให้อจจจยระบุเวลาท่ม่ใช้ในการตอบค้ถม ตอนนั้นด้วย เนื่องจกข้อสอบอ้ตนัยประยุคต์ม่มีการดำเนิน ของสถานการณ้ผู้ป่วยท่ม่กำหนดให้โดยม่มีการให้ข้อมลท่ม่ละ ส่วน ผู้เข้สอบจ้เป็นท่ม่จะต้งรู้เวลาท่ม่ม่ในการทำข้อสอบ แต่ละต่อนก่อนท่ม่จะต้งส่งค้ตอบและสถานการณ้ผู้ป่วย ดำเนินต้อไป นการกำหนดเวลาในการทำข้อสอบแต่ละ ตอนให้อจจจยผู้ออกข้อสอบพิจารณาจกท้งเวลาท่ม่ ต้องใช้ในการอ่าน และเวลาท่ม่ต้องใช้ในการเขยนค้ตอบ นข้อสอบตอนท่ม่ต้องอ่านเนื่อหาจทย์มก หรือต้องเขยน ค้ตอบหลายบรรท้ด ควรต้งม่มีการให้เวลาในการทำ ข้อสอบมกพอ หากเป็นไปได้ควรได้ม่มีการลองทำการ อ่านจทย์และเขยนค้ตอบโดยตัวอจจจยผู้ออกข้อสอบ เองหรือเพ่อนอจจจยแล้วลองจับเวลาที่อจจจยใช้ในการ ทำข้อสอบตอนนั้น ๆ เวลาท่ม่ได้จะเป็นเวลาที่ผู้เข้วชชาญใช้ แก่ปัญหาผู้ป่วยนสถานการณ้ดงกล่ว หากให้นค้ศึกษา ทำ ควรเพิ่มเวลาให้ร้อยละ 30 – 50 ของเวลาที่อจจจยใช้

5. กำหนดเกณฑ์การให้คะแนน

ข้ันตอนสุดท้ายนการสร้างข้อสอบอ้ตนัย ประยุคต์คือการกำหนดเกณฑ์การให้คะแนน ซ่เป็นข้ัน ตอนท่ม่ม่ความทำทาย และสร้างควมล้บากใจให้แก่ อจจจยผู้ออกข้อสอบหลายท่น เนื่องด้วยเกรงว่จะเฉลย ค้ตอบม่ครอบคลุมล่ท่ม่ผู้เข้สอบจะเขยนตอบมา หรือ เกิดความม่เป็นธรรมข้ัน นนท่ม่ผู้นพนธ์ขอเสนอแ่นน้ แนว ทางนการกำหนดเกณฑ์ให้คะแนนดงน้

- แ่นน้ให้กำหนดคะแนนเต็มนการแก้ปัญห

เวชบันทึกศิรราช

บทความทั่วไป

สถานการณ์หนึ่ง ๆ เป็น 100 คะแนน เท่ากันในทุกสถานการณ์ เพื่อให้ไม่ต้องทำการปรับคะแนนสอบหลังการตรวจข้อสอบ

- กรณีที่มีคำตอบที่ถูกต้องยอมรับได้เพียงคำตอบเดียว เช่น ข้อมูลจากโจทย์มีความชัดเจนว่าผู้ป่วยเป็นโรคอะไร แล้วโจทย์ให้ผู้เข้าสอบตอบชื่อโรค หากผู้เข้าสอบตอบตรงตามเฉลยที่ตั้งไว้ให้ได้คะแนนเต็ม หากตอบอื่นนอกจากนั้นไม่ได้คะแนน

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

- คำตอบบางลักษณะมีการเขียนเนื้อหาที่มีความครบถ้วนสมบูรณ์แตกต่างกันได้ การกำหนดเกณฑ์สามารถกำหนดให้คำตอบที่มีความสมบูรณ์ได้คะแนนเต็ม ส่วนคำตอบที่ไม่สมบูรณ์จะได้คะแนนลดลงไปตามความเหมาะสม (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 ไม่สมควรปฏิบัติแก่ผู้ป่วยโดยทั่วไปแล้วพิจารณาไม่ให้เป็นคะแนน ซึ่งก็จัดเป็นการทำโทษในระดับหนึ่งแล้ว เพราะผู้สอบมีสิทธิเขียนคำตอบได้จำนวนจำกัด การที่ไม่ให้คะแนนในคำตอบที่ไม่เหมาะสม ก็จะทำให้คะแนนสูงสุดที่ผู้สอบจะทำได้ลดลงไปแล้ว การปฏิบัติที่ไม่ถูกต้องที่มีผลเสียรุนแรงต่อผู้ป่วยเท่านั้นที่ควรพิจารณาให้คะแนนติดลบ และแม้มีการให้คะแนนติดลบก็ไม่ควรมีการติดลบข้ามไปถึงข้อสอบข้ออื่นในชุดข้อสอบนั้น

- การกำหนดเกณฑ์การให้คะแนน ไม่ควรใช้อาจารย์ท่านเดียวในการกำหนด เพราะมักได้คำตอบที่ไม่ครอบคลุม ควรใช้ทีมคณาจารย์หลายท่านช่วยกันคิด คำตอบที่ผู้เข้าสอบอาจจะตอบได้ในสถานการณ์ดังกล่าว ซึ่งจะได้เกณฑ์การให้คะแนนที่สมบูรณ์กว่า อย่างไรก็ตาม ถึงแม้ว่าจะใช้คณาจารย์หลายท่านช่วยกันคิดคำตอบแล้วก็ตาม จะพบว่าในการตรวจข้อสอบอัตโนมัติประยุกต์หลายครั้ง จะพบคำตอบที่ผู้เข้าสอบตอบมาที่ น่าจะได้คะแนนแต่อาจารย์ผู้ออกข้อสอบไม่ได้กำหนดเกณฑ์คะแนนไว้ล่วงหน้าอยู่ประปราย ดังนั้นในการนำข้อสอบอัตโนมัติประยุกต์ที่สร้างขึ้นใหม่มาใช้ในการสอบ 2-3 รอบแรกแนะนำให้อาจารย์ผู้ออกข้อสอบและมีความเชี่ยวชาญชำนาญในการดูแลผู้ป่วยในสถานการณ์นั้น ๆ เป็นผู้ทำการตรวจข้อสอบ เพื่อให้สามารถพิจารณาได้ว่าคำตอบใดที่น่าจะเพิ่มเข้าไปในเกณฑ์การให้คะแนนด้วย ซึ่งเมื่อทำไป 2-3 รอบการสอบแล้วมักจะได้เกณฑ์การให้คะแนนที่มีความครอบคลุมคำตอบที่ผู้สอบจะตอบมาได้ทั้งหมด แล้วจึงมอบหมายให้อาจารย์ท่านอื่นช่วยตรวจให้คะแนนข้อสอบต่อไป

เมื่อทำการกำหนดเกณฑ์การให้คะแนนในข้อสอบเสร็จทุกข้อย่อยแล้วกระบวนการขั้นตอนสุดท้ายในการสร้างข้อสอบอัตโนมัติประยุกต์คือการกำหนดเกณฑ์ผ่านของโจทย์สถานการณ์นั้น กล่าวคือจากคะแนนเต็ม 100 คะแนน ผู้สอบต้องทำคะแนนได้อย่างน้อยที่สุดกี่คะแนนจึงจะจัดว่าสอบผ่านในการแก้ปัญหาสถานการณ์นั้น ๆ วิธีการตั้งเกณฑ์ผ่านทำได้หลายวิธี แต่วิธีที่เป็นที่นิยมมากที่สุดสำหรับข้อสอบอัตโนมัติประยุกต์ และเป็นวิธีที่คณะแพทยศาสตร์ศิริราชพยาบาลใช้เป็นประจำในการตัดสินผลสอบอัตโนมัติประยุกต์คือวิธี Modified Angoff ซึ่งมีขั้นตอนที่สำคัญสามขั้นตอนคือ

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

(2) ไล่ดูโจทย์คำถามทีละข้อพร้อมเฉลย แล้วทำสัญลักษณ์ * ไว้ในคำตอบที่คาดว่าผู้ที่มีความรู้ ความสามารถคาบเส้นจะตอบในข้อสอบแต่ละตอน

(3) ทำการรวมค่าคะแนนที่ได้รับการทำสัญลักษณ์ * ไว้ตั้งแต่ข้อแรกจนถึงข้อสุดท้าย จะได้คะแนนเกณฑ์ผ่านในการแก้ปัญหาสถานการณ์นั้น ๆ²²

แนวทางการพัฒนาข้อสอบอัตนัยประยุกต์ในคณะแพทยศาสตร์ศิริราชพยาบาล

คณะแพทยศาสตร์ศิริราชพยาบาลมีการใช้ข้อสอบอัตนัยประยุกต์ในการประเมินความรู้ของนักศึกษาแพทย์ชั้นคลินิกมานานแล้ว โดยเริ่มต้นจากการสอบของแต่ละภาควิชา และต่อมาเมื่อศูนย์ประเมินและรับรองความรู้ความสามารถในการประกอบวิชาชีพเวชกรรมกำหนดให้การสอบอัตนัยประยุกต์เป็นส่วนหนึ่งของการประเมินขั้นตอนที่ 3 ในการขอใบประกอบวิชาชีพเวชกรรมตั้งแต่ปีการศึกษา 2550 ทางคณะแพทยศาสตร์ศิริราชพยาบาลก็ได้มีการจัดสอบประมวลความรู้ทางการแพทย์สหสาขาวิชา ด้วยข้อสอบอัตนัยประยุกต์ (comprehensive MEQ examination) ในนักศึกษาแพทย์ปีที่ 6 อย่างต่อเนื่อง ตลอดช่วงเวลาที่มีการใช้ข้อสอบอัตนัยประยุกต์ในคณะฯ ได้มีการพัฒนาข้อสอบประเภทนี้อย่างต่อเนื่อง จากเดิมเคยจัดสอบข้อสอบอัตนัยประยุกต์ในรูปแบบข้อสอบกระดาษ จนพัฒนาให้จัดสอบอัตนัยประยุกต์ด้วยการนำเสนอข้อมูลผู้ป่วยบนจอภาพคอมพิวเตอร์ ร่วมกับการเขียนคำตอบในกระดาษคำตอบ ตั้งแต่ปีการศึกษา 2552 จนถึงปัจจุบัน แต่ถึงแม้ว่าฝ่ายการศึกษาจะมีการพัฒนาระบบจัดสอบข้อสอบอัตนัยประยุกต์ให้มีประสิทธิภาพมากขึ้น อำนวยความสะดวกให้ผู้เข้าสอบมากขึ้น และเพิ่มความพึงพอใจในประสบการณ์การสอบขึ้นอย่างต่อเนื่อง จากการเก็บรวบรวมข้อมูลการวิเคราะห์ข้อสอบ วิเคราะห์คะแนน และแบบสำรวจความพึงพอใจของผู้สอบที่ผ่านมาผู้นิพนธ์มีความเห็นว่าการจัดสอบประมวลความรู้ทางการแพทย์ด้วยข้อสอบอัตนัยประยุกต์ของนักศึกษาแพทย์ยังสามารถพัฒนาให้มีคุณภาพดีขึ้นได้อีกในหลายด้าน ดังนี้

(1) เนื้อหาข้อสอบ

ข้อสอบอัตนัยประยุกต์ที่ใช้ในการสอบประมวลความรู้ทางการแพทย์ของคณะแพทยศาสตร์ศิริราชพยาบาลที่ผ่านมามีหลายข้อเป็นเนื้อหาวิชาที่ยากและมีความรู้ลึกในระดับผู้เชี่ยวชาญเฉพาะทาง แนวทางการ

พัฒนาการสอบอัตนัยประยุกต์อันดับแรกคือการพัฒนาเนื้อหาให้เหมาะสมกับการประเมินความรู้ของแพทย์เวชปฏิบัติทั่วไป

เนื้อหาข้อสอบอัตนัยประยุกต์สำหรับการสอบประมวลความรู้ไม่ควรมุ่งเน้นเนื้อหาที่เป็นสหสาขาวิชา กล่าวคือต้องอาศัยองค์ความรู้ที่นักศึกษาได้ศึกษาจากหลายภาควิชามาช่วยกันแก้ปัญหาผู้ป่วย ข้อสอบอัตนัยประยุกต์ที่นำมาสอบนักศึกษาแพทย์ทุกข้อในปัจจุบันล้วนมีความเป็นสหสาขาวิชาทั้งสิ้น มีอาจารย์จากหลากหลายภาควิชามาร่วมกันออกข้อสอบ แต่อย่างไรก็ตามข้อสอบบางข้ออาจมีลักษณะการใช้ความรู้สหสาขาวิชาแบบแยกเป็นส่วน ๆ กล่าวคืออาจารย์ต่างภาควิชากันใช้การแบ่งงานออกเป็นส่วน ๆ อาจารย์ภาควิชาที่หนึ่งออกข้อสอบในตอนหนึ่งกับสอง อาจารย์ภาควิชาที่สองออกข้อสอบในตอนสามกับสี่ และอาจารย์ภาควิชาที่สามออกข้อสอบในตอนห้ากับหก ข้อสอบลักษณะนี้มักจะยากมาก เนื่องจากเป็นการใช้ความรู้เชิงลึกของแต่ละภาควิชาที่ละเรื่อง เช่น ชักประวัติ ตรวจร่างกายแล้วก็ไม่สามารถวินิจฉัยโรคได้ ต้องส่งต่อไปทำการตรวจเพิ่มเติมในอีกภาควิชาหนึ่ง ซึ่งผลการตรวจเพิ่มเติมก็แปลผลได้ยาก เมื่อได้ข้อสรุปแล้วก็ต้องส่งต่อไปให้แพทย์อีกสาขาวิชาหนึ่งทำการรักษา เมื่อรักษาแล้วก็มีภาวะแทรกซ้อนต้องส่งต่อให้แพทย์อีกสาขาวิชาหนึ่งทำการแก้ไขภาวะแทรกซ้อนให้ เป็นต้น โดยทั่วไปแล้วข้อสอบอัตนัยประยุกต์ที่ใช้ความรู้สหสาขาวิชาที่เป็นที่ต้องการในการสอบประมวลความรู้ไม่ควรเป็นการประเมินความรู้ในเชิงลึกที่ละวิชาในข้อสอบแต่ละตอน แต่ควรเป็นการผสมผสานความรู้จากหลากหลายสาขาวิชาในทุกขั้นตอน เช่น หญิงอายุ 30 ปี ปวดท้องน้อยตื้อ ๆ ตลอดเวลา 6 ชั่วโมง มีไข้ต่ำ ๆ คลื่นไส้เล็กน้อย โจทย์ให้ผู้สอบซักประวัติเพื่อการวินิจฉัยโรคซึ่งผู้สอบที่จะตอบคำถามได้ดีต้องอาศัยความรู้ทั้งโรคในระบบทางเดินอาหาร ทางเดินปัสสาวะ ภาวะสืบพันธุ์สตรี กระดูกและกล้ามเนื้อ เป็นต้น

ข้อแนะนำในเรื่องเนื้อหาที่สำคัญคืออาจารย์ผู้ออกข้อสอบต้องตระหนักว่าการสอบนี้เป็นการประเมินความรู้เวชปฏิบัติทั่วไป มิใช่การประเมินความรู้เชิงลึกในศาสตร์ของแต่ละสาขาวิชา โรคหรือภาวะที่นำมาออก

ข้อสอบส่วนใหญควรอยู่ในเกณฑ์มาตรฐานผู้ประกอบวิชาชีพเวชกรรมในกลุ่มที่ 1 หรือ 2 (โรคหรือภาวะที่แพทย์เวชปฏิบัติทั่วไปสามารถให้การดูแลด้วยตนเองได้ และพิจารณาส่งต่อในกรณีที่โรครุนแรงหรือซับซ้อน) โรคหรือภาวะที่อยู่ในเกณฑ์มาตรฐานฯ กลุ่มที่ 3 (โรคหรือภาวะที่แพทย์เวชปฏิบัติทำการดูแลเบื้องต้นแล้วให้ส่งต่อไปยังผู้เชี่ยวชาญ) ควรนำมาออกข้อสอบไม่มากนัก หากจะนำโรคหรือภาวะในเกณฑ์มาตรฐานฯ กลุ่มที่ 3 มาออกสอบ ต้องมุ่งเน้นการดูแลรักษาเบื้องต้นที่แพทย์เวชปฏิบัติทั่วไปพึงทำได้ ไม่ควรมุ่งประเด็นไปที่การรักษาโดยผู้เชี่ยวชาญเฉพาะสาขามากเกินไป

(2) รูปแบบคำถาม

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

(3) จำนวนสถานการณ์ผู้ป่วยที่ใช้สอบ

ในการสอบประมวลความรู้ด้วยข้อสอบอัตนัยประยุกต์ของคณะแพทยศาสตร์ศิริราชพยาบาลที่ผ่านมามีการใช้สถานการณ์ผู้ป่วยในข้อสอบตั้งแต่ 5 ถึง 8 ราย ถึงแม้ว่าจำนวนสถานการณ์ในการสอบระยะหลังมี

แนวโน้มเพิ่มขึ้น แต่หากพิจารณาในแง่ของความจำเพาะต่อบริบทของผู้ป่วย (case specificity) ที่ได้อภิปรายไปก่อนหน้านี้แล้วจะเห็นได้ว่าการที่ผู้สอบแก้ปัญหาผู้ป่วยได้ 5 ถึง 8 รายนี้ น่าจะยังคงครอบคลุมประเด็นปัญหาทางคลินิกได้ไม่มากเพียงพอ และคะแนนสอบที่ได้มาน่าจะพัฒนาให้มีความเที่ยงสูงขึ้นได้อีกหากในการสอบมีจำนวนสถานการณ์มากขึ้น เนื่องด้วยรูปแบบข้อสอบอัตนัยประยุกต์ที่ใช้ในการสอบของคณะฯ ยังเน้นการสอบถามการจัดการปัญหาของผู้ป่วยตลอดตั้งแต่ต้นจนจบ (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 ปัญหาสำคัญประการหนึ่งที่สร้างความลำบากให้กับผู้สอบคือปริมาณข้อมูลที่น่าเสนอให้ผู้สอบอ่านในสถานการณ์ผู้ป่วยแต่ละรายนั้นมีมาก ทำให้ผู้สอบต้องใช้เวลาในการอ่านมากและเหลือเวลาสำหรับเขียนคำตอบน้อย ถึงแม้ว่าในการนำเสนอข้อมูลของข้อสอบอัตนัยประยุกต์จะได้มีการแยกข้อมูลเดิมที่เคยนำเสนอไปก่อนหน้านี้ ออกจากข้อมูลใหม่ที่เพิ่มเติมขึ้นมาในการนำเสนอข้อสอบแต่ละตอนแล้วก็ตาม ด้วย

รายละเอียดที่นำเสนอมีมาก ผู้สอบก็ยังคงมีความจำเป็นต้องประมวลผลข้อมูลปริมาณมากอยู่ดี จากการทบทวนเนื้อหาของข้อสอบอัตโนมัติประยุกต์ที่ได้จัดสอบไปหลายครั้งพบว่าข้อสอบหลายข้อใช้ข้อมูลเพียงส่วนน้อยของที่นำเสนอเท่านั้นก็สามารถนำไปสู่การแก้ปัญหาและการตัดสินใจเลือกการส่งตรวจหรือให้การรักษาผู้ป่วยได้อย่างถูกต้อง ดังนั้นแนวทางในการพัฒนาคุณภาพของข้อสอบอัตโนมัติประยุกต์อีกทางหนึ่งคือการที่อาจารย์ผู้ออกข้อสอบพึงตระหนักถึงข้อจำกัดเรื่องเวลาในการทำข้อสอบของนักศึกษาและเขียนสถานการณ์ผู้ป่วยให้มีความกระชับ นำเสนอเฉพาะข้อมูลที่มีความจำเป็นในการตัดสินใจให้การดูแลรักษาผู้ป่วยเท่านั้น ในการนำเสนอข้อมูลแต่ละตอนควรต้องทบทวนว่าข้อมูลเก่าที่เคยให้ในขั้นตอนก่อนหน้านั้นมีความจำเป็นต้องนำเสนอซ้ำทั้งหมดหรือไม่ หากทำได้ควรทำการสรุปข้อมูลให้ผู้เข้าสอบ และตัดทอนข้อมูลที่ไม่จำเป็นในการแก้ปัญหาขั้นตอนนั้น ๆ ออกไป ตัวอย่างเช่น ในข้อสอบตอนที่หนึ่งมีการนำเสนอประวัติผู้ป่วยสั้น ๆ แล้วมีโจทย์ถามถึงประวัติที่จะชักเพิ่มเติม และการตรวจร่างกายที่จะทำเพื่อนำไปสู่การวินิจฉัยโรค ในข้อสอบตอนที่สองอาจารย์นำเสนอประวัติและผลการตรวจร่างกายเพิ่มเติมให้ แล้วมีโจทย์ถามถึงการวินิจฉัยโรค และการส่งตรวจทางห้องปฏิบัติการที่เหมาะสม ในข้อสอบตอนที่สามอาจารย์นำเสนอข้อมูลการวินิจฉัยโรคของผู้ป่วยพร้อมผลการตรวจทางห้องปฏิบัติการ แล้วถามแนวทางการรักษา การนำเสนอข้อสอบในลักษณะนี้ในข้อสอบหลายข้อมีการนำเสนอข้อมูลของโจทย์ซ้ำเดิมและค่อย ๆ เพิ่มข้อมูลขึ้นในทุกขั้นตอน ในข้อสอบตอนที่สองก็นำเสนอข้อมูลที่เสนอในตอนหนึ่งกับสอง ในข้อสอบตอนที่สามก็นำเสนอข้อมูลที่เสนอในตอนหนึ่ง สอง และ สาม ซึ่งเมื่อผ่านการสอบไปหลายตอนจะมีข้อมูลสะสมจำนวนมากที่ผู้สอบต้องอ่าน การนำเสนอข้อสอบที่มีประสิทธิภาพมากกว่าควรมีการสรุปข้อมูลอย่างเหมาะสม ในข้อสอบตอนที่สาม หากได้ข้อสรุปการวินิจฉัยโรคแล้ว จะถามแนวทางการรักษาโรค อาจารย์ควรพิจารณาตัดข้อมูลประวัติและการตรวจร่างกายออก หากการสั่งการรักษาจำเป็นต้องทราบข้อมูลจากประวัติ หรือการตรวจร่างกายบางอย่าง เช่น น้ำหนักตัว หรือ โรคร่วมที่ส่งผลต่อการ

วางแผนการรักษา ก็ให้นำเสนอเฉพาะข้อมูลที่ส่งผลต่อการตัดสินใจในขั้นตอนนั้นเท่านั้น

การนำเสนอข้อสอบอัตโนมัติประยุกต์ด้วยระบบคอมพิวเตอร์ก็เป็นอีกแนวทางหนึ่งที่คณะแพทยศาสตร์ศิริราชพยาบาลเห็นความสำคัญ และได้ดำเนินการพัฒนาอย่างต่อเนื่อง คณะแพทยศาสตร์ศิริราชพยาบาลมีความพร้อมในการพัฒนาด้านนี้มากพอสมควร เนื่องด้วยมีห้องคอมพิวเตอร์ที่มีจำนวนคอมพิวเตอร์มากพอที่จะจัดให้ผู้เข้าสอบทุกคนมีจอคอมพิวเตอร์ส่วนตัว มีการวางระบบเครือข่ายให้มีการส่งผ่านข้อมูลระหว่างเครื่องคอมพิวเตอร์ได้ดี และมีความเสถียรของระบบพอสมควร มีการวางมาตรการรักษาความปลอดภัยของข้อมูลในระบบที่ดี สามารถควบคุมการเข้าออกของข้อมูลจากระบบเครือข่ายคอมพิวเตอร์ได้ จึงส่งผลให้คณะได้ปรับปรุงแบบการจัดสอบอัตโนมัติประยุกต์จากระบบสอบด้วยข้อสอบกระดาษมาเป็นการนำเสนอข้อสอบบนจอคอมพิวเตอร์ ตั้งแต่ปีการศึกษา 2552 ซึ่งจากการสำรวจความเห็นของนักศึกษาผู้เข้าสอบได้รับการตอบรับดีมาก นักศึกษาพึงพอใจกับการสอบในระบบนี้ในระดับมากถึงมากที่สุด อย่างไรก็ตามระบบการสอบนี้ยังมีโอกาสที่จะพัฒนาให้ดีขึ้นได้อีก ในระบบการจัดสอบปัจจุบันของคณะฯ ยังคงเป็นรูปแบบที่ไม่ได้ใช้คอมพิวเตอร์อย่างเต็มรูปแบบ ยังคงให้ผู้สอบเขียนคำตอบลงในกระดาษคำตอบและเก็บกระดาษในตอนท้ายของการสอบในแต่ละสถานการณ์ผู้ป่วย การใช้ประโยชน์ของคอมพิวเตอร์ในการสอบปัจจุบันเน้นไปในการนำเสนอข้อมูลที่ทำให้ผู้สอบสามารถเห็นภาพถ่ายรังสี ภาพการตรวจทางห้องปฏิบัติการ แผนภาพ ตาราง รวมถึงรูปของผู้ป่วยได้ โดยผู้สอบทุกคนเห็นภาพที่มีความละเอียดสูงเท่าเทียมกัน และทำให้การบริหารการสอบทำได้มีประสิทธิภาพมากขึ้น ตัดปัญหาผู้สอบลืกลองเปิดดูข้อสอบในตอนต่อไปล่วงหน้า หรือทำข้อสอบในบางตอนเกินเวลา การแสดงเวลาที่เหลือในการทำข้อสอบแต่ละตอนบนหน้าจอทำให้ผู้สอบบริหารเวลาในการทำข้อสอบได้ดีขึ้น

ระบบจัดสอบอัตโนมัติประยุกต์ด้วยคอมพิวเตอร์อย่างเต็มรูปแบบที่ไม่ต้องมีการเขียนตอบในกระดาษเลยนั้นมีการจัดทำในต่างประเทศ^{12,23} แต่ต้องยอมรับว่าการ

สร้างระบบการทดสอบอัตโนมัติด้วยคอมพิวเตอร์อย่างเต็มรูปแบบนั้นเป็นงานที่ซับซ้อนและมีความท้าทายหลายอย่าง ทั้งในด้านผู้จัดสอบ ระบบเครือข่ายคอมพิวเตอร์ และผู้เข้าสอบ ในอนาคตอันใกล้นี้ทางฝ่ายการศึกษาฯ ยังไม่มีแนวทางที่จะพัฒนาการสอบอัตโนมัติประยุกต์เป็นระบบคอมพิวเตอร์อย่างเต็มรูปแบบ ด้วยข้อจำกัดสำคัญสามประการคือ ความพร้อมของผู้เข้าสอบ ความพร้อมของผู้ตรวจข้อสอบ และความพร้อมของระบบการสื่อสารระหว่างผู้ใช้กับคอมพิวเตอร์ กล่าวคือ ผู้เข้าสอบจำนวนไม่น้อยยังไม่คุ้นเคยกับการพิมพ์คำตอบที่มีทั้งภาษาไทยและภาษาอังกฤษผสมกันภายในเวลาที่จำกัด อาจารย์ผู้ตรวจข้อสอบจำนวนไม่น้อยยังไม่สะดวกที่จะทำการตรวจข้อสอบและกรอกคะแนนบนหน้าจอคอมพิวเตอร์ในสถานที่และเวลาที่กำหนด และการสร้างระบบการสื่อสารระหว่างคอมพิวเตอร์กับผู้ใช้ให้ทั้งนำเสนอข้อมูลผู้ป่วยที่มีรายละเอียดมาก พร้อมกับตอบรับคำตอบที่มีทั้งอักษร ตัวเลข และสัญลักษณ์พิเศษ ที่ผู้เข้าสอบจะพิมพ์เข้าเครื่องพร้อม ๆ กันหลายร้อยคนโดยมีการควบคุมเวลาอย่างรัดกุมด้วย ยังเป็นสิ่งที่ทำได้ยากในระบบเครือข่ายคอมพิวเตอร์ในปัจจุบัน ดังนั้นในอนาคตอันใกล้นี้ทิศทางการพัฒนาระบบการทดสอบข้อสอบอัตโนมัติคงยังมุ่งเน้นไปในรูปแบบการนำเสนอข้อสอบผ่านจอภาพคอมพิวเตอร์ ร่วมกับการเขียนตอบในกระดาษคำตอบอยู่

แต่ถึงแม้ว่าจะคงการทดสอบอัตโนมัติในรูปแบบผสมผสานเช่นนี้ ผู้นิพนธ์ก็ยังเห็นว่าสิ่งที่จะระบบการนำเสนอข้อมูลผ่านจอคอมพิวเตอร์สามารถทำให้ดีขึ้นได้ เช่นการทำให้ภาพมีรายละเอียดสูงขึ้น การเปิดโอกาสให้ผู้เข้าสอบสามารถขยายภาพเพื่อดูรายละเอียดในบางส่วน การปรับรูปแบบการนำเสนออักษร และพื้นหลังของจอภาพให้ผู้เข้าสอบอ่านข้อมูลได้ง่ายขึ้น เป็นต้น ซึ่งสิ่งเหล่านี้จะได้มีการศึกษาหาแนวทางในการพัฒนาในการสอบอัตโนมัติประยุกต์ครั้งต่อไป แต่อย่างไรก็ตามด้วยศักยภาพของระบบการทดสอบในปัจจุบัน ผู้นิพนธ์ยังมีความเห็นว่าอาจารย์ผู้ออกข้อสอบก็ยังไม่ได้ใช้ศักยภาพของระบบอย่างเต็มที่ ยังมีข้อสอบหลายข้อที่ใช้การบรรยายสิ่งตรวจพบที่สามารถมองเห็นเป็นภาพได้

แต่นำมาเขียนเป็นอักษรบรรยายสิ่งตรวจพบดังกล่าวซึ่งทำให้ผู้เข้าสอบไม่ได้คิดวิเคราะห์และแปลผลการตรวจด้วยตนเอง แนวทางการพัฒนาข้อสอบอัตโนมัติประยุกต์ที่สมควรได้รับการส่งเสริมในระบบการทดสอบปัจจุบันคือการใช้สื่อที่เป็นรูปภาพในข้อสอบให้มากขึ้น ไม่ว่าจะเป็นการตรวจร่างกายจากการดู การดูภาพรังสี การดูคลื่นไฟฟ้าหัวใจ การดูสิ่งส่งตรวจด้วยกล้องจุลทรรศน์ ล้วนแล้วแต่ควรนำเสนอเป็นรูปภาพทั้งสิ้น

บทสรุป

ในบทความนี้ผู้นิพนธ์ได้กล่าวถึงความรู้พื้นฐานในการสร้างข้อสอบอัตโนมัติประยุกต์ โดยได้สรุปลักษณะพื้นฐานของข้อสอบอัตโนมัติประยุกต์ พัฒนาการของข้อสอบประเภทนี้จากรูปแบบการจัดการปัญหาผู้ป่วยเป็นการแก้ปัญหาสำคัญ มีการสรุปขั้นตอนสำคัญในการสร้างข้อสอบอัตโนมัติประยุกต์ห้าขั้นตอน ได้แก่ (1) ตั้งกลุ่มพัฒนาข้อสอบ, (2) เลือกปัญหาทางคลินิก, (3) กำหนดปัญหาสำคัญ, (4) เขียนโจทย์คำถาม, และ (5) กำหนดเกณฑ์การให้คะแนน และในตอนท้ายได้มีการนำหลักการพัฒนาข้อสอบต่าง ๆ ที่กล่าวมาแล้วมาวิเคราะห์สถานการณ์การทดสอบอัตโนมัติประยุกต์สำหรับนักศึกษาแพทย์คณะแพทยศาสตร์ศิริราชพยาบาลและเสนอแนะแนวทางในการพัฒนาคุณภาพการสอบอัตโนมัติประยุกต์สี่แนวทาง ได้แก่ (1) เนื้อหาข้อสอบ, (2) รูปแบบคำถาม, (3) จำนวนสถานการณ์ผู้ป่วย, และ (4) การนำเสนอข้อสอบ ผู้นิพนธ์เชื่อมั่นว่าหากการทดสอบอัตโนมัติประยุกต์ได้รับการพัฒนาอย่างเหมาะสมจะนำไปสู่การประเมินความรู้ 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 medical education. Dordrecht: Kluwer Academic Publishers, 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 questions: 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 Professor of Physic, Cambridge University School of clinical Medicine, 1985:8-27.
14. Page G, Bordage G. The Medical Council of Canada's key features project: A more valid written examination of clinical decision-making skills. *Acad Med* 1995;70:104-10.
15. Neufeld VR, Norman GR, Barrows HS, Feightner JW. Clinical problem solving by medical students: A longitudinal and cross-sectional analysis. *Med Educ* 1981;15:315-22.
16. Perkins DN, Salomon G. Are cognitive skills context-bound? *Educ Researcher* 1989;18:16-25.
17. van der Vleuten CPM, Swanson DB. Assessing clinical skills with standardized patients: The state of the art. *Teach Learn Med* 1990;2 (58-76).
18. Eva KW. On the generality of specificity. *Med Educ* 2003;37(7): 587-88.
19. Bordage G, Page G. An alternate approach to PMPs, the key feature concept. In: Hart I, Harden R, editors. *Further developments in assessing clinical competence*. Montreal: Can-Heal Publications, 1987:57-75.
20. Page G, Bordage G, Allen T. Developing key features problems and examinations to assess clinical decision-making skills. *Acad Med* 1995;70:194-201.
21. Norman G, Bordage G, Page G, Keane D. How specific is case specificity? *Med Educ* 2006;40:618-23.
22. Hambleton RK, Pitoniak MJ. Setting performance standards. In: Brennan RL, editor. *Educational measurement, 4th ed*. Westport, CT: Praeger publishers, 2006:433-70.
23. Federation of State Medical Boards of the United States, National Board of Medical Examiners. USMLE Step 3: Content description and general information, Available from http://www.usmle.org/pdfs/step-3/2014content_Step3.pdf. June 2014.

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

ขอแก้ไขเป็น

เวชบันทึกศิริราช

ปีที่ 7 ฉบับที่ 2 กรกฎาคม-ธันวาคม 2557 หน้า 74-83 เรื่อง

“หน้ากากครอบกล่องเสียง Laryngeal Mask Airway (LMA)” โดย อังศุมาศ หวังดี

และได้ทำการแก้ไข pdf เรียบร้อยแล้ว

รศ. ดร.นพ.เชิดศักดิ์ ไอรณนรัตน์

หัวข้อ : Summary

Summary

นพ. เชิดศักดิ์ ไอรณนรัตน์
ภาควิชาศัลยศาสตร์ คณะแพทยศาสตร์ศิริราชพยาบาล
มหาวิทยาลัยมหิดล

Experiential Learning Theory

Experimentation (Apply) → Experience → Reflection → Conceptualization → Experimentation

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

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

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

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

Atkins and Murphy (1995)

Five Levels of Reflection

Reconstructing
Reasoning
Relating
Responding
Reporting

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

Summary of the Workshop

- Morning
 - MCQ item development
 - MCQ item analysis
- Afternoon
 - CR item development

5

Shee.si.mahidol.ac.th

เอกสารประกอบการอบรม



16 March 2018

Part 3 : Practical examination

รศ. ดร.นพ.เชิดศักดิ์ ไอรณิรัตน์

หัวข้อ : OSCE item development

OSCE Item
Development

เชิดศักดิ์ ไอรณิรัตน์

ภาควิชาศัลยศาสตร์ คณะแพทยศาสตร์ศิริราชพยาบาล

มหาวิทยาลัย มหิดล

OSCE

- Objective
- Structured
- Clinical
- Examination
- มีวัตถุประสงค์ที่ชัดเจน
- มีการจัดโครงสร้างเป็นสถานีย่อย
- ประเมินทักษะทางคลินิก
- การสอบ

History

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

OSCE

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

Components of an OSCE item

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

Scenario

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

Scenario 1

หัวข้อ : การตรวจร่างกายผู้ป่วยที่มีอาการปวดท้อง

Objective : นักศึกษาแพทย์สามารถแสดงวิธีการตรวจร่างกายผู้ป่วยที่มีอาการปวดท้องเฉียบพลัน และให้การวินิจฉัยที่ถูกต้องได้

ผู้สอบ: นักศึกษาแพทย์ชั้นปีที่ 6

สถานการณ์: สมบูรณ์ อายุ 35 ปี มีอาการปวดท้องใต้ชายโครงด้านซ้าย 6 ชั่วโมง ปวดตื้อๆตลอดเวลา

คำสั่ง : จงแสดงวิธีการตรวจหน้าท้องผู้ป่วย บรรยายสิ่งที่ตรวจพบและให้การวินิจฉัยโรคที่คิดถึงมากที่สุด 1 โรค

เวลา : 5 นาที (ตรวจร่างกาย 4 นาทีครึ่ง บอกสิ่งที่พบและวินิจฉัยครึ่งนาที)

Scenario 1 (cont.)

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

Instruction for Examinees

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

Standardized Patient (SP)

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

History

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

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

Instruction for SPs

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

Instruction for SPs

- โจทย์:** นักศึกษาจะทำการซักประวัติท่านเพื่อให้การวินิจฉัยโรคให้ท่านให้ข้อมูลต่อไปนี้
- ข้อมูลจากโจทย์:** ท่านเป็นผู้ป่วยชายไทย อายุ 40 ปี มีอาการปวดขาหนีบข้างขวา 1 วัน
- การตรวจ:** ตรวจร่างกายเบื้องต้น เป็นเสีย กางเกงที่สามารถเปิดหน้าท้องได้สะดวก
- การตรวจ:** ตรวจร่างกายเบื้องต้น ไม่มี
- ข้อมูลที่นักศึกษาจะซักถามจากท่าน**
- ตำแหน่งที่ปวดท้อง: ปวดบริเวณขาหนีบด้านขวา
 - ลักษณะของอาการปวด: ช่วงแรกปวดหน่วงๆ ตลอดเวลา
 - มีอาการปวดร้าวไปที่อื่นหรือไม่: ไม่มี
 - ลักษณะของอาการปวดตอนเริ่มแรก เป็นอย่างไร เป็นทันทีทันใดหรือค่อยๆปวดเพิ่มขึ้นบ้าง เป็นที่ตำแหน่งเดียวกันหรือมีการย้ายที่ปวด: ค่อยๆปวดเพิ่มขึ้นบ้าง ไม่มีการย้ายที่ปวด
 - มีปัจจัยใดที่ทำให้ปวดเปลี่ยนแปลงหรือไม่: ปวดเพิ่มมากขึ้นในขณะขึ้นหรือออ

Instruction for SPs (cont.)

- อาการร่วมอื่นๆ
 - ทั่วไป: มีไข้ต่ำๆ
 - ระบบทางเดินอาหาร: มีอาการปวดท้องบีบๆเป็นพักๆ คลื่นไส้และอาเจียน
- ประวัติอดีต
 - ประวัติการมีก้อนที่ขาหนีบ
สังเกตมีก้อนที่ขาหนีบข้างขวา มา 2 ปี
 - ประวัติการเปลี่ยนแปลงของก้อนที่ขาหนีบ
ขนาดก้อนเท่าๆเดิม จะโตมากเวลาขึ้นหรือเบ่ง เวลานอนแล้ว ก้อนจะยุบได้เอง
- ...
- ประวัติส่วนตัว: อาชีพ การสูบบุหรี่ การดื่มสุรา
ทำงานเป็นเสมียน สูบบุหรี่วันละ 2 ซองมา 10 ปี ไม่ดื่มสุรา

Components of an OSCE item

- Scenario (ภาพรวมสถานการณ์)
- Instruction for examinees (คำแนะนำผู้เข้าสอบ)
- Instruction for SPs (คำแนะนำผู้ป่วยมาตรฐาน)
- Scoring rubric (ใบให้คะแนน +/- คำแนะนำอาจารย์)

Scoring Rubric General Format

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

Scoring Rubric

- กระชับ ได้ใจความ สื่อความหมายตรงกัน
- กำหนดประเด็นที่สำคัญ หรือเป็นจุดที่มักทำผิดพลาด
- บรรยายพฤติกรรมที่ผู้ประเมินสังเกตได้
- กำหนดน้ำหนักคะแนนตามความสำคัญ

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

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

เชิดศักดิ์ ไอรมนรัตน์

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

1. สิ่งแรกที่ต้องคำนึงถึงคือวัตถุประสงค์ของการสอบ เนื่องจาก OSCE เป็นการสอบที่ต้องใช้ทรัพยากรมาก ควรตั้งวัตถุประสงค์การสอบเพื่อประเมินความรู้ความสามารถที่ไม่สามารถประเมินได้ด้วยวิธีอื่น เช่น ทักษะในการสื่อสารกับผู้ป่วย ทักษะการให้คำแนะนำแก่ผู้ป่วย ทักษะการทำหัตถการ เป็นต้น ไม่ควรใช้ OSCE เพื่อวัดความรู้ผิวเผินที่สามารถวัดได้ด้วยข้อสอบ MCQ
2. วางแบบแปลนของเนื้อหาข้อสอบ (test blueprint) ที่ครอบคลุมเนื้อหาวิชาในทุกด้าน และทุกทักษะที่ต้องการประเมินอย่างเท่าเทียมกัน มีการระบุว่าในการสอบ OSCE นี้ทดสอบความรู้เรื่องใดบ้าง (โรคปอด โรคหัวใจ โรคไต ฯลฯ) และใช้ทักษะใดบ้าง (การซักประวัติ การตรวจร่างกาย การให้คำแนะนำ ฯลฯ) อย่างละเอียด ระวังอย่าให้เนื้อหาข้อสอบมีน้ำหนักในเรื่องใดเรื่องหนึ่งมากกว่าเรื่องอื่น
3. ในการเขียนโจทย์ OSCE แต่ละข้อ ต้องเขียนให้ครอบคลุมรายละเอียดทุกด้านของการสอบ ได้แก่ คำชี้แจงสำหรับนักเรียน สำหรับผู้ป่วยมาตรฐาน และสำหรับอาจารย์ผู้คุมสอบ สถานการณ์ผู้ป่วยจำลอง ประวัติและผลการตรวจร่างกายที่ผู้ป่วยมาตรฐานต้องแสดงออก อุปกรณ์ประกอบที่ต้องใช้ ระยะเวลาที่ต้องใช้ แบบฟอร์มให้คะแนน และเกณฑ์การให้คะแนน
4. การเขียนโจทย์ผู้ป่วยควรนำข้อมูลมาจากผู้ป่วยจริง ซึ่งจะทำให้โจทย์มีความเหมือนจริง ไม่ขาดรายละเอียดในเนื้อหาของโจทย์ และประหยัดเวลาในการแต่งโจทย์ นอกจากนี้ยังทำให้มีแฟ้มประวัติและผลการตรวจเพิ่มเติมรวมทั้งฟิล์มที่สามารถนำมาใช้เสริมโจทย์ได้ง่าย
5. โจทย์สำหรับแต่ละสถานี่ควรมีความยาวเหมาะสม โจทย์ที่ใช้เวลานานสามารถให้ข้อมูลเกี่ยวกับความสามารถของนักเรียนในเรื่องนั้นๆ ได้ละเอียด แต่ก็ทำให้มีโอกาสดัดความสามารถของนักเรียนได้น้อยเรื่อง เนื่องจากทักษะทางการแพทย์หลายด้านมีความเจาะจงต่อภาวะโรค (นักเรียนที่ซักประวัติโรคเลือดได้ดีอาจซักประวัติผู้ป่วยโรคซึมเศร้าไม่คล่องได้) โดยทั่วไปแนะนำให้จัดเวลาที่ใช้สอบในแต่ละสถานี่ ให้นักเรียนได้มีโอกาสสอบในอย่างน้อย 8 – 10 สถานี่ (ยิ่งมีสถานี่สอบมาก 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 ไชยมณีรัตน์

ในบทความนี้จะขอเสนอเกร็ดความรู้เกี่ยวกับการใช้ผู้ป่วยมาตรฐาน (Standardized patients) ใน OSCE ก่อนอื่นผมขอกล่าวถึงนิยามของศัพท์ที่สำคัญในการใช้ผู้ป่วยในการสอบก่อน เราเรียกคนปกติที่ไม่มีภาวะเจ็บป่วย แต่แสดงบทบาทเป็นผู้ป่วยว่า ผู้ป่วยสมมติ (simulated patient) ซึ่งผู้ป่วยสมมติเหล่านี้อาจแสดงออกไม่สม่าเสมอ เมื่อได้พบกับนักเรียนแต่ละคน หากเราทำการฝึกให้ผู้ป่วยสมมติ (หรือ ผู้ป่วยจริง) แสดงออกซึ่งอาการและอาการแสดงอย่างสม่าเสมอ เป็นมาตรฐานเดียวกันไม่ว่าจะได้พบกับนักเรียนคนใด เราจะได้ ผู้ป่วยมาตรฐาน (standardized patient) การสอบ OSCE ให้ได้ผลการประเมินที่แม่นยำนั้นต้องใช้ผู้ป่วยมาตรฐาน (standardized patient, SP)

1. ผู้ป่วยมาตรฐานต้องได้รับการฝึกฝนอย่างจริงจังมั่นใจว่าการแสดงออกซึ่งอาการและอาการแสดงได้มาตรฐานในทุกครั้งที่แสดงบทบาท การฝึกฝนนี้ต้องเริ่มต้นจากการมีบท (script) ที่ดี มีความละเอียดครอบคลุมข้อมูลทุกด้านที่เกี่ยวข้องกับภาวะโรคที่สนใจ และมีการฝึกซ้อมและตรวจแก้ไขโดยอาจารย์ผู้แต่งโจทย์เพื่อให้มั่นใจว่าความเข้าใจบทบาทของผู้ป่วยมาตรฐานถูกต้องตามความตั้งใจของผู้แต่งโจทย์ โดยทั่วไปเมื่อได้รับการฝึกฝนแล้วผู้ป่วยมาตรฐานสามารถแสดงออกซึ่งอาการและอาการแสดงได้อย่างถูกต้องมากกว่า 90%
2. ในการสอบใหญ่บางครั้งมีความจำเป็นต้องใช้ผู้ป่วยมาตรฐานหลายคนเพื่อแสดงบทบาทเดียวกัน มีหลายการศึกษาแสดงว่าการใช้ผู้ป่วยมาตรฐานหลายคนในลักษณะนี้ไม่ลดความแม่นยำของผลสอบ ตรงเท่ากับที่เราใช้สถานีสอบ OSCE มากเพียงพอ และผู้ป่วยมาตรฐานได้ถูกสุ่มกระจายตัวอยู่ตามสถานีสอบอย่างไม่ลำเอียง (randomly distributed)
3. หลายการศึกษาที่วิเคราะห์การสอบที่มีความจำเป็นต้องใช้ผู้ป่วยมาตรฐานชุดเดิมสอบนักเรียนหลายชุดต่อเนื่องกัน พบว่านักเรียนที่สอบรอบหลังไม่ได้ทำคะแนนได้ดีกว่านักเรียนที่สอบรอบแรก แสดงว่านักเรียนที่สอบก่อนไม่ให้ข้อมูลเกี่ยวกับการสอบที่เป็นประโยชน์แก่นักเรียนที่สอบรอบหลัง หรือหากนักเรียนให้ข้อมูลแก่กัน ข้อมูลเพียงที่ได้รับเกี่ยวกับคำชี้แจงโจทย์โดยไม่มีข้อมูลรายละเอียดของเกณฑ์การให้คะแนนนั้นไม่ได้ก่อให้เกิดความได้เปรียบในการสอบแก่นักเรียนรอบหลัง
4. นอกจากจะใช้ผู้ป่วยมาตรฐานเพื่อวัดทักษะของนักเรียนที่เกี่ยวข้องกับผู้ป่วยโดยตรง (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 ไชรมณีรัตน์

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 “นักเรียนอธิบายเหตุการณ์ที่จะทำได้ชัดเจน” : 1 ไม่เห็นด้วยอย่างยิ่ง, 2 ไม่เห็นด้วย, 3 เห็นด้วย, 4 เห็นด้วยอย่างยิ่ง) การให้คะแนนด้วย checklist จะได้ผลการประเมินที่ผู้ให้คะแนน (rater) มีความเห็นพ้องกัน (inter-rater agreement) มากกว่า แต่สามารถแยกแยะความแตกต่างระหว่างนักเรียนที่มีความสามารถต่างกันได้ดีไม่เท่ากับการให้คะแนนด้วย rating scale ควรใช้ checklist สำหรับให้คะแนนโจทย์ที่ประเมินความครบถ้วนของเนื้อหาหรือขั้นตอน (เช่น ชักประวัติ ตรวจร่างกาย) แต่ควรใช้ rating scale สำหรับให้คะแนนโจทย์ที่ประเมินคุณภาพของทักษะหรือกระบวนการปฏิบัติ (เช่น ทักษะการสื่อสาร ทักษะการทำหัตถการ)
2. ไม่มีความจำเป็นต้องใช้ผู้ให้คะแนน (rater) มากกว่า 1 คน ต่อ 1 สถานี หากมีทรัพยากรบุคคลมากพอ เราควรที่จะเพิ่มจำนวนสถานีสอบ มากกว่า เพิ่มจำนวนผู้ให้คะแนนต่อสถานี การเพิ่มจำนวนสถานีสอบ ส่งผลให้คะแนนสอบ OSCE มีความแม่นยำเพิ่มขึ้นมากกว่า การเพิ่มจำนวนผู้ให้คะแนนต่อสถานี
3. นอกจากเราจะให้อาจารย์แพทย์เป็นผู้ให้คะแนนแล้ว เรายังสามารถฝึกให้ผู้ป่วยมาตรฐาน (standardized patient) ทำการให้คะแนนได้ด้วย พบว่าเมื่อได้รับการอธิบายเกณฑ์การให้คะแนนและฝึกปฏิบัติแล้ว ผู้ป่วยมาตรฐาน สามารถให้คะแนนที่มีความแม่นยำสูงไม่แพ้อาจารย์แพทย์ ข้อดีของการให้ผู้ป่วยมาตรฐานเป็นผู้ให้คะแนนคือสะดวก และประหยัด ในทางกลับกันการให้อาจารย์แพทย์เป็นผู้ให้คะแนนมีข้อได้เปรียบคืออาจารย์สามารถชี้แนะข้อบกพร่อง และแนะนำแนวทางการปรับปรุงแก้ไขทักษะและวิธีคิดของนักเรียนได้ทันที
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 ดังนั้นเราไม่ควรแจ้งรายละเอียด ข้อถูก ข้อผิด ของนักเรียนแต่ละคนในทุกสถานี แต่แจ้งเพียงผลสอบว่าผ่านหรือไม่ผ่าน

หัวข้อ : Long case examination

Long-case Examination

PORNPAN KOOMANACHAI, MD
FACULTY OF MEDICINE SIRIRAJ HOSPITAL

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

Advantages and Disadvantages

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

- The candidate
 - 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 candidate

Assessment Objectives

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

- Psychomotor skills

- Attitudes

Long-case Examination

- อาจารย์เคยมีประสบการณ์คุมสอบรายยาวหรือการจัดสอบรายยาวหรือไม่
- อาจารย์ประสบปัญหาใดในการคุมสอบหรือจัดสอบบ้าง
- อาจารย์มีแนวทางแก้ไขปัญหอย่างไร

Long-case Examination

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

“Luck of the draw; different examiners examine different candidates on different patients”

Stokes, 1974

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

*Noricine, 2002
Int J Health Sci (Qassim). 2008; 2(2):3-7*

Long-case Examination

- ให้อาจารย์แต่ละท่านเขียนลักษณะของทักษะและคะแนนที่ต้องการประเมินผู้เรียนจากการสอบรายยาว โดยให้คะแนนเต็มของการสอบเป็น 100 คะแนน (เวลา 5 นาที)

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

OBJECTIVE STRUCTURED LONG EXAMINATION RECORD (OSLER)		DATE: _____
CANDIDATE'S Name: _____	EXAMINATION NO. _____	
<small>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 the co-examiner at follow.</small>		
<small>GRADE</small> P+ = Very good/excellent P = Pass/Borderline pass P- = Below pass	<small>MARKS</small> (50-80) = see cover page (30-50) = for specific (15-45) = mark details	<small>CD-EXAMINER:</small> _____
PRESENTATION OF HISTORY	GRADE	AGREED GRADE
REGULARITY	_____	_____
COMMUNICATION PROCESS	_____	_____
SYSTEMATIC PRESENTATION	_____	_____
CORRECT FACTS ESTABLISHED	_____	_____
PHYSICAL EXAMINATION		
SYSTEMIC	_____	_____
TECHNIQUE (including attitude to patient)	_____	_____
CORRECT FINDINGS ESTABLISHED	_____	_____
APPROPRIATE INVESTIGATIONS IN A LOGICAL SEQUENCE (communication process option)		
CLINICAL ACUMEN (problem identification/problem solving ability)	_____	_____
ADDITIONAL COMMENTS:		
Please Tick (V) for CASE DIFFICULTY Standard _____ Individual examiner _____ Difficult _____ Very difficult _____		
GRADE Individual examiner OVERALL GRADE _____ MARK _____		Standard case: 1 problem Difficult: up to 3 problems Very difficult: > 3 problems
OVERALL GRADE _____ MARK _____		PRICES OF EXAMINERS AGREED GRADE _____ AGREED MARKS _____

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

Long-case Examination

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

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. MEO
 3. Long case exam

Long Case Examination

- ข้อกำหนดของ ศร. ในการสอบ long case ข้อกำหนดของ ศร. 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)

Long-case Examination

Questions & Comments

ผศ. นพ.ตรีภพ เลิศบรรณพงษ์

หัวข้อ : Portfolio

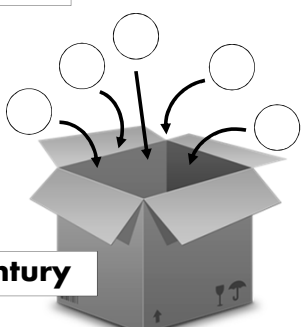
Portfolio
The 21st century assessment tool

LERTBUNNAPHONG T, M.D.

The story of Portfolio

เก่ง และ ดี
 รู้ได้อย่างไร

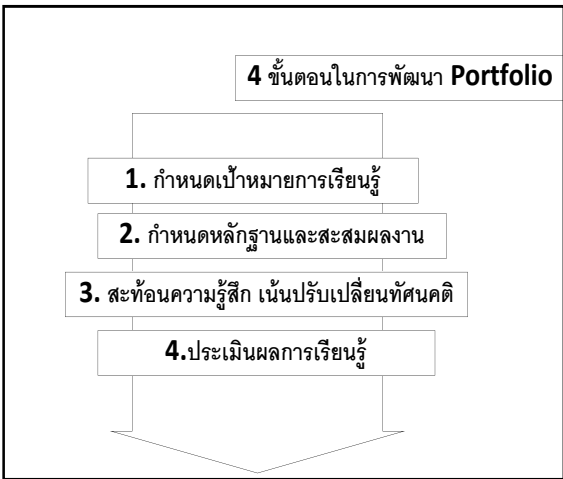
The story of Portfolio



21st century

What is PORTFOLIO?





1. กำหนดเป้าหมาย

VISION

MISSION

POLICY

DESIRED DOCTOR

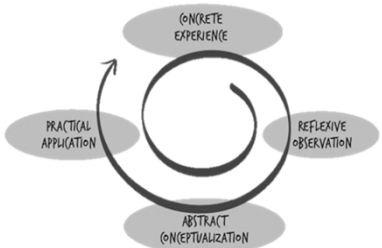
2. กำหนดหลักฐาน

ต้องสะท้อนเป้าหมาย

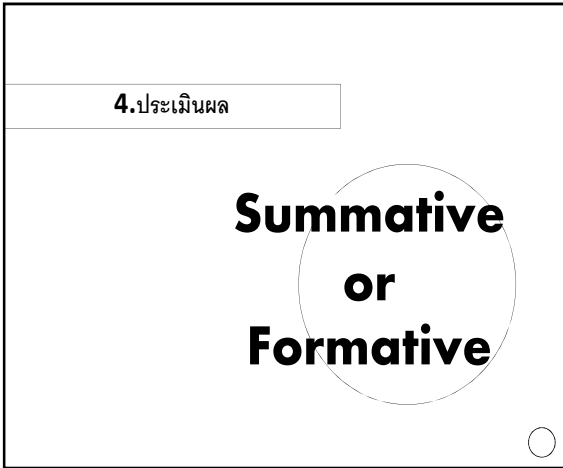


3. สะท้อนความรู้สึก ปรับเปลี่ยนทัศนคติ

CRITICAL REFLECTION



Learning without reflection is waste
Confucius







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 และพฤติกรรมการทำงาน ในรูปแบบ Portfolio ดังผลสรุปในเอกสารฉบับนี้

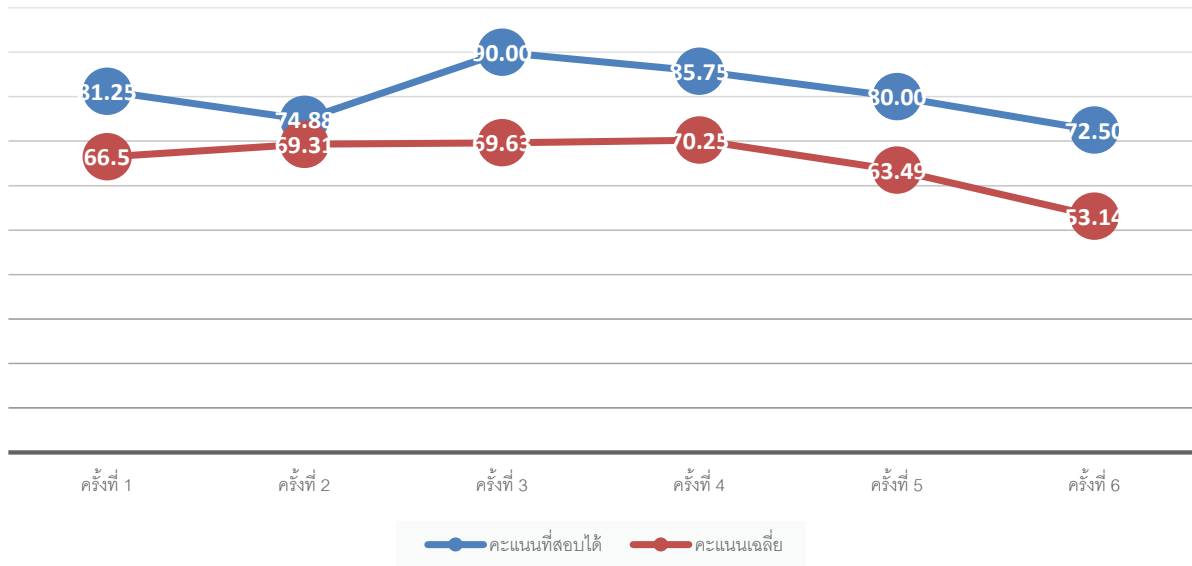
ภาควิชาฯ ขออำนวยการให้ท่านประสบความสำเร็จในการดำเนินชีวิตครอบครัว และหน้าที่การงาน ตลอดไป

ศาสตราจารย์คลินิก นายแพทย์ชาญชัย วันทนาศิริ
หัวหน้าภาควิชาสูติศาสตร์-นรีเวชวิทยา
คณะแพทยศาสตร์ศิริราชพยาบาล มหาวิทยาลัยมหิดล

การประเมินความรู้ทางสูติศาสตร์-นรีเวชวิทยา

(Knowledge assessment)

ผลการสอบวัดระดับความรู้ทางวิชาการ



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

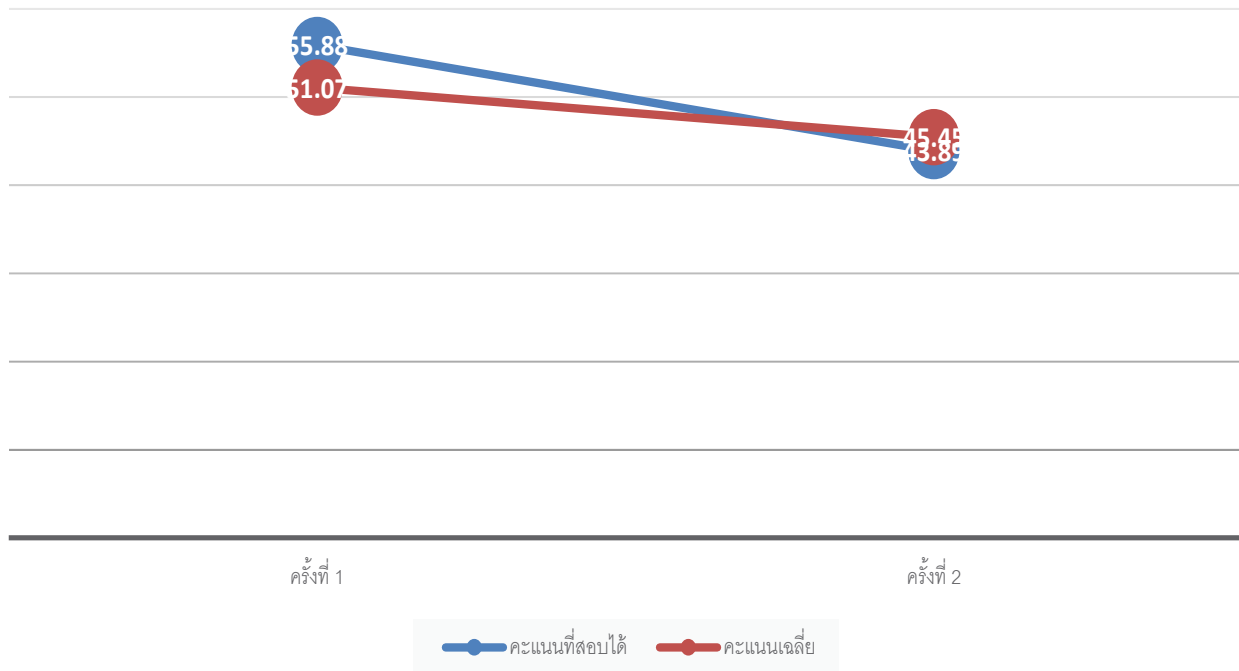
การสอบวิชาภาษาอังกฤษและกฎหมายทางการแพทย์ครั้งที่ 1

ผ่าน ไม่ผ่าน

การสอบวิชาภาษาอังกฤษและกฎหมายทางการแพทย์ครั้งที่ 2 (กรณีสอบไม่ผ่านครั้งแรก)

ผ่าน ไม่ผ่าน

ผลการสอบวัดระดับความรู้ทางวิชาการ



ชั้นปีที่ 2

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

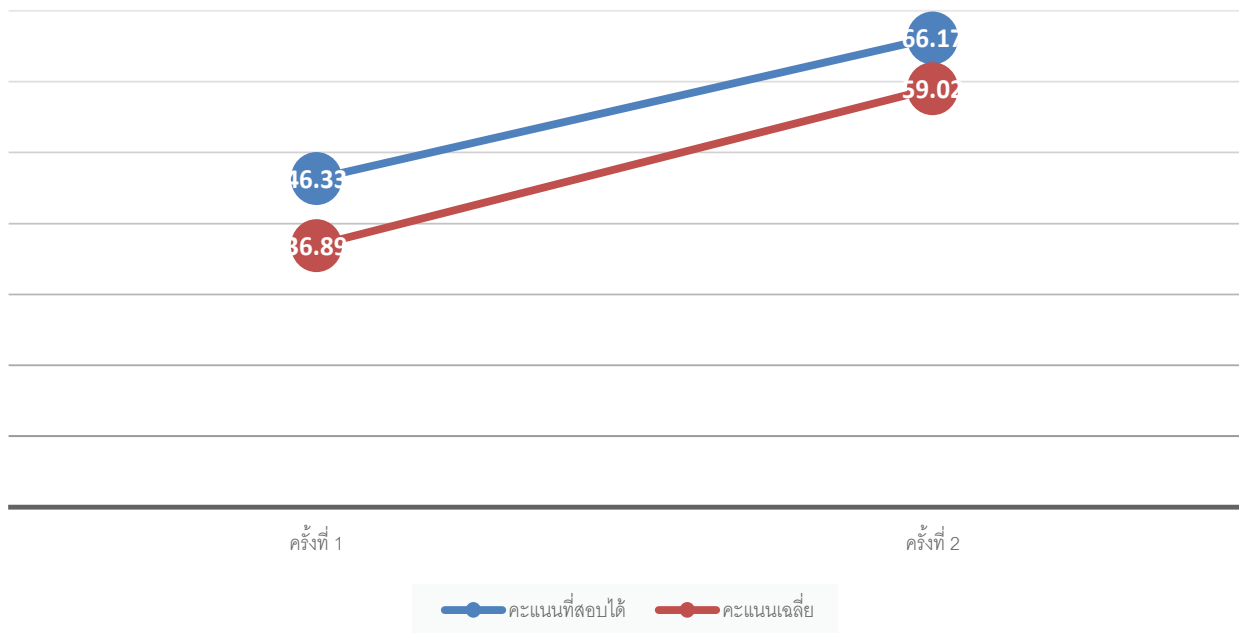
การสอบ OSLER ในสถาบัน ครั้งที่ 1

ผ่าน ไม่ผ่าน

การสอบ Basic science ราชวิทยาลัยสูตินรีแพทย์แห่งประเทศไทย

ผ่าน ไม่ผ่าน

ผลการสอบวัดระดับความรู้ทางวิชาการ



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 ราชวิทยาลัยสูตินรีแพทย์แห่งประเทศไทย ครั้งที่ 1

ผ่าน ไม่ผ่าน

การสอบ OSLER ราชวิทยาลัยสูตินรีแพทย์แห่งประเทศไทย ครั้งที่ 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 salpingoophorectomy	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 เนื่องจากอยู่ระหว่างกระบวนการพัฒนาและปรับปรุงระบบเก็บข้อมูลหัตถการ
แพทย์ประจำบ้าน ภาควิชาสูติศาสตร์-นรีเวชวิทยา

การทำงานวิจัยระดับแพทย์ประจำบ้าน
(Research competency)

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

อาจารย์ผู้ควบคุมผู้ช่วยศาสตราจารย์นายแพทย์ธันยรัตน์ วงศ์วานานุรักษ์

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>

ผลการประเมินเจตคติและพฤติกรรมการทำงานของแพทย์ประจำบ้าน (Multisources feedback)

แพทย์ประจำบ้านจะได้รับการประเมินในประเด็นต่อไปนี้

1. ความรู้ความสามารถด้านวิชาการ

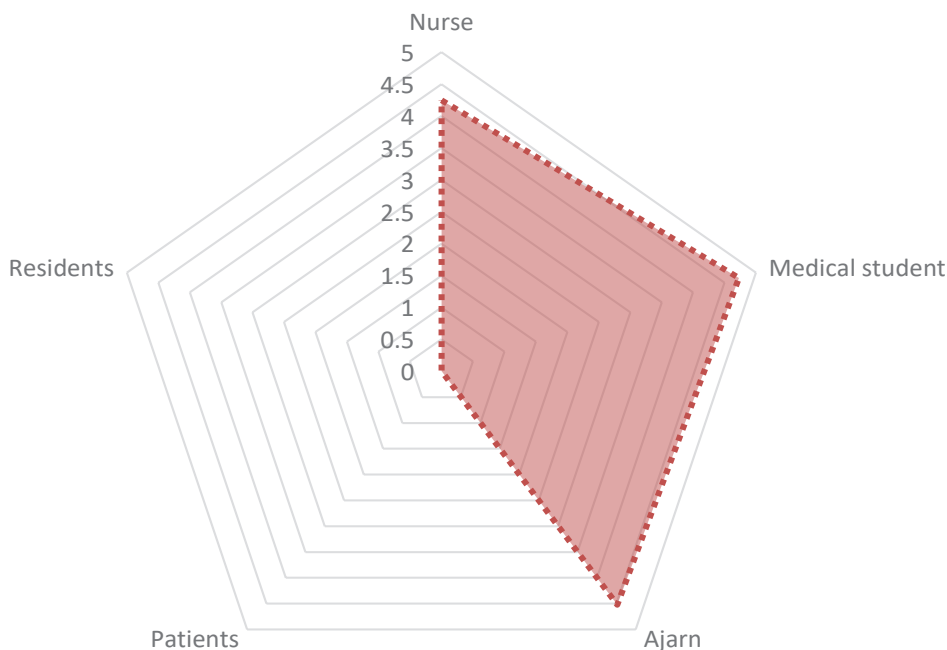
2. ทักษะพื้นฐานในการปฏิบัติงาน

ได้แก่ ทักษะการสื่อสารกับเพื่อนร่วมงานและผู้ป่วย/ญาติ การบันทึกรายงานผู้ป่วย การทำงานร่วมกับผู้อื่น และบุคลิกภาพขณะปฏิบัติงาน

3. คุณธรรมและจริยธรรม

ได้แก่ ความรับผิดชอบ ความเสียสละ ความตรงต่อเวลา ความซื่อสัตย์ การปฏิบัติตามระเบียบข้อบังคับ และอภัย/น้ำใจ/ความเอื้อเฟื้อต่อผู้อื่น

MULTISOURCES FEEDBACK 2011

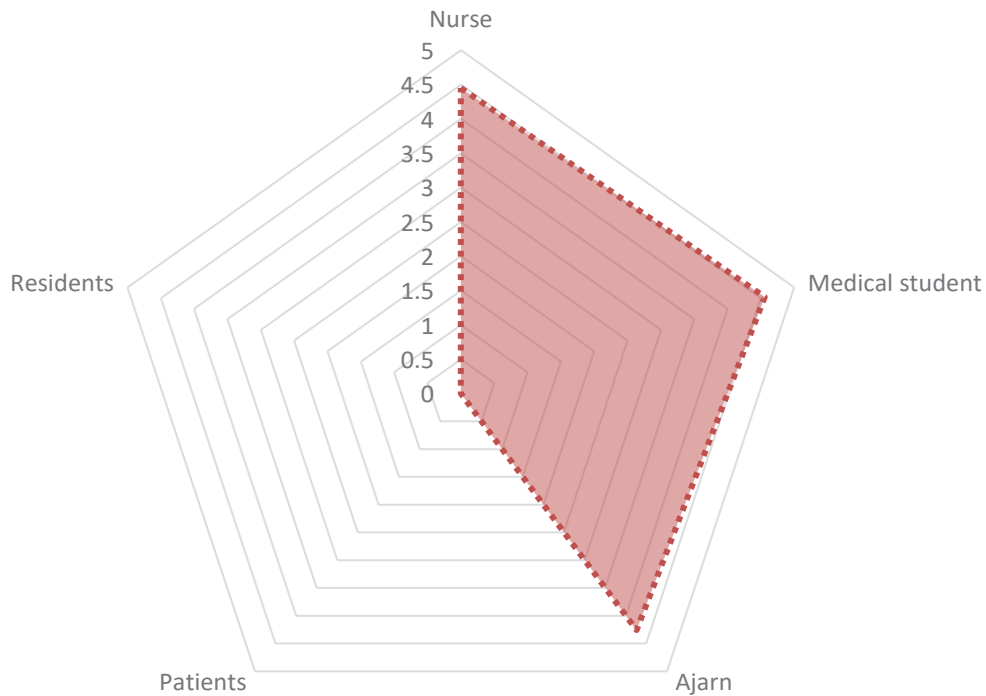


ชั้นปีที่ 1 ปีการศึกษา 2554

Rotation	อาจารย์ (5 คะแนน)	นักศึกษาแพทย์ (5 คะแนน)	พยาบาล (5 คะแนน)	แพทย์ ประจำบ้าน	ผู้รับบริการ
พระศรีฯ 9/2			4.61		
LR เข้า		5.00	3.76		
LR พิเศษเข้า			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		

*ยังไม่มีผลการประเมินจากแพทย์ประจำบ้านและผู้รับบริการในปีการศึกษา 2554

MULTISOURCES FEEDBACK 2012

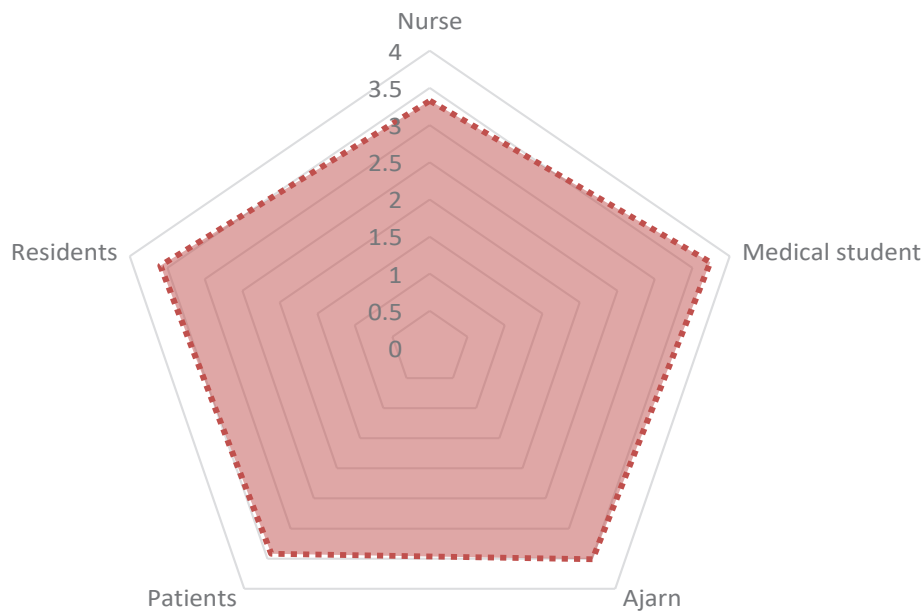


ชั้นปีที่ 2 ปีการศึกษา 2555

Rotation	อาจารย์ (5 คะแนน)	นักศึกษาแพทย์ (5 คะแนน)	พยาบาล (5 คะแนน)	แพทย์ ประจำบ้าน	ผู้รับบริการ
นรีเวช 1	3.93	4.00	4.53		
เลือดสิน	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		

*ยังไม่มีผลการประเมินจากแพทย์ประจำบ้านและผู้รับบริการในปีการศึกษา 2555

MULTISOURCES FEEDBACK 2013



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 เข้า		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 และพฤติกรรมการทำงาน ในรูปแบบ Portfolio ดังผลสรุปในเอกสารฉบับนี้

ภาควิชาฯ ขออัญวยพรให้ท่านประสบความสำเร็จในการดำเนินชีวิตครอบครัว และหน้าที่การงาน ตลอดไป

ศาสตราจารย์คลินิก นายแพทย์ชาญชัย วันทนาศิริ

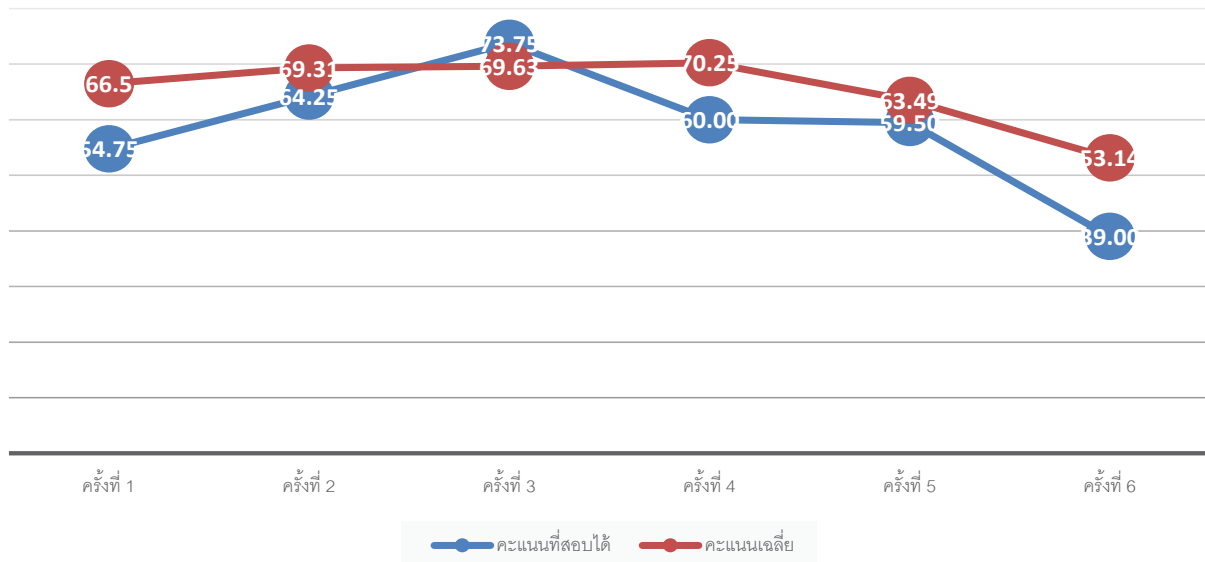
หัวหน้าภาควิชาสูติศาสตร์-นรีเวชวิทยา

คณะแพทยศาสตร์ศิริราชพยาบาล มหาวิทยาลัยมหิดล

การประเมินความรู้ทางสูติศาสตร์-นรีเวชวิทยา

(Knowledge assessment)

ผลการสอบวัดระดับความรู้ทางวิชาการ



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

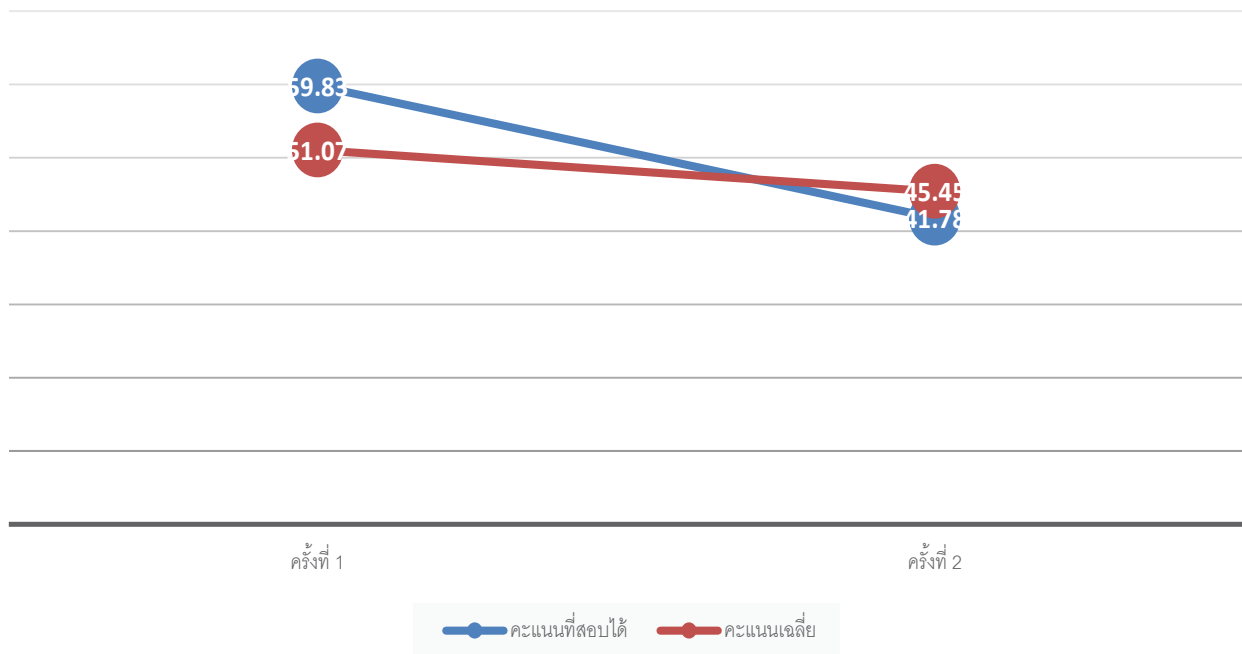
การสอบวิชาภาษาอังกฤษและกฎหมายทางการแพทย์ครั้งที่ 1

ผ่าน ไม่ผ่าน

การสอบวิชาภาษาอังกฤษและกฎหมายทางการแพทย์ครั้งที่ 2 (กรณีการสอบครั้งที่ 1 ไม่ผ่าน)

ผ่าน ไม่ผ่าน

ผลการสอบวัดระดับความรู้ทางวิชาการ



ชั้นปีที่ 2

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

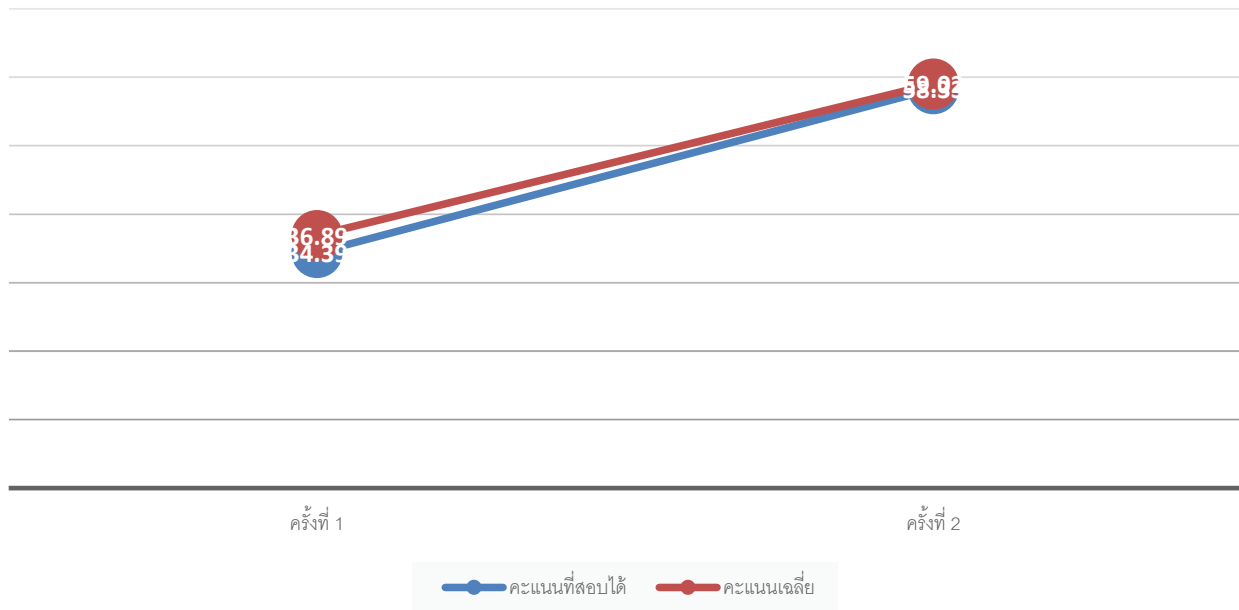
การสอบ OSLER ในสถาบัน ครั้งที่ 1

ผ่าน ไม่ผ่าน

การสอบ Basic science ราชวิทยาลัยสูตินรีแพทย์แห่งประเทศไทย

ผ่าน ไม่ผ่าน

ผลการสอบวัดระดับความรู้ทางวิชาการ



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 ราชวิทยาลัยสูตินรีแพทย์แห่งประเทศไทย ครั้งที่ 1

ผ่าน ไม่ผ่าน

การสอบ OSLER ราชวิทยาลัยสูตินรีแพทย์แห่งประเทศไทย ครั้งที่ 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 salpingoophorectomy	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 เนื่องจากอยู่ระหว่างกระบวนการพัฒนาและปรับปรุงระบบเก็บข้อมูลหัตถการ
 แพทย์ประจำบ้าน 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

- ผ่าน SIRB เมื่อ 28 สิงหาคม 2555
เลขที่ 415/2555(EC3)
- ประกวดการนำเสนองานวิจัยในการประชุมราชวิทยาลัยสูตินรีแพทย์แห่งประเทศไทย
วันที่ 26 พฤศจิกายน 2556
 เข้าร่วมนำเสนอ ไม่ได้รับรางวัล
 เข้าร่วมนำเสนอ ได้รับรางวัล
- การตีพิมพ์ในวารสารวิชาการ
 ไม่ได้ตีพิมพ์
 ได้รับการตีพิมพ์ (ระบุรายละเอียดวารสาร)

ผลการประเมินเจตคติและพฤติกรรมการทำงานของแพทย์ประจำบ้าน (Multisources feedback)

แพทย์ประจำบ้านจะได้รับการประเมินในประเด็นต่อไปนี้

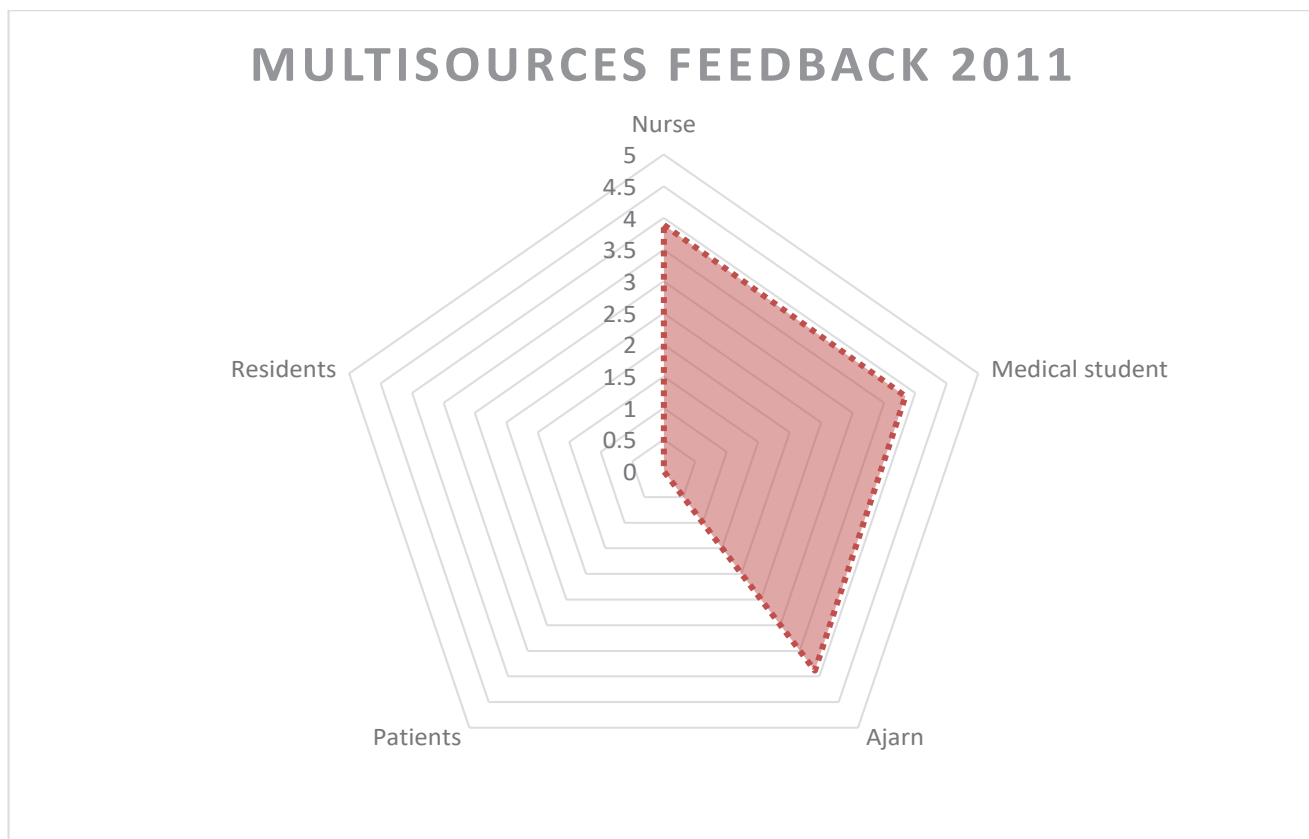
1. ความรู้ความสามารถด้านวิชาการ

2. ทักษะพื้นฐานในการปฏิบัติงาน

ได้แก่ ทักษะการสื่อสารกับเพื่อนร่วมงานและผู้ป่วย/ญาติ การบันทึกรายงานผู้ป่วย การทำงานร่วมกับผู้อื่น และบุคลิกภาพขณะปฏิบัติงาน

3. คุณธรรมและจริยธรรม

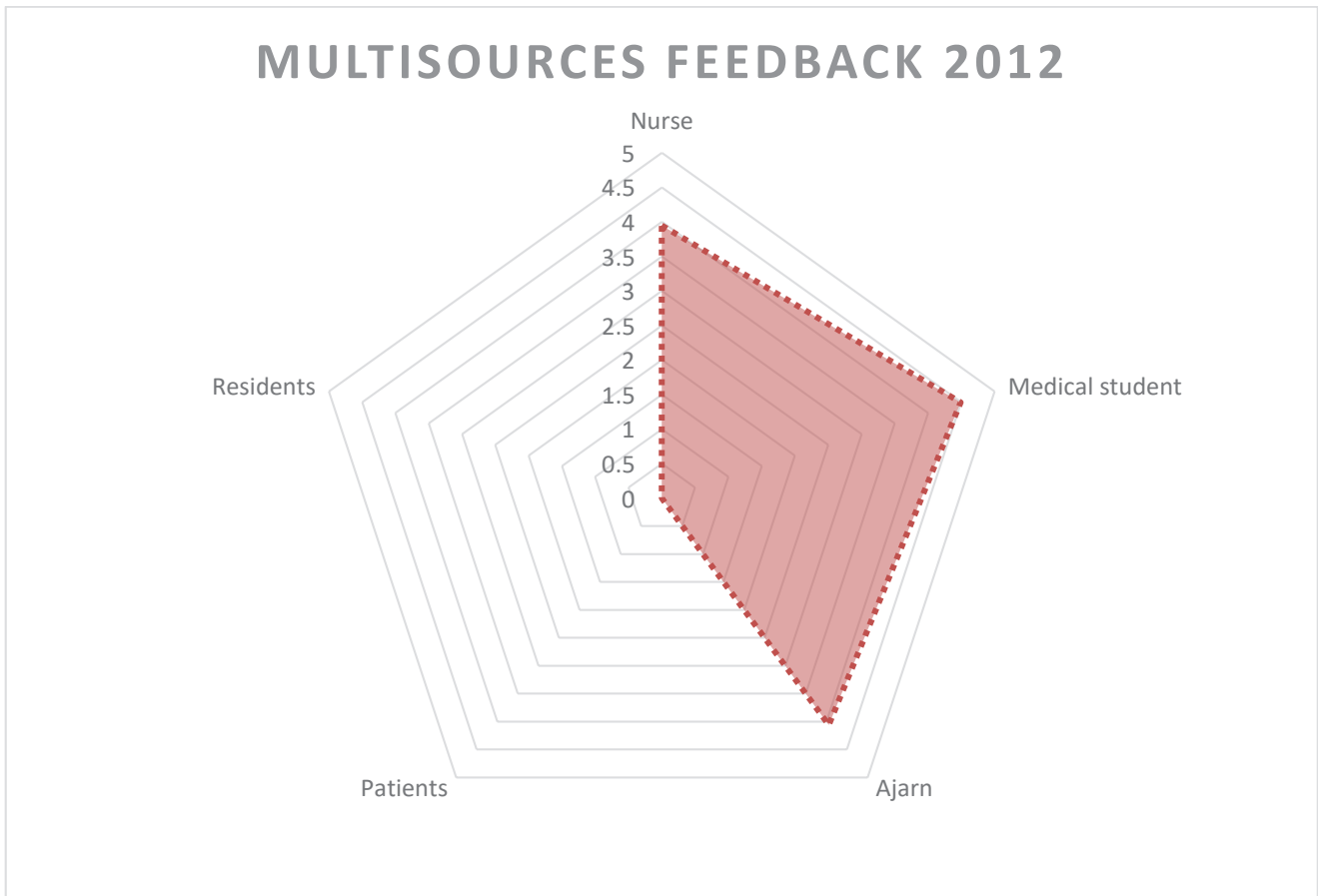
ได้แก่ ความรับผิดชอบ ความเสียสละ ความตรงต่อเวลา ความซื่อสัตย์ การปฏิบัติตามระเบียบข้อบังคับ และอัธยาศัย/น้ำใจ/ความเอื้อเฟื้อต่อผู้อื่น



ชั้นปีที่ 1 ปีการศึกษา 2554

Rotation	อาจารย์ (5 คะแนน)	นักศึกษาแพทย์ (5 คะแนน)	พยาบาล (5 คะแนน)	แพทย์ ประจำบ้าน	ผู้รับบริการ
พระศรีฯ 9/2			4.00		
LR เข้า		3.58	4.00		
LR พิเศษเข้า			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		

*ยังไม่มีผลการประเมินจากแพทย์ประจำบ้านและผู้รับบริการในปีการศึกษา 2554

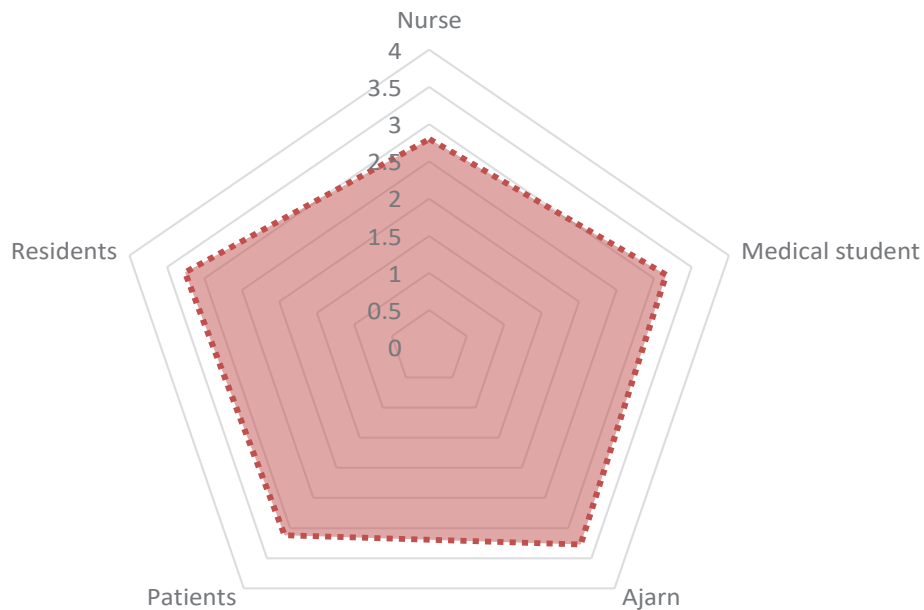


ชั้นปีที่ 2 ปีการศึกษา 2555

Rotation	อาจารย์ (5 คะแนน)	นักศึกษาแพทย์ (5 คะแนน)	พยาบาล (5 คะแนน)	แพทย์ ประจำบ้าน	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		

*ยังไม่มีผลการประเมินจากแพทย์ประจำบ้านและผู้รับบริการในปีการศึกษา 2555

MULTISOURCES FEEDBACK 2013



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 เข้า		3.10	2.95	2.63	
Surgery	3.62				
คะแนนเฉลี่ย	3.27	3.26	2.80	3.16	3.12

*เริ่มการประเมินจากนักศึกษาแพทย์และผู้รับบริการ ในปีการศึกษา 2556



Portfolios for Assessment and Learning

Jan van Tartwijk
Erik W Driessen

AMEE GUIDE
Assessment

45






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

© AMEE 2010

ISBN: 978-1-903934-57-9

Contents

Abstract	1
Introduction	2
Portfolio goals, content, and organization	4
Portfolios as a multipurpose instrument	4
Electronic portfolios	7
Portfolios and learning from experience	9
Theoretical background	9
Reflection and professional development	10
Using portfolios as tools for assessment	14
Factors influencing the success of the introduction of a portfolio	21
People	21
Academic leadership	23
Infrastructure	23
Concluding remarks	24
References	25

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.

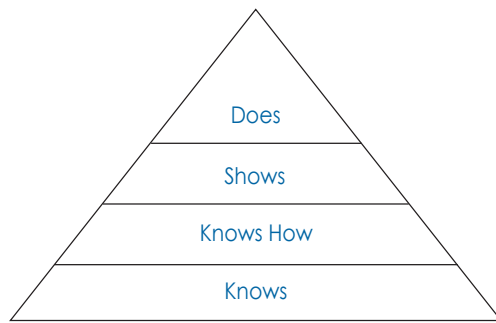
However, stimulating the development of competencies (Box 1) and the assessment of its result is complicated. Already in 1990, Miller described the challenges involved in assessing clinical competence. He presented a framework for clinical assessment, shaped like a pyramid (Figure 1), whose layers from 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.

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

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.

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

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

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)

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

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, 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 instrument¹**: 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.

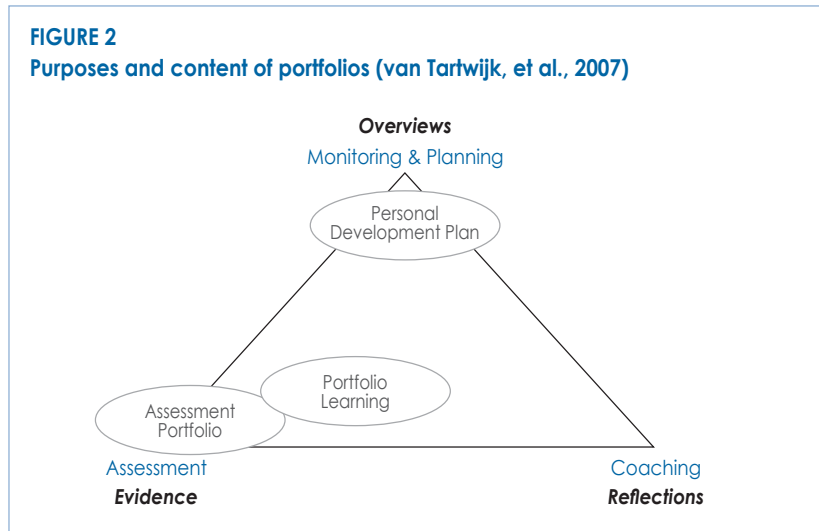
Obviously, a portfolio can be used to achieve more than one goal. When a portfolio is to serve a combination of goals, its position in the triangle will shift towards the centre because its strengths have to be distributed more evenly over evidence, overviews and reflections. In practice, the majority 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

A portfolio can also stimulate reflection...

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 (Doran 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.

In a study comparing web-based and paper-based portfolios (Driessen et al., 2007a), not only did learners add more personal touches to content and form and invest 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

...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.

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.

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 reflection 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

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

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

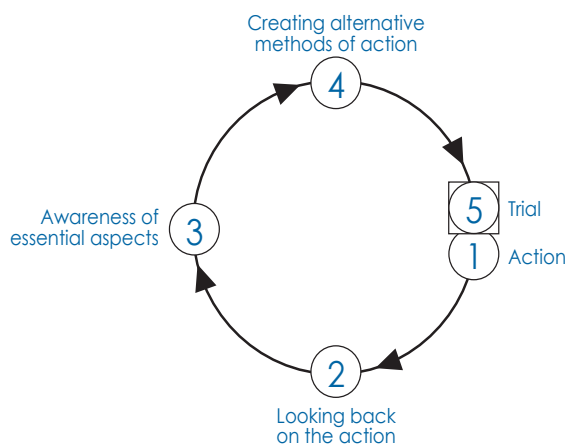
not be viewed as hierarchical. Different (educational) contexts and situations may lend themselves more to one kind of reflection than to another.

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 (**A**ction, **L**ooking back, **A**wareness, **C**reating alternative methods, **T**rial) (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).

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.

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

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.

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.

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

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).

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).

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

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.

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).

The evidence held by a portfolio is often not standardised and its meaning often depends on the context from which it originates.

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	<p>Slow in taking a history and performing a physical examination. Considers irrelevant aspects.</p> <p>Slow in making a diagnosis. Misses important conclusions.</p> <p>Frequently unable to formulate management plan and needs considerable guidance.</p>	<p>Adequate speed in taking a history and performing a physical examination. Relevant aspects are considered.</p> <p>Adequate speed in making a diagnosis. Diagnosis contains important conclusions.</p> <p>Formulates an adequate management plan for simple clinical presentations.</p> <p>Needs some guidance.</p> <p>Achieves these goals in the second half of the internship.</p>	<p>Conducts an adequate and efficient history and physical examination.</p> <p>Arrives at an accurate diagnosis within adequate time.</p> <p>Formulates an adequate management plan for simple clinical presentations.</p> <p>Needs little guidance.</p> <p>Has achieved these goals at the start of the internship.</p>
Professionalism (for instance as judged by 360 degree feedback)	<p>Does not keep commitments.</p> <p>Occasionally fails to ask for supervision when this is necessary. Reacts defensively to feedback.</p> <p>Is unable to cope with stress</p> <p>Does not pay attention to his/her personal appearance.</p> <p>Frequently shows awkward behaviour or behaves disrespectfully.</p>	<p>Keeps commitments.</p> <p>Asks for supervision when this is necessary.</p> <p>Needs help in reflecting and considering alternatives and responds adequately to feedback.</p> <p>Occasionally needs help in coping with stress.</p> <p>Appropriate personal appearance; behaves respectfully.</p>	<p>Keeps commitments.</p> <p>Asks for supervision when this is necessary.</p> <p>Is able to reflect critically; responds adequately to feedback and is prepared to acknowledge errors.</p> <p>Is able to cope with stress adequately.</p> <p>Looks well cared for and behaves respectfully.</p>
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.	<p>Incomplete, limited or one-sided description of strengths and weaknesses in performance (e.g. only strengths or only weaknesses, limited to one competency).</p> <p>No explanations only lists of facts or situations.</p> <p>No learning goals, learning goals do not match the analysis or are not specific.</p>	<p>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)</p> <p>Some of the learning goals are not specified.</p>	<p>Above expectation (authentic, recognizable, and well explained). A good analysis of strengths and weaknesses. Also internal attributions and references to evidence in the portfolio.</p> <p>Logical, detailed (based on the analysis) and attainable learning goals.</p>

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 & Lincoln's (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).

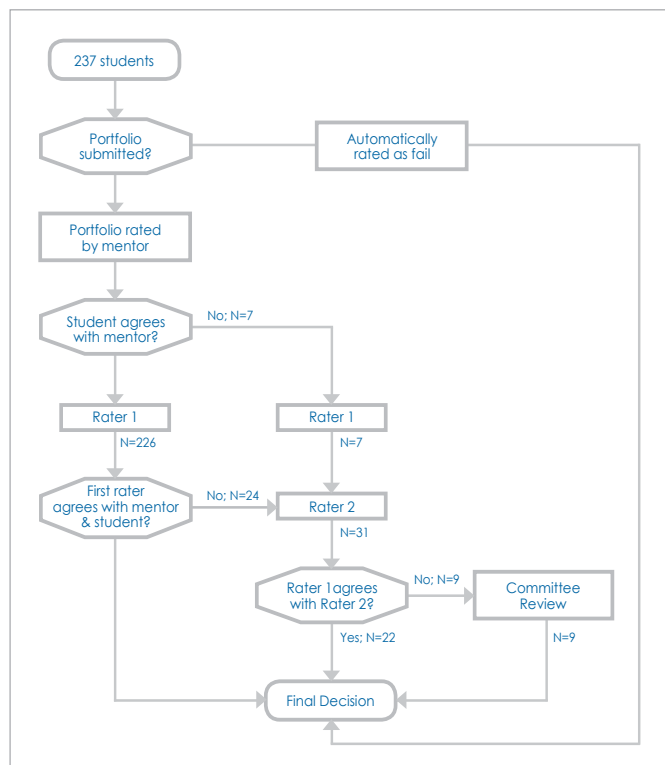
The tension between assessment for support and assessment for high stakes decision making will never disappear.

BOX 10
A procedure for portfolio assessment (Driessen et al., 2005a)

The student 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.

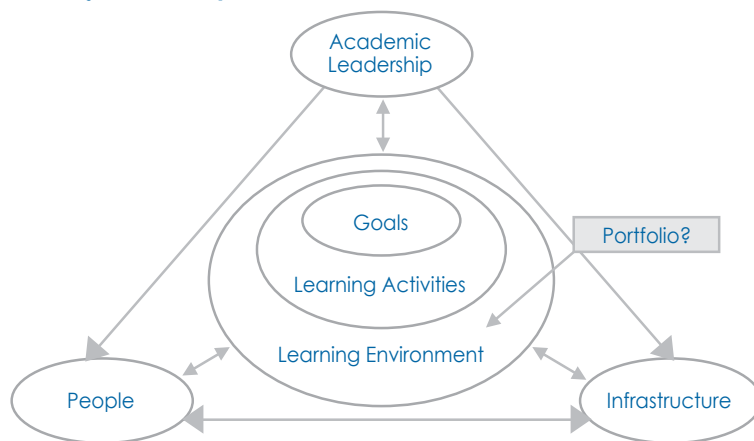
Factors influencing the success of the introduction of a portfolio²

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).

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

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.

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 is 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.

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".

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.

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.

"..... introducing portfolios is just like buying shoes: the best choice depends on purpose and comfort comes with wearing".

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, HAMMERNESSE 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, TABATABA D, GOLD SL & WINER LR (2006). Learning while evaluating: the use of an electronic evaluation portfolio in a geriatric medicine clerkship. *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 reflection 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).
- HAMMERNES 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 health 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, TUIHOF 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 Champion-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**
Moirá 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

Scottish Charity No. SC 031618

รศ. ดร.นพ.เชิดศักดิ์ ไอรมนิรัตน์

หัวข้อ : Clinical performance ratings

Clinical Performance Ratings

เชิดศักดิ์ ไอรมนิรัตน์
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

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

Outline

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

3

4

Clinical Performance Ratings

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

Clinical Performance Ratings

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

Rater Errors

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

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

7

Rater Errors

Leniency/Severity

- difference in the levels of severity between raters

Rater inconsistency

- instability of the level of severity within each rater

Halo

- rater's tendency to let the rating of one trait influence his/her ratings on other traits

Restriction of range

- clustering of ratings around a particular point on the rating scale

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

8

Improving Raters

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

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

Assessment

**แบบประเมินการปฏิบัติงานของนักศึกษาแพทย์ปี 6
คณะแพทยศาสตร์ศิริราชพยาบาล**

น.ศ. พ.
ฝึกปฏิบัติงานที่
หอผู้ป่วย

รหัส
ภาควิชา/แผนก
ช่วงเวลาปฏิบัติงาน

ถึง

หัวข้อการประเมิน	%	ดีเยี่ยม (10)	ดี (8-9)	ผ่าน (6-7)	ไม่ผ่าน (<6)	หมายเหตุ
1. ความรู้		มีความรู้พื้นฐานที่สอดคล้องอย่างดี และสามารถนำมาประยุกต์ใช้ในการดูแลผู้ป่วยเป็นอย่างดี	มีความรู้พื้นฐานที่สอดคล้องอย่างดี แต่ยังไม่สามารถนำมาประยุกต์ใช้ในการดูแลผู้ป่วยได้ดีนัก	มีความรู้พื้นฐานที่สอดคล้องแต่ไม่สามารถนำมาประยุกต์ใช้ในการดูแลผู้ป่วย	ขาดความรู้พื้นฐานที่สำคัญ	
2. ทักษะ						
2.1 การแก้ปัญหาทางคลินิก		รวบรวมข้อมูลปัญหาได้สมบูรณ์ คิดวิเคราะห์แก้ปัญหาได้ด้วยตนเอง	รวบรวมข้อมูลปัญหาได้สมบูรณ์ คิดวิเคราะห์แก้ปัญหาได้ด้วยตนเอง	รวบรวมข้อมูลปัญหาได้สมบูรณ์ แต่คิดวิเคราะห์	การรวบรวมข้อมูลปัญหาและการคิดวิเคราะห์บกพร่อง	
2.2 ความสามารถในการดูแลผู้ป่วยและการตัดสินใจ		เลือกการสืบค้นและการรักษาได้ถูกต้อง สามารถบอกเหตุผล และคำอธิบายถึงปัจจัยรอบด้าน	เลือกการสืบค้นและการรักษาได้ถูกต้อง สามารถบอกเหตุผล แต่ยังขาดการคำนึงถึงปัจจัยรอบด้าน	เลือกการสืบค้นและการรักษาได้ถูกต้อง แต่ไม่สามารถบอกเหตุผลได้ชัดเจน	ไม่สามารถเลือกการสืบค้นและการรักษาได้อย่างถูกต้อง	
2.3 การบันทึกเวชระเบียน		มีข้อมูลสำคัญครบถ้วน เป็นระเบียบ อ่านง่าย ลงรายละเอียดและรหัส	มีข้อมูลสำคัญครบถ้วน แต่ไม่เป็นระเบียบ อ่านยาก หรือ ไม่ลงรายละเอียดชื่อ/รหัส	ขาดข้อมูลสำคัญบางอย่าง เช่น ประวัติสังคม ประวัติการผ่าตัด	ขาดข้อมูลที่สำคัญหลายอย่าง ไม่เขียน progress note	
2.4 การทำหัตถการ		ทำหัตถการที่สำคัญได้อย่างมีประสิทธิภาพ และปลอดภัย มีความชำนาญในการใช้เครื่องมือ และติดตามดูแลผู้ป่วยหลังทำหัตถการอย่างเหมาะสม	สามารถทำหัตถการที่สำคัญได้ แต่ความปลอดภัยไม่ดีพอ ต้องมีความช่วยเหลือในบางขั้นตอน มีการติดตามดูแลผู้ป่วยหลังทำหัตถการอย่างเหมาะสม	สามารถทำหัตถการที่สำคัญได้ แต่ต้องได้รับความช่วยเหลือจากผู้ช่วยหรือติดตามดูแลผู้ป่วยหลังผ่าตัด	ไม่สามารถทำหัตถการที่สำคัญได้ แม้จะได้รับการชี้แนะแล้ว ไม่รู้ขั้นตอนการทำหัตถการ และ/หรือขาดทักษะพื้นฐานในการทำหัตถการ	
2.5 ทักษะการนำเสนอ		เป็นขั้นตอนดีมาก เข้าใจง่าย	เป็นขั้นตอน พังเข้าใจ โดยอาจต้องถามเพิ่มเติมเล็กน้อย	ไม่เป็นที่พอใจในเนื้อหาบางส่วน	นำเสนอมาก นึกเรียนไม่มีความเข้าใจในเนื้อหาบางส่วน หรือใช้ภาษาไม่เหมาะสม หรือสร้างความสับสนให้แก่ผู้ป่วยและญาติ	
2.6 การสื่อสารกับผู้ป่วย/ญาติ		ดีมาก ผู้ป่วยและญาติพึงพอใจมาก	ดี ผู้ป่วยและญาติเข้าใจโรคที่เป็น	ผู้ป่วยและญาติบางคนไม่เข้าใจโรค		
3. ความเป็นวิชาชีพแพทย์						
3.1 ความสามารถในการเรียนรู้ด้วยตนเอง		แสดงถึงความใฝ่รู้ ค้นคว้าเพิ่มเติมได้ด้วยความตั้งใจ	แสดงถึงความใฝ่รู้ ค้นคว้าเพิ่มเติมได้โดยต้องขอร้องหรือชี้แนะ	ต้องการคำแนะนำและวิธีการจะค้นคว้าเพิ่มเติม	ขาดความใฝ่รู้ แม้จะได้รับคำแนะนำและชี้แนะ	
3.2 การวางตัวที่เหมาะสม		ตรงต่อเวลา บุคลิกภาพ ความประพฤติ การแต่งกายเหมาะสม เป็นส่วนใหญ	ตรงต่อเวลา บุคลิกภาพ ความประพฤติ การแต่งกายเหมาะสม เป็นส่วนใหญ่	ไม่ตรงต่อเวลา บุคลิกภาพ การแต่งกายเหมาะสมเป็นส่วนใหญ่	มีพฤติกรรมที่ไม่เหมาะสม และไม่ปรับปรุงหลังจากได้รับคำติชม	
3.3 ความรับผิดชอบ		รับผิดชอบมาก หรือ ได้รับคำชมในการดูแลผู้ป่วยและครอบครัว	รับผิดชอบดีในการดูแลผู้ป่วยและการอยู่เวร	ไม่มีข้อร้องเรียนเรื่องความรับผิดชอบในการดูแลผู้ป่วยและการอยู่เวร	ไม่รับผิดชอบ หรือมีข้อร้องเรียนในการดูแลผู้ป่วยและการอยู่เวร	
3.4 เจตคติและจริยธรรม		ดูแลผู้ป่วยทั้งร่างกายและจิตใจ อย่างดี เคารพสิทธิของผู้ป่วย	ดูแลผู้ป่วยทั้งร่างกายและจิตใจ เคารพสิทธิของผู้ป่วย	การดูแลผู้ป่วยขาดมิติด้านจิตใจ แต่ยังคงเคารพสิทธิของผู้ป่วย	การดูแลผู้ป่วยขาดมิติด้านจิตใจ และไม่เคารพสิทธิของผู้ป่วย	
3.5 มนุษยสัมพันธ์กับผู้ร่วมงาน		มีมนุษยสัมพันธ์ดีมาก การทำงานเป็นทีมดีมาก	มีมนุษยสัมพันธ์ดี ทำงานร่วมกับผู้อื่นได้	ขาดมนุษยสัมพันธ์ หรือมีปัญหาในการทำงานร่วมกับผู้อื่น	มนุษยสัมพันธ์ไม่ดี และไม่สามารถทำงานร่วมกับผู้อื่นได้	
เวลาปฏิบัติงาน		ครบ	ป่วย.....วัน	ลา.....วัน	ขาด.....วัน	
ความคิดเห็นเพิ่มเติม						
หมายเหตุ กรุณาใส่คะแนนในช่องสี่เหลี่ยมทแยงมุมของตาราง (ไม่มีจุดทศนิยม) , NA = ไม่สามารถประเมินได้						

ผู้ประเมิน
วันที่

ผศ. พญ.กชณา รักขมนัน

หัวข้อ : Workplace-based assessment

CLINICAL TEACHING MADE EASY

Workplace-based assessment

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

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

What is workplace-based assessment?

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

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

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

'Knows' forms the base of Miller's pyramid, the entry point in the development of expertise. This tier is best assessed using simple knowledge tests such as multiple choice questions. The next tier up 'knows

how' seeks to measure understanding or application of knowledge and is assessed using instruments such as unfolding patient management problems, extended matching or short essay questions. Higher up, objective structured clinical examinations assess at the 'shows how' level where students are required to demonstrate not only knowledge and understanding, but that they can bring together and manipulate relevant knowledge, skills and attitudes in a controlled situation.

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

Is it useful?

The utility, or usefulness, of an assessment has been defined as a product of its reliability, validity, cost-effectiveness, acceptability and educational impact (van der Vleuten, 1996). Utility can be applied to an entire assessment system or to an individual assessment method or component of the system. The concept is important in that no single element should be regarded

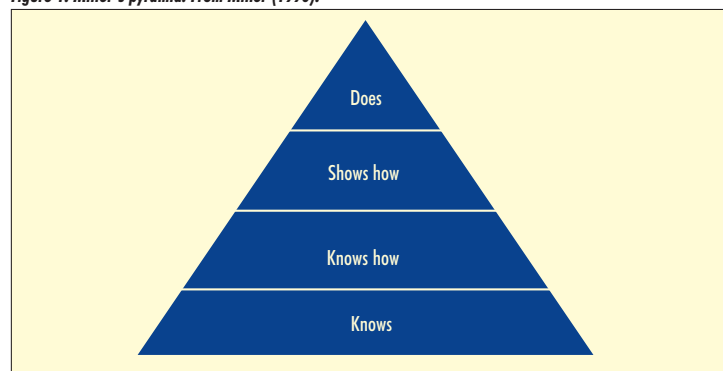
as predominant. Assessment design then inevitably leads to a trade off between individual elements. Thus, traditional approaches to maximize the reliability or reproducibility of assessments can have a negative educational impact on the learner by reducing the opportunity for meaningful developmental feedback. Workplace-based assessments offer high educational impact but might not be as reliable as other highly structured tests such as multiple choice questions.

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

From methods to programmes

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

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



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

Correspondence to: Dr T Swanwick

CLINICAL TEACHING MADE EASY

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

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

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

What to assess?

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

Competency-based

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

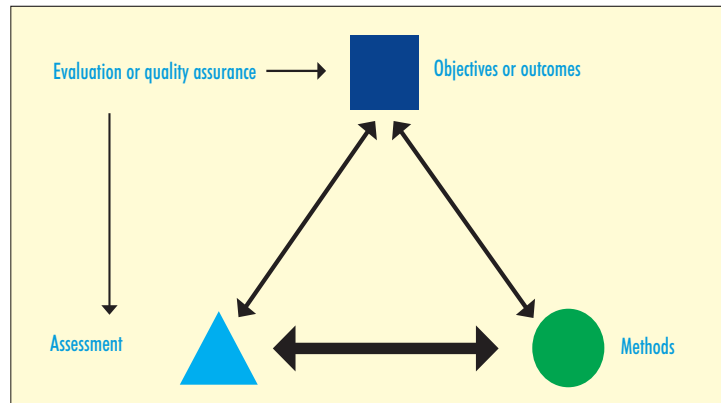


Figure 2. The educational paradigm: integrating teaching, learning and assessment.

Blueprinted

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

Developmental

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

So, workplace-based assessment provides useful formative and developmental

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

How much evidence is enough?

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

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

Confidence in the reproducibility of judgments made on the basis of work-

CHECK
ORIGINAL

CLINICAL TEACHING MADE EASY

place-based assessment can be improved through triangulation. This involves using a range of different methods to collect evidence using multiple raters over a sustained period of time. Triangulation with other assessments external to the workplace is also important and an overarching assessment strategy for each training programme, in which workplace-based assessment is supported by other test methods – such as those of ‘knowledge’ and ‘skills for clinical method’, is essential.

Which methods?

The methods for used for providing feedback and gathering workplace evidence in current use tend to be variations on one of four themes; observations of clinical activities, discussion of clinical cases, analysis of performance data and multi-source feedback.

Observations of clinical activities

Traditionally, clinical skills have been assessed by the ‘long case’ presentation. The problem of case specificity using this technique, limiting the potential to sample widely, has given rise to the mini-clinical evaluation exercise or mini-CEX (Norcini et al, 1995). This tool has been developed to assess the clinical skills that trainees most often use in real patient encounters. It is based on assessment of multiple complete or partial clinical encounters observed by an educational supervisor or other clinician.

The direct observation of procedural skills (DOPS) is another widely used tool, and one of a number of similar instruments based around the assessment of real-life activities where the focus is on the skill with which the activity was performed. ‘The consistent feature is that one or more assessors, who are trained in the assessment of that skill, make a judgment about a real life performance’ (Postgraduate Medical Education and Training Board, 2007).

A raft of other observational tools encompassing a wide range of workplace activities are in also current use including the procedure-based assessment of the Intercollegiate Surgical Curriculum, the mini-imaging interpretation exercise of the Royal College of Radiologists and the assessment of teaching of the Royal College of Psychiatrists.

Discussion of clinical cases

The origin of the use of case-based discussion in UK training assessment systems stemmed from their use in the General Medical Council’s performance procedures (Southgate et al, 2001) deriving originally from chart-stimulated recall oral assessments used in the USA and Canada. Case-based discussion is one of the evidence gathering tools used in workplace-based assessment in the UK foundation programme and is also being used in specialty training programmes such as in medicine, paediatrics and general practice.

Analysis of performance data

Norcini (2003) describes the basis for making a judgment on clinical performance data as having three potential sources; outcomes, process and volume. Outcomes of care, while being the most desirable measure, are limited by problems of attribution (to the individual), complexity, case mix and numbers. This is a particular problem in the assessment of trainee performance.

The process of care is more directly attributable to the individual doctor but effective processes do not necessarily mirror the best patient outcomes. The use of volumes of activity is premised on the basis that the more of a given activity that a doctor performs, the better their quality of care is likely to be. This basis for judgment is typified by the log books of the craft specialties such as surgery.

Multi-source feedback

The aim of using multi-source feedback to assess doctors in the workplace is to view a person’s work from a variety of perspectives. In medical settings, physician colleagues (peers), co-workers and patients can be asked to complete surveys about the doctor. The person being assessed receives feedback based on his/her own aggregate ratings, usually along with average ratings of others being assessed at the same time. There is also a clear opportunity for comparing self-assessment data with those provided by raters.

Multi-source feedback tools can be subdivided into peer-rating tools, such as the mini-PAT (mini peer-rating assessment tool) used in foundation training, and patient satisfaction questionnaires, a significant number of which are in use in the UK (Chisholm and Askham, 2006).

Portfolios

Workplace-based assessments are usually collected within a structured portfolio. A portfolio comprises a dossier of evidence collected over time, which demonstrates a doctor’s education and practice achievements (Wilkinson et al, 2002). There are many portfolio models (Webb et al, 2002) but in essence, if well constructed, a portfolio should chronicle the journey of a learner towards the attainment of professional expertise. A portfolio:

- Aims to serve as the reflective learning log of the learner, available to be shared with his/her educational supervisor
- Demonstrates the learner’s progress towards covering the breadth and depth of the curriculum
- Acts as a repository for assessments
- Provides a framework for learning agreements between learners and teachers
- Charts a learner’s progression and can help in making career choices and decisions.

The majority of portfolios used in medical education are web-based although with significant differences in structure and design between specialties and stage of training.

Quality assurance

Returning to the concept of utility, workplace-based assessment has huge strengths in the area of validity by virtue of its assessment of real or authentic material. Potentially it may have significant educational impact because of the reconnection of teaching and learning. Acceptability and cost-effectiveness are also potential winners but depend largely on how programmes are implemented. There are, however, significant issues with reliability as understood by traditional psychometric approaches. As Southgate et al (2001) point out, ‘establishing the reliability of assessments of performance in the workplace is difficult because they rely on expert judgements of unstandardised material’.

In workplace-based assessment there are several specific threats to reliability:

- Inter-observer variation: the tendency for one observer to mark consistently higher or lower than another
- Intra-observer variation: variation in an observer’s performance for no apparent reason (the ‘good day/bad day’ phenomenon)

CLINICAL TEACHING MADE EASY

- Case specificity: variation in the candidate's performance from one challenge to another, even when they seem to test the same attribute.

In the context of workplace-based assessment it is therefore helpful to reframe reliability as an attempt to maximize 'consistency and comparability'. Baker et al (1992) propose a number of activities that can help to do this, namely:

- Specification of standards, criteria, scoring guides
- Calibration of assessors and moderators
- Moderation of results, particularly those on the borderline
- Training of assessors, with retraining where necessary
- Verification and audit through the collection of assessment data.

It is clear, then, that the implementation of a successful workplace-based assessment programme will require training for assessors, arrangements for calibration, a procedure for the moderation of results and a raft of quality control checks. The more that teachers can be engaged in assessment, for example in selecting methodologies, generating standards and discussing criteria, the more the educational benefits of this powerful form of assessment can be realized.

Conclusions

Workplace-based assessment offers the opportunity to connect teaching, learning and assessment, provides a means for assessment of problematic areas that require evaluation of real performance in practice and is a useful component of an overall assessment programme. In order for its benefits to be realized there needs to be clarity about what is being assessed through the identification of holistically described professional competencies; attention given to the developmental nature of the assessment; a variety of assessment tools used to gather evidence from multiple clinical contexts using multiple raters; and processes in place by which evidence can be collated, synthesized and judged at regular intervals by an educational supervisor to assess the learner's progress with consistency and comparability across assessment programmes maximized through a robust programme of quality assurance. **BJHM**

Conflict of interest: none.

Baker E, O'Neil H, Linn R (1992) Policy and validity prospects for performance-based assessment. *Am Psychol* **48**(12): 1210-18
Chisholm A, Askham J (2006) *What Do You Think of Your Doctor? A review of questionnaires for gathering patients' feedback about their doctor.*

Picker Institute, Europe
Dreyfus H, Dreyfus S (1986) *Mind over machine. The Power of Human Intuition Expertise in the Era of the Computer.* Basil Blackwell, Oxford
Goncz A (1994) Competency based assessment in the professions in Australia. *Assessment in Education* **1**(1): 27-44
Miller G (1990) The assessment of clinical skills/competence/performance. *Acad Med* **65**(Suppl): S63-7
Norcini J (2003) ABC of learning and teaching in medicine. Work based assessment. *BMJ* **326**: 753-5
Norcini J, Blank L, Arnold G, Kimball H (1995) The mini-CEX: a preliminary investigation. *Ann Intern Med* **125**: 795-9
Postgraduate Medical Education and Training Board (2007) *Developing and Maintaining an Assessment System - a guide to good practice.* Postgraduate Medical Education and Training Board, London
Rethans J, Norcini J, Baron-Maldonado M, Blackmore D, Jolly B, La Duca T (2002) The relationship between competence and performance: implications for assessing practice performance. *Med Educ* **36**: 901-9
Southgate L, Cox J, David T et al (2001) The assessment of poorly performing doctors: the development of the assessment programmes for the General Medical Council's Performance Procedures. *Med Educ* **35**(Suppl 1): 2-8
Schuwirth L, van der Vleuten C (2004) Changing education, changing assessment, changing research. *Med Educ* **38**: 805-12
Talbot M (2004) Monkey see, monkey do: a critique of the competency model in graduate medical education. *Med Educ* **38**: 1-7
van der Vleuten C (1996) The assessment of professional competence: developments, research and practical implications. *Advances in Health Science Education* **1**: 41-67
van der Vleuten C, Schuwirth L (2005) Assessing professional competence: from methods to programmes. *Med Educ* **39**: 309-17
Webb C, Gray M, Jasper M, Miller C, McMullan M, Scholes J (2002) Models of portfolios. *Med Educ* **36**(10): 897-8
Wilkinson TJ, Challis M, Hobma SO, Newble DI, Parboosingh JT, Sibbald JG, Wakeford R (2002) The use of portfolios for assessment of the competence and performance of doctors in practice. *Med Educ* **36**: 918-24
William D, Black P (1996) Meanings and consequences: a basis for distinguishing formative and summative functions of assessment? *Br Educ Res J* **22**: 537-48

KEY POINTS

- Workplace-based assessment is now widespread across all specialities and all stages of training.
- Workplace-based assessment offers the opportunity to connect teaching, learning and assessment.
- Workplace-based assessment has a dual function of offering focussed and timely feedback to trainees as well as providing data to support more long range judgments about trainee progress.
- Workplace-based assessment requires new ways of thinking about reliability based on maximizing consistency and comparability.

London Deanery

This series of articles for clinical teachers was originally commissioned as a suite of e-learning modules for the London Deanery. Both the series and e-learning modules were designed and edited by Judy McKimm and Tim Swanwick.

The London Deanery e-learning modules for clinical teachers are open access and available at www.londondeanery.ac.uk/facultydevelopment Each module takes 30-60 minutes to complete and proof of completion is available in the form of a printed certificate.

รศ. ดร.นพ.เชิดศักดิ์ ไอรณิรัตน์

หัวข้อ : Summary

Summary

นพ. เชิดศักดิ์ ไอรณิรัตน์
 ภาควิชาศัลยศาสตร์ คณะแพทยศาสตร์ศิริราชพยาบาล
 มหาวิทยาลัยมหิดล

Experiential Learning Theory

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

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

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

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

Atkins and Murphy (1995)

Five Levels of Reflection

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

Summary of the Workshop

- Morning
 - OSCE
 - Long case exam
- Afternoon
 - Portfolio
 - Clinical performance ratings
 - Workplace-based assessment

Shee.si.mahidol.ac.th

กระดาษบันทึก

กระดาษบันทึก

กระดาษบันทึก

กระดาษบันทึก

กระดาษบันทึก

กระดาษบันทึก

► Question & Comments

ศูนย์ความเป็นเลิศด้านการศึกษาวิทยาศาสตร์สุขภาพ (ศศว)
Siriraj Health science Education Excellence center (SHEE)

ฝ่ายการศึกษาก่อนปริญญา คณะแพทยศาสตร์ศิริราชพยาบาล

สำนักงาน: ตึกอตุลยเดชวิกรม ชั้น 6 (ห้อง 656)

Tel. 02 419 9978, 02 419 96637 Fax. 02 412 3901



shee.si.mahidol.ac.th



shee.mahidol@gmail.com



mahidol.shee



SHEE FC



Siriraj Health science Education Excellence center