



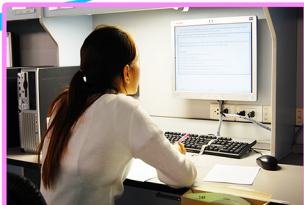
Mahidol University  
Faculty of Medicine  
Siriraj Hospital

ศูนย์ความเป็นเลิศด้านการศึกษาวิทยาศาสตร์สุขภาพ (ศตว)  
คณะแพทยศาสตร์ศิริราชพยาบาล มหาวิทยาลัยมหิดล

# Advances in Competency-based Assessment

การประเมินที่ครอบคลุม ถูกต้อง  
เป็นธรรม ทันสมัย และได้มาตรฐาน

## เอกสารประกอบการอบรม



## สารบัญ

	หน้า
กำหนดการ .....	1
รายชื่อผู้ร่วมอบรม Part I .....	3
รายชื่อผู้ร่วมอบรม Part II .....	5
เอกสารประกอบการอบรม	
24 Apr 2018.....	7
หัวข้อ : Future directions for medical competency assessment.....	9
หัวข้อ : New trends in written exam .....	15
หัวข้อ : IT in assessment.....	23
หัวข้อ : Advanced item analytic techniques.....	27
หัวข้อ : Setting a passing standard.....	33
หัวข้อ : Entrust able Professional Activities (EPA).....	39
25 Apr 2018.....	45
หัวข้อ : Technical skills assessment .....	47
หัวข้อ : Communication skills assessment .....	53
หัวข้อ : Assessing leadership and team management.....	113
หัวข้อ : Rubric scale development .....	127
กระดาษบันทึก .....	146
ช่องทางการติดต่อสื่อสาร.....	149



## (ร่าง) กำหนดการอบรมเชิงปฏิบัติ เรื่อง Advances in competency-based assessment

วันที่ 24 - 25 เมษายน 2561 ณ ห้องบรรยาย 3A01 ชั้น 3A อาคารศรีสวรินทิรา คณะแพทยศาสตร์ศิริราชพยาบาล

วันอังคารที่ 24 เมษายน พ.ศ. 2561 Part I: หลักการวัดและประเมินผล (อบรมภาคทฤษฎี)		
09.00 – 10.00 น.	ทิศทางการเปลี่ยนแปลงในการประเมินความสามารถทางการแพทย์ Future directions for medical competency assessment	ศ. พญ.บุญมี สถาปัตยกรรมศาสตร์*
10.15 – 11.00 น.	แนวทางใหม่ในการสอบข้อเขียน New trends in written exam	รศ.ดร. นพ. เชิดศักดิ์ ไอรอมณีรัตน์
11.00 - 12.00 น.	การใช้เทคโนโลยีสารสนเทศในการประเมินผล IT in assessment	คุณทศพร มาสวัสดิ์ รศ.ดร. นพ. เชิดศักดิ์ ไอรอมณีรัตน์
12.00 – 13.00 น.	พักรับประทานอาหารกลางวัน	
13.00 – 14.00 น.	การวิเคราะห์คะแนนสอบขั้นสูง Advanced item analytic techniques	รศ.ดร. นพ. เชิดศักดิ์ ไอรอมณีรัตน์
14.00 – 15.00 น.	การตั้งเกณฑ์ผ่านการสอบ Setting a passing standard	รศ.ดร. นพ. เชิดศักดิ์ ไอรอมณีรัตน์
15.15 – 16.15 น.	การกำหนดกิจกรรมทางวิชาชีพที่ไว้ใจให้ผู้เรียนทำได้ Entrust able Professional Activities (EPA)	ผศ. พญ.กษณา รักษมณี
วันพุธที่ 25 เมษายน พ.ศ. 2561 Part II: แนวทางการประเมินทักษะขั้นสูง (อบรมภาคทฤษฎี + workshop)		
09.00 – 10.00 น.	การประเมินทักษะการทำหัตถการ Technical skills assessment	ผศ. พญ.กษณา รักษมณี
10.15 – 11.00 น.	การประเมินทักษะการสื่อสาร Communication skills assessment	อ. พญ.กมลทิพย์ เลิศชัยสถาพร* รศ.ดร. นพ. เชิดศักดิ์ ไอรอมณีรัตน์
11.00 - 12.00 น.	การประเมินความเป็นผู้นำและการบริหารทีม Assessing leadership and team management	ผศ. พญ.ธัชววรรณ จิระดิวานนท์
12.00 – 13.00 น.	พักรับประทานอาหารกลางวัน	
13.00 – 14.00 น.	แนวทางการสร้างแบบประเมิน Rubric scale development	ผศ. นพ.ตรีภพ เลิศบรรณพงษ์
14.00 – 15.00 น.	กิจกรรมกลุ่ม: สร้างแบบประเมิน Group activity: Rubric scale development	รศ.ดร. นพ. เชิดศักดิ์ ไอรอมณีรัตน์ ผศ. นพ.ตรีภพ เลิศบรรณพงษ์ ผศ. พญ.ธัชววรรณ จิระดิวานนท์ ผศ. พญ.กษณา รักษมณี คุณทศพร มาสวัสดิ์
15.15 – 16.15 น.	การนำเสนอเครื่องมือประเมิน Presentation of rating instruments	รศ.ดร. นพ. เชิดศักดิ์ ไอรอมณีรัตน์ ผศ. นพ.ตรีภพ เลิศบรรณพงษ์ ผศ. พญ.ธัชววรรณ จิระดิวานนท์ ผศ. พญ.กษณา รักษมณี คุณทศพร มาสวัสดิ์

หมายเหตุ: กำหนดการอาจมีการเปลี่ยนแปลงตามความเหมาะสม





## รายชื่อผู้ร่วมอบรม

## Part I: หลักการวัดและประเมินผล

ลำดับ	คำนำหน้า	ชื่อ	สกุล	สังกัด	หน่วยงาน/ภาควิชา
1	พญ.	กนกรัตน์	สุวรรณละอออง	ศูนย์แพทยศาสตรศึกษาชั้นคลินิก โรงพยาบาลสงขลา	กลุ่มงานอายุรกรรม
2	พญ.	กนกวรรณ	ศรีรักษา	โรงพยาบาลขอนแก่น	กุมารเวชกรรม
3	ผศ.ดร.	กรแก้ว	จันทภาษา	คณะเภสัชศาสตร์ มหาวิทยาลัยขอนแก่น	เภสัชกรรมชุมชน
4	นพ.	กฤษ	หาญชาญชัยกุล	โรงพยาบาลสรรพสิทธิประสงค์	กลุ่มงานสูตินรีเวชกรรม
5	ดร.	ก่อเกียรติ	ธีระกิตต์ธนากุล	คณะแพทยศาสตร์ มหาวิทยาลัยราชภัฏวชิราวุฒวิทยาลัย	ฝ่ายวิจัยและเวชศาสตร์ชุมชน
6	พญ.	กัลยา	หวังเรืองสถิตย์	โรงพยาบาลพุทธชินราช พิษณุโลก	กลุ่มงานวิสัญญีวิทยา
7	พญ.	กัลยาณี	อาสนศักดิ์	ศูนย์แพทยศาสตรศึกษาชั้นคลินิก โรงพยาบาลสงขลา	กลุ่มงานกุมารเวชกรรม
8	นพ.	กิตติพงศ์	มาศเกษม	โรงพยาบาลเจริญกรุงประชารักษ์	กลุ่มงานกุมารเวชกรรม
9	ผศ. นพ.	กิติกุล	สีละวงศ์	คณะแพทยศาสตร์โรงพยาบาลรามาธิบดี มหาวิทยาลัยมหิดล	ภาควิชาจักษุวิทยา
10	ผศ. นพ.	จปรัฐ	ปรีชาพานิช	คณะแพทยศาสตร์ศิริราชพยาบาล	ภาควิชาสูติศาสตร์-นรีเวชวิทยา
11	พญ.	จุไรรัตน์	บัวภิบาล	สถาบันสิรินธรเพื่อการฟื้นฟูสมรรถภาพทางการแพทย์แห่งชาติ	สาขา เวชศาสตร์ฟื้นฟู
12	รศ.ดร. นพ.	ชัยเลิศ	พิชิตพรชัย	คณะแพทยศาสตร์ศิริราชพยาบาล	ภาควิชาสูติศาสตร์
13	พญ.	ติยารัตน์	ชยันกิจ	โรงพยาบาลพุทธชินราช พิษณุโลก	หน่วยจิตเวชศาสตร์
14	ผศ. พญ.	ทานตะวัน	อวิรุทธ์วรกุล	คณะแพทยศาสตร์โรงพยาบาลรามาธิบดี มหาวิทยาลัยมหิดล	จิตเวชศาสตร์
15	อ. นพ.	ธัญชัย	เพชรภาค	คณะแพทยศาสตร์โรงพยาบาลรามาธิบดี มหาวิทยาลัยมหิดล	ภาควิชาอายุรศาสตร์
16	นพ.	ธัญ	ลักษณะานนท์	โรงพยาบาลแพร่	เวชศาสตร์ฉุกเฉิน
17	อ. นพ.	ธีรพงศ์	โตเจริญโชค	คณะแพทยศาสตร์ศิริราชพยาบาล	ภาควิชาศัลยศาสตร์
18	นพ.	ธีรวุฒิ	รัตนพิชญชัย	วิทยาลัยแพทยศาสตร์ มหาวิทยาลัยรังสิต	ภาควิชาจิตเวชศาสตร์
19	พญ.	นันทรา	สุวันทรัตน์	วิทยาลัยแพทยศาสตร์นานาชาติจุฬาภรณ์ มหาวิทยาลัยธรรมศาสตร์	สาขา อนุสาขาอายุรศาสตร์โรคติดเชื้อ
20	ดร.	นิโรบล	กนกสุนทรรัตน์	คณะแพทยศาสตร์โรงพยาบาลรามาธิบดี มหาวิทยาลัยมหิดล	สาขาการพยาบาลผู้ใหญ่และผู้สูงอายุ
21	พญ.	นิตากร	ไวดาบ	โรงพยาบาลเจริญกรุงประชารักษ์	กลุ่มงานกุมารเวชกรรม
22	นพ.	บุรภัทร	สังข์ทอง	คณะแพทยศาสตร์ มหาวิทยาลัยสงขลานครินทร์	ภาควิชาศัลยศาสตร์
23	ดร.	เบญจพร	ศิลาภิรักษ์	โรงพยาบาลขอนแก่น	กลุ่มงานเภสัชกรรม
24	ผศ. พญ.	ปรีชญา	วงษ์กระจำง	คณะแพทยศาสตร์ศิริราชพยาบาล	ภาควิชาพยาธิวิทยาคลินิก
25	ดร. พญ.	ปวีณา	พิทักษ์สุรชัย	คณะแพทยศาสตร์ศิริราชพยาบาล	ภาควิชาโสต นาสิก ลาริงซ์วิทยา
26	พญ.	ปองทอง	ปุรานิธิ	คณะแพทยศาสตร์โรงพยาบาลรามาธิบดี มหาวิทยาลัยมหิดล	กุมารเวชศาสตร์
27	พญ.	ปิยนุช	บุรณพร	ศูนย์แพทยศาสตรศึกษาชั้นคลินิก โรงพยาบาลสงขลา	กลุ่มงานโสต คอ นาสิก
28	รศ. พญ.	พนัสยา	เอียรธาดากุล	คณะแพทยศาสตร์ศิริราชพยาบาล	ภาควิชาพยาธิวิทยาคลินิก
29	รศ. พญ.	พรจิรา	ปรีวีชรากุล	คณะแพทยศาสตร์ศิริราชพยาบาล	ภาควิชาจิตเวชศาสตร์
30	รศ.ดร. นพ.	พรพรต	ลัมประเสริฐ	คณะแพทยศาสตร์เทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง	ภาควิชาพยาธิวิทยา
31	พญ.	พวงเพชร	ศิริเลิศธนาพันธ์	ศูนย์แพทยศาสตรศึกษาชั้นคลินิก โรงพยาบาลสงขลา	กลุ่มงานเวชกรรมฟื้นฟู
32	นางสาว	พิริยาพร	พลอยทิพย์	คณะแพทยศาสตร์ศิริราชพยาบาล	โรงเรียนผู้ช่วยพยาบาล
33	พญ.	เพ็ญพรชา	อุดมทรัพย์	ศูนย์แพทยศาสตรศึกษาชั้นคลินิก โรงพยาบาลนครพิงค์	เวชศาสตร์ชุมชน เวชศาสตร์ครอบครัว
34	นพ.	มนต์ชัย	ศิริบำรุงวงศ์	โรงพยาบาลเลิดสิน	ภาควิชาอายุรศาสตร์
35	นพ.	ราศิน	วรวงศากุล	คณะแพทยศาสตร์โรงพยาบาลรามาธิบดี มหาวิทยาลัยมหิดล	ภาควิชารังสีวิทยาสาขารังสีรักษาและมะเร็งวิทยา
36	พญ.	รุจิรา	ลีธนาภรณ์	ศูนย์แพทยศาสตรศึกษาชั้นคลินิก โรงพยาบาลสงขลา	กลุ่มงานอายุรกรรม
37	พญ.	ลัดดา	จันทร์แรม	โรงพยาบาลร้อยเอ็ด	งานเวชศาสตร์ครอบครัว
38	รศ. พญ.	วรพรรณ	เสนาณรงค์	คณะแพทยศาสตร์ศิริราชพยาบาล	ภาควิชาอายุรศาสตร์
39	อ. พญ.	วรรณภา	อาจจงค์	โรงพยาบาลพุทธชินราช พิษณุโลก	กุมารเวชกรรม
40	ดร. นพ.	วรุฒม์	พงศาพิชญ์	คณะแพทยศาสตร์ศิริราชพยาบาล	ภาควิชาโสต นาสิก ลาริงซ์วิทยา
41	นพ.	วสุ	เดชะวัฒนากุล	ศูนย์แพทยศาสตรศึกษาชั้นคลินิก โรงพยาบาลพะเยา	อุบัติเหตุและฉุกเฉิน
42	พญ.	วันหนึ่ง	คณานนท์	โรงพยาบาลพุทธชินราช พิษณุโลก	กลุ่มงานวิสัญญีวิทยา
43	พญ.	วันวิสาข์	สินธุประสิทธิ์	โรงพยาบาลขอนแก่น	วิสัญญีวิทยา
44	ดร. นพ.	วิษั	เกษมทรัพย์	คณะแพทยศาสตร์โรงพยาบาลรามาธิบดี มหาวิทยาลัยมหิดล	ภาควิชาเวชศาสตร์ชุมชน
45	นพ.	ศรีลค์	สวัสดิ์วินิช	ศูนย์แพทยศาสตรศึกษาชั้นคลินิก โรงพยาบาลสงขลา	กลุ่มงานวิสัญญีวิทยา
46	พญ.	ศศิธร	ธนศรีภักติกุล	โรงพยาบาลขอนแก่น	วิสัญญีวิทยา
47	พญ.	ศิดาญ	สุริยะ	คณะแพทยศาสตร์ มหาวิทยาลัยนครสวรรค์	อายุรศาสตร์
48	นางสาว	ศิริประภา	ฤชัย	โรงพยาบาลบำรุงราษฎร์	ศูนย์พัฒนาและฝึกอบรมบุคลากร

ลำดับ	คำนำหน้า	ชื่อ	สกุล	สังกัด	หน่วยงาน/ภาควิชา
49	รศ. พญ.	ศิริพร	ปิติมานะอารี	คณะแพทยศาสตร์ศิริราชพยาบาล	ภาควิชาวิสัญญีวิทยา
50	พญ.	ศิวะพร	เกียรติธนะบำรุง	คณะแพทยศาสตร์โรงพยาบาลรามาธิบดี มหาวิทยาลัยมหิดล	ภาควิชา โสต ศอ นาสิกวิทยา
51	นาง	สมฤทัย	เพชรประยูร	คณะแพทยศาสตร์ศิริราชพยาบาล	โรงเรียนผู้ช่วยพยาบาล
52	รศ. นพ.	สามารถ	ภคกษมา	คณะแพทยศาสตร์โรงพยาบาลรามาธิบดี มหาวิทยาลัยมหิดล	กุมารเวชศาสตร์
53	รศ. นพ.	สิทธิพร	ศรีนวลนิต	คณะแพทยศาสตร์ศิริราชพยาบาล	ภาควิชาศัลยศาสตร์ สาขาวิชาศัลยศาสตร์ยูโร
54	พญ.	สุธิดา	สัมฤทธิ์	คณะแพทยศาสตร์โรงพยาบาลรามาธิบดี มหาวิทยาลัยมหิดล	เวชศาสตร์ครอบครัว
55	ผศ. พญ.	สุภาวดี	มากะนัดถ์	คณะแพทยศาสตร์ มหาวิทยาลัยนครสวรรค์	อายุรศาสตร์
56	ผศ. พญ.	สุภาวรรณ	เศรษฐบรรจง	คณะแพทยศาสตร์ศิริราชพยาบาล	ภาควิชานิติเวชศาสตร์
57	พญ.	สุมาลิน	ชมคช	ศูนย์แพทยศาสตร์ศึกษาชั้นคลินิก โรงพยาบาลสงขลา	กลุ่มงาน อุบัติเหตุและฉุกเฉิน
58	นพ.	สุรศักดิ์	รื่นสุข	โรงพยาบาลราชบุรี	กลุ่มงานนิติเวชศาสตร์
59	พญ.	หทัยทิพย์	ต่างงาม	ศูนย์แพทยศาสตร์ศึกษาชั้นคลินิก โรงพยาบาลนครพิงค์	กุมารเวชศาสตร์
60	นพ.	องอาจ	สิกขมาน	วิทยาลัยแพทยศาสตร์ มหาวิทยาลัยรังสิต	ภาควิชาเวชศาสตร์ครอบครัว
61	รศ. พญ.	อดิพร	ดวงทอง	คณะแพทยศาสตร์ศิริราชพยาบาล	ภาควิชาจักษุวิทยา
62	อ. นพ.	อนิรุต	วรวาท	คณะแพทยศาสตร์ศิริราชพยาบาล	นิติเวชศาสตร์
63	นาง	อมรรัตน์	จวบสมัย	คณะแพทยศาสตร์ศิริราชพยาบาล	โรงเรียนผู้ช่วยพยาบาล
64	ดร.	อรุทัย	พรหมสงค์	คณะแพทยศาสตร์ มหาวิทยาลัยราชภัฏวชิราวุธานุสรินทร์	ชีวการแพทย์
65	รศ. พญ.	อรุโณทัย	ศิริอัครกุล	คณะแพทยศาสตร์ศิริราชพยาบาล	ภาควิชาวิสัญญีวิทยา
66	รศ.ดร. นพ.	อัษฎา	เมธเศรษฐ	คณะแพทยศาสตร์ศิริราชพยาบาล	ภาควิชาศัลยศาสตร์
67	นาง	อุมาพร	ภูมิศรี	คณะแพทยศาสตร์ศิริราชพยาบาล	โรงเรียนผู้ช่วยพยาบาล

รายชื่อผู้ร่วมอบรม

Part II: แนวทางการประเมินทักษะขั้นสูง

กลุ่มที่ 1					
ลำดับ	คำนำหน้า	ชื่อ	สกุล	สังกัด	หน่วยงาน/ภาควิชา
1	พญ.	กนกวรรณ	ศรีรักษา	โรงพยาบาลขอนแก่น	กุมารเวชกรรม
2	พญ.	กัลยาณี	อาสาศักดิ์	ศูนย์แพทยศาสตรศึกษาชั้นคลินิก โรงพยาบาลสงขลา	กลุ่มงานกุมารเวชกรรม
3	ผศ. นพ.	จปรัฐ	ปรีชาพานิช	คณะแพทยศาสตร์ศิริราชพยาบาล	ภาควิชาสูติศาสตร์-รีเวชวิทยา
4	พญ.	ปองทอง	ปุราณีย์	คณะแพทยศาสตร์โรงพยาบาลรามธิบดี มหาวิทยาลัยมหิดล	กุมารเวชศาสตร์
5	อ. พญ.	วรรณณา	อาจองค์	โรงพยาบาลพุทธชินราช พิษณุโลก	กุมารเวชกรรม
6	ผศ.ดร.	กรแก้ว	จันทภาษา	คณะเภสัชศาสตร์ มหาวิทยาลัยขอนแก่น	เภสัชกรรมชุมชน
7	พญ.	ติยารัตน์	ชัยนิกิจ	โรงพยาบาลพุทธชินราช พิษณุโลก	หน่วยจิตเวชศาสตร์
8	นพ.	ธีรวัฒน์	รัตนพิชญชัย	วิทยาลัยแพทยศาสตร์ มหาวิทยาลัยรังสิต	ภาควิชาจิตเวชศาสตร์
9	ดร. นพ.	วิรัช	เกษมทรัพย์	คณะแพทยศาสตร์โรงพยาบาลรามธิบดี มหาวิทยาลัยมหิดล	ภาควิชาเวชศาสตร์ชุมชน
10	พญ.	สุธิดา	สัมฤทธิ์	คณะแพทยศาสตร์โรงพยาบาลรามธิบดี มหาวิทยาลัยมหิดล	เวชศาสตร์ครอบครัว
11	นพ.	องอาจ	สิกขมาน	วิทยาลัยแพทยศาสตร์ มหาวิทยาลัยรังสิต	ภาควิชาเวชศาสตร์ครอบครัว

กลุ่มที่ 2					
ลำดับ	คำนำหน้า	ชื่อ	สกุล	สังกัด	หน่วยงาน/ภาควิชา
1	พญ.	กนกรัตน์	สุวรรณละออง	ศูนย์แพทยศาสตรศึกษาชั้นคลินิก โรงพยาบาลสงขลา	กลุ่มงานอายุรกรรม
2	พญ.	นันตรา	สุวันทรัตน์	วิทยาลัยแพทยศาสตร์นานาชาติจุฬาภรณ์ มหาวิทยาลัยธรรมศาสตร์	สาขา อนุสาขาศัลยกรรมโรคติดเชื้อ
3	นพ.	มนต์ชัย	ศิริบำรุงวงศ์	โรงพยาบาลเลิดสิน	ภาควิชาอายุรศาสตร์
4	พญ.	รุจิรา	ลิธนาภรณ์	ศูนย์แพทยศาสตรศึกษาชั้นคลินิก โรงพยาบาลสงขลา	กลุ่มงานอายุรกรรม
5	พญ.	ศิดาญ	สุริยะ	คณะแพทยศาสตร์ มหาวิทยาลัยนเรศวร	อายุรศาสตร์
6	พญ.	สุมาลิน	ชุมคช	ศูนย์แพทยศาสตรศึกษาชั้นคลินิก โรงพยาบาลสงขลา	กลุ่มงาน อุบัติเหตุและฉุกเฉิน
7	พญ.	จุไรรัตน์	บัวภิบาล	สถาบันสิรินธรเพื่อการฟื้นฟูสมรรถภาพทางการแพทย์แห่งชาติ	สาขา เวชศาสตร์ฟื้นฟู
8	พญ.	ทวงเพชร	ศิริเลิศรณานนท์	ศูนย์แพทยศาสตรศึกษาชั้นคลินิก โรงพยาบาลสงขลา	กลุ่มงานเวชกรรมฟื้นฟู
9	นพ.	บุรภัทร	สังข์ทอง	คณะแพทยศาสตร์ มหาวิทยาลัยสงขลานครินทร์	ภาควิชาศัลยศาสตร์
10	นพ.	สุศักดิ์	รินสุข	โรงพยาบาลราชบุรี	กลุ่มงานนิติเวชศาสตร์
11	อ. นพ.	อนันต์	วรวาท	คณะแพทยศาสตร์ศิริราชพยาบาล	นิติเวชศาสตร์

กลุ่มที่ 3					
ลำดับ	คำนำหน้า	ชื่อ	สกุล	สังกัด	หน่วยงาน/ภาควิชา
1	ผศ. นพ.	กิติกุล	ลิละวงศ์	คณะแพทยศาสตร์โรงพยาบาลรามธิบดี มหาวิทยาลัยมหิดล	ภาควิชาจักษุวิทยา
2	ดร. พญ.	ปวีณา	พิทักษ์สุรชัย	คณะแพทยศาสตร์ศิริราชพยาบาล	ภาควิชาโสต นาสิก ลาริงซ์วิทยา
3	พญ.	ปิยนุช	บุรณพร	ศูนย์แพทยศาสตรศึกษาชั้นคลินิก โรงพยาบาลสงขลา	กลุ่มงานโสต ศอ นาสิก
4	ดร. นพ.	วรุฒม์	พงศาพิชญ์	คณะแพทยศาสตร์ศิริราชพยาบาล	ภาควิชาโสต นาสิก ลาริงซ์วิทยา
5	พญ.	ศิวะพร	เกียรติฉัตรบำรุง	คณะแพทยศาสตร์โรงพยาบาลรามธิบดี มหาวิทยาลัยมหิดล	ภาควิชา โสต ศอ นาสิกวิทยา
6	รศ. พญ.	อติพร	ดวงทอง	คณะแพทยศาสตร์ศิริราชพยาบาล	ภาควิชาจักษุวิทยา
7	นพ.	ราศิน	วรวงศากุล	คณะแพทยศาสตร์โรงพยาบาลรามธิบดี มหาวิทยาลัยมหิดล	ภาควิชารังสีวิทยา สาขารังสีรักษาและมะเร็งวิทยา
8	นพ.	ศรีลัก	สวัสดิ์นิช	ศูนย์แพทยศาสตรศึกษาชั้นคลินิก โรงพยาบาลสงขลา	กลุ่มงานวิสัญญีวิทยา
9	รศ. พญ.	ศิริพร	ปิติมานะอารี	คณะแพทยศาสตร์ศิริราชพยาบาล	ภาควิชาวิสัญญีวิทยา
10	พญ.	วันวิสาข์	สินธุประสิทธิ์	โรงพยาบาลขอนแก่น	วิสัญญีวิทยา
11	พญ.	ศศิธร	ธนศรีภักติกุล	โรงพยาบาลขอนแก่น	วิสัญญีวิทยา

กลุ่มที่ 4					
ลำดับ	คำนำหน้า	ชื่อ	สกุล	สังกัด	หน่วยงาน/ภาควิชา
1	รศ.ดร. นพ.	ชัยเลิศ	พิชิตพรชัย	คณะแพทยศาสตร์ศิริราชพยาบาล	ภาควิชาสรีรวิทยา
2	ดร.	นิโรบล	กนกสุนทรรัตน์	คณะแพทยศาสตร์โรงพยาบาลรามธิบดี มหาวิทยาลัยมหิดล	สาขาการพยาบาลผู้ใหญ่และผู้สูงอายุ
3	ดร.	เบญจพร	ศิลาภิรักษ์	โรงพยาบาลขอนแก่น	กลุ่มงานเภสัชกรรม
4	ผศ. พญ.	ปรีชญา	วงศ์กระจ่าง	คณะแพทยศาสตร์ศิริราชพยาบาล	ภาควิชาพยาธิวิทยาคลินิก
5	รศ. พญ.	พนัสยา	เอียรธาดากุล	คณะแพทยศาสตร์ศิริราชพยาบาล	ภาควิชาพยาธิวิทยาคลินิก
6	นางสาว	พิริยาพร	พลอยทิพย์	คณะแพทยศาสตร์ศิริราชพยาบาล	โรงเรียนผู้ช่วยพยาบาล
7	นางสาว	ศิริประภา	ฤกษ์ชัย	โรงพยาบาลบำรุงราษฎร์	ศูนย์พัฒนาและฝึกอบรมบุคลากร
8	นาง	สมฤทัย	เพชรประยูร	คณะแพทยศาสตร์ศิริราชพยาบาล	โรงเรียนผู้ช่วยพยาบาล
9	นาง	อมรรัตน์	จวบสมัย	คณะแพทยศาสตร์ศิริราชพยาบาล	โรงเรียนผู้ช่วยพยาบาล
10	นาง	อุมาพร	ภูฎีศรี	คณะแพทยศาสตร์ศิริราชพยาบาล	โรงเรียนผู้ช่วยพยาบาล



## เอกสารประกอบการอบรม



24 Apr 2018



24 Apr 2018

หัวข้อ : Future directions for medical competency assessment

**FUTURE DIRECTIONS FOR  
MEDICAL COMPETENCY  
ASSESSMENT**

Boonmee Sathapatavongs  
April 24, 2018

**Assessment**

? WHY  
? WHAT  
? HOW  
? WHEN  
? WHO

**Assessment Methods (Tools)**

- Validity
- Reliability
- Feasibility
- Impact on future learning and practice (educational & catalytic effect)
- Acceptability to learners and faculty members

**Does**  
Assessment in work environment; focus on overall performance, not components  
Direct observation of learner performance, portfolios, clinical triple jump, 360 assessment, clinical competency exams, videotaping with follow-up review

**Shows How**  
Assessment in controlled situations  
OSCEs, simulations, lab practicals, standardized patients

**Knows How**  
Assess capacity for clinical-context application  
Essays, triple jump, case-based MCQs

**Knows**  
Test factual recognition  
Context-free MCQs, reports written by students, oral exams

**Trends and Opportunities in Medical Education:  
Aligning to Societal Needs and Expectations**

- Organizations and institutions develop and adopt educational frameworks eg. 7-star doctor, CanMeds framework
- Shifting from apprenticeship models of medical education to competency-based medical education, outcome-based medical education requiring the need to appropriately communicate specific goals and objectives to both learners and faculty regarding educational expectations
- Patients and communities calling for increased compassion and caring from healthcare providers
- The Globalization of medical education, where educational frameworks, accreditation standards(WFME), curricular methods, and assessment techniques are spilling across borders and adopted

Arch Med Health Sci 2017;5:154-6

**What is seven star doctor?**

Examples of Learning Outcomes

1. Care Provider
2. Decision-maker
3. Communicator
4. Community Leader
5. Manager
6. Researcher
7. Life-long Learner

THE CANMEDS ROLES FRAMEWORK

### Professional Standards of Medical Practitioners – Thai Medical Council 2012

- Professional habits, attitudes, moral and ethics
- Communication and interpersonal skills
- Scientific knowledge of Medicine
- Patient care
- Health promotion and health care system: individual, community and population health
- Continuous professional development

### Why do we assess our students?

- Are we interested in what outcomes a student has met or how a student can *better* meet the desired outcomes? Key differentiating factors between the two approaches are the difference in time required to undertake the assessment and the provision of comprehensive feedback.
- **assessment for learning VS assessment of learning**

### Assessment for Learning

- Primarily aimed at aiding learning through constructive feedback that identifies areas for development. Alternative terms are **Formative or Low-stakes assessment**. **Lower reliability** is acceptable for individual assessments as they can and should be **repeated frequently**. This increases their reliability and helps to document progress. Such assessments are ideally undertaken in the workplace
- Guiding future learning, providing reassurance, promoting reflection and shaping values

### Assessment of Learning

- Primarily aimed at determining a level of competence to permit progression of training or certification. Such assessments are undertaken infrequently (for example, examinations) and must have **high reliability** as they often form the basis of pass/fail decisions. Alternative terms are **Summative or Highstakes** assessment
- Making an overall judgment about competence, fitness to practice, or qualification for advancement to higher levels of responsibility

### What are we trying to assess?

- Overall professional competence: medical knowledge, technical skill, clinical reasoning, professionalism, communication and interpersonal skills including teamwork, and reflection (To be noted that doctor's communications skills, or lack of, have accounted for the most common complaint to the Medical Council)
- Assessment of competence should provide insight into actual performance in the clinical setting as well as the capacity to adapt to change and generate new knowledge
- Postgraduate Medical training have introduced **Mini Case Based Discussion (Mini-Cx)** and **Directly Observed Practical Skill (DOPS)** to their routine evaluations, to reflect the complexity of real life medical scenarios using concepts of entrustable professional activities (**EPA**) and **milestones**

### Who should assess students?

- Traditionally, teachers as assessors
- Trends are to move towards multisource assessment, incorporating self and peer assessment, so that deeper and more authentic learning can occur
- Self-assessment fosters reflection and revision, leading to a more desirable learning outcome through collaboration and a mutual understanding of expectations
- Students develop the capacity to make judgements about their own work and that of others in order to become effective continuing learners and practitioners
- This form of assessment depends on trust, needing time to develop



### Workplace-Based Assessment (WPBA) or Work-Based Assessment (WBA)

The assessment of working practices based on what trainees actually **do** in the workplace.

Various methods such as :

- Mini-Clinical Evaluation Exercise (mini-CEX)
- Direct Observation of Procedural Skills (DOPS)/Clinical Encounters(DOCE)
- Case-Based Discussions (CbD)
- Mini-Peer Assessment tool (Mini-PAT)
- Multi-Source Feedback (MSF)
- Patient Survey (PS)
- Portfolios

GMC, UK: A guide for implementation April 2010

### Mini-Clinical Evaluation Exercise(Mini-CEX)

- Mini-CEX is intended to facilitate formative assessment of core clinical skills. It can be used by faculty as a routine, seamless evaluation of trainees in real life setting.
- The Mini-CEX is a 10- to 20-minute direct observation assessment or "snapshot" of a trainee-patient interaction. Faculty are encouraged to perform at least one per clinical rotation. To be most useful, faculty should provide timely and specific feedback to the trainee after each assessment of a trainee-patient encounter. The results must be recorded, preferably in the portfolio for further follow-up

### PORTFOLIO DEFINITION

- A purposeful collection of student work that exhibits the student's efforts, progress, and achievements in one or more areas. The collection must include student participation in selecting contents, the criteria for selection, the criteria for judging merit, and **evidence of learner self-reflection** (Paulson FL, Paulson PR, Meyer CA (1991) What makes a portfolio a portfolio? *Educational Leadership* 48: 60-3.)
- A **documentation of learning and an articulation of what has been learned** (Snadden D, Thomas MI (1998) Portfolio learning in general practice vocational training- does it work? *Medical Education* 32: 401-6)

### What is not a Portfolio?

A portfolio is not the same as:

- a logbook, which simply records specific activities undertaken
- a CV, which just provides a summary of an individual's employment history and qualifications
- a course log, which records training and targeted activities for a specific course
- a training folder, which collects evidence of participation in training (e.g. certificates, programs).

**Portfolio must contain what learner has learned and is able to apply that in real life practice "self reflection"**

### Portfolio : Formative vs. Summative

- Writing the portfolio itself requires engagement in a process of reflection and critical self-awareness. Its creation therefore constitutes an educational process, and this aspect needs to be recognized over and above the outcomes of learning that are identified and evidenced in the physical material contained in the portfolio (Challis M, *Medical Teacher*1999; 21: 370-86)
- Limitations: time-consuming process for learners and teachers, unrecognition of the relevance of reflective learning to their practice , by learners, interrater reliability, variability of content and criterion-related validity

### Multi-source Feedback

- An important tool for obtaining evidence about interpersonal and communication skills, judgement, professional behaviour and clinical practice.
- All those working with a trainee (including trainers, fellow trainees and senior nurses/allied health professionals) are asked to rate the trainee's performance in various domains such as teamwork, communication and decision-making towards the end of a training placement.
- These ratings are collated and fed back to the trainee by their supervisor. This forms an important part of the **appraisal** process. Alternative terms are peer-review or **360° feedback**

### Workplace-Based Assessment (WPBA)

- Not sufficiently reliable to stand alone and that it should be used together with endpoint high stakes 'know how' and 'show how' assessments of learning.
- It is important to avoid the danger that assessments in the workplace are seen as simply opportunistic. They need to be appropriately utilized by both trainee and trainer/assessor through dialogue and properly structured learning plans

GMC, UK: A guide for implementation April 2010

### Entrustable Professional Activity (EPA)

- Entrustment refers to the ability to effectively perform a professional activity **without supervision**
- Bring trust and supervision into assessment which are intuitive for faculty working with trainees
- Entrustment decisions allow inference about a learner's competence
- "Trust reflects a dimension of competence that reaches further than observed ability. It includes the real outcome of training – the quality of care"

### Attributes of Entrustable Professional Activities (EPA)

- Essential professional work in a given context
- Requiring adequate KAP through training
- Leading to recognized output of professional work
- Usually confined to qualified personnels
- Independently executable within a time frame
- Observable and measurable in the process and outcome
- Reflecting one or more competencies to be acquired

### Entrustable Professional Activity (EPA) cont.

- One of the key features of EPAs is the link between authentic, often-everyday tasks of a profession and the opportunities to observe and assess learners' performance completing those tasks
- Entrustment requires that faculty and assessors make a judgment that integrates learner performance with assessor expectations and the nature of the task/EPA
- EPAs provide a framework for granting responsibility as soon as learners are ready for it
- **Variability among faculty judges is a threat to the validity of these entrustment judgments**

### Level of Supervision: Milestone setting

- Level 1: not allowed to practice the EPA
- Level 2: practice with full supervision
- Level 3: practice with supervision on demand
- Level 4: "unsupervised" practice allowed
- Level 5: supervision task may be given

### Performance Data

- The basis of making a judgment
  - Outcome of care
  - Process of care
  - Volume of activity information
- Methods of collecting data
  - Audit of medical records
  - Use of administrative databases
    - e.g. cost-effectiveness, medical errors
  - Log diaries
  - Direct observation

### Enhanced Requirements for Assessment in a Competency-Based, Time-Variable Medical Education System

- Time variability demands on the assessment data that are so necessary for making decisions about learner progress, either formative (e.g. feedback for improvement) or summative (e.g., decisions about advancing a student)
- It requires data from multiple sources that are gathered more frequently and on variable schedules, a greater level of data sharing, management, and communication than the traditional assessment systems

Larry D. Gruppen, Academic Medicine, Vol. 93, No. 3 / March 2018 Supplement

### Assessment in a Competency-Based, Time-Variable Medical Education System (cont.)

- Context specificity requires that multiple assessments be done by multiple observers over multiple cases in a variety of contexts to obtain a meaningful and trustworthy estimate of performance
- E-portfolios and mobile technology could help capture natural encounters in the workplace to provide feedback, formative assessment, and summative decisions regarding progress in various contexts

### Assessment in a Competency-Based, Time-Variable Medical Education System (cont.)

- a series of formative assessments, each of which separately serves to stimulate learning (an approach labeled *assessment for learning*), together may serve to make summative decisions (or what is called *assessment of learning*)
- Similarly, entrustment decisions have been identified as part of ad hoc assessments about learners for training purposes in health care tasks and as part of summative decisions about certification for health care tasks

### Assessment in a Competency-Based, Time-Variable Medical Education System (cont.)

- Summative judgments require not only solid assessment data but also a standard or criterion for performance that defines the decision (competent vs. not competent)
- EPAs have emerged as a key aspect of many CBME systems
- EPAs have been defined as units of professional practice to be entrusted to learners for unsupervised execution once they have demonstrated adequate performance

### Systems for Making Assessment Judgments

- Clinical competency committees (CCCs or entrustment committees) provide an example of innovation in how faculty collaborate around making assessment decisions, they collect review, and synthesize assessment data from various sources and times, organized around defined competencies, balancing the risks and benefits of the decisions about trust and progression through the program

### Be reminded!

- As the students of today are more inclined to question and challenge grades, there is a growing emphasis on ensuring that assessment methods are reliable, valid and able to sustain legal scrutiny
- The positive role of feedback and how it supports and facilitates learning the formative assessment requires a substantial amount of time and a commitment from staff to provide timely and detailed feedback. However, it is this valuable feedback that makes the learning encounter engaging and worthwhile
- The balance between formative and summative assessment

### References

- Assessment in medical education. Epstein R. NEJM 2007; 356:387-96
- Assessment in Medical Education: What Are We Trying to Achieve? Dr Ferris & O' Flynn. International Journal of Higher Education 2015; 4(2):139-44
- Trends and Opportunities in Medical Education: Aligning to Societal Needs and Expectations. Maniate JM. Arch Med Health Sci 2017; 5:154-6
- Entrustment Decisions: Bringing the Patient into the Assessment Equation. ten Cate O. Acad Med. 2017; 92:736-8
- Time-Variable Training in Medicine: Theoretical Considerations. ten Cate O et al. Acad Med. 2018;93:56-511

24 Apr 2018

หัวข้อ : New trends in written exam

## แนวทางใหม่ในการสอบข้อเขียน

(New Trends in Written Exam)

รศ. นพ.เชิดศักดิ์ ไอรอมณีรัตน์

คณะแพทยศาสตร์ศิริราชพยาบาล มหาวิทยาลัยมหิดล

ข้อสอบข้อเขียนเป็นรูปแบบการสอบที่มีใช้กันอย่างแพร่หลายในโรงเรียนวิทยาศาสตร์สุขภาพ เนื่องจากเป็นรูปแบบการวัดผลสัมฤทธิ์ทางการศึกษาที่ทำได้ง่าย เชื่อมโยงกับวัตถุประสงค์การเรียนรู้ได้ชัดเจน ครอบคลุมเนื้อหาวิชาการที่ต้องการประเมินในผู้เรียนได้มากในเวลาที่จำกัด และมีความเป็นธรรมสูง โดยทั่วไปแล้วการสอบข้อเขียนสามารถแบ่งออกเป็นสองรูปแบบหลักคือ (1) ข้อสอบเลือกคำตอบ (selected response items) และ (2) ข้อสอบที่ผู้สอบสร้างคำตอบเอง (constructed response items) โดยมีคุณสมบัติแตกต่างกัน ดังสรุปในตารางที่ 1

ตารางที่ 1 คุณสมบัติของข้อสอบข้อเขียน<sup>1</sup>

	ข้อสอบเลือกคำตอบ (Selected Response items)	ข้อสอบผู้สอบสร้างคำตอบเอง (Constructed response items)
ความรู้ที่ทำการประเมิน	ความจำ การแปลผลขั้นพื้นฐาน การประยุกต์ใช้ความรู้	การคิดวิเคราะห์ที่ซับซ้อน การแก้ปัญหา การแปลผล การตัดสินใจ
การสร้างข้อสอบ	ง่าย	ยาก ซับซ้อน
ค่าใช้จ่ายในการตรวจ	ต่ำ	สูง
ลักษณะการให้คะแนน	กติกากการให้คะแนนชัดเจน ปราศจากอคติ	ต้องใช้ดุลยพินิจของผู้ตรวจ
ผลกระทบของผู้ตรวจ	ไม่มีผล ไม่ว่าใครตรวจก็ได้คะแนนเท่ากัน	มีผลมาก คำตอบเดียวกันตรวจโดยผู้ตรวจคนละคนอาจได้คะแนนไม่เท่ากัน
ความเที่ยงของคะแนน	สูง	ต่ำ

ถึงแม้ข้อสอบข้อเขียนทั้งสองรูปแบบจะเป็นที่ใช้กันอย่างแพร่หลายในวงการศึกษานในประเทศไทย แต่ข้อสอบสองรูปแบบนี้ก็ยังมีข้อจำกัดในการใช้งาน โดยเฉพาะอย่างยิ่งเมื่อต้องการประเมินทักษะการคิดวิเคราะห์ที่ซับซ้อน จึงได้มีความพยายามพัฒนาข้อสอบข้อเขียนในรูปแบบใหม่ๆ ขึ้นมา ในบทความนี้ผู้นิพนธ์จะได้นำเสนอ

ข้อสอบข้อเขียนในรูปแบบเหล่านี้ เพื่อให้อาจารย์ผู้เกี่ยวข้องกับการวัดและประเมินผลได้มีทางเลือกในการสร้างข้อสอบที่หลากหลายมากขึ้น

### 1. ข้อสอบจับคู่ (Extended matching items)

จุดอ่อนของข้อสอบปรนัยที่มีผู้วิพากษ์กันมากคือการที่ผู้สอบที่ไม่มีความรู้สามารถเดาสุ่มแล้วมีโอกาสตอบถูกมากพอสมควร (หากมีห้าตัวเลือก โอกาสเดาถูกก็จะเป็น 0.2) การสร้างข้อสอบจับคู่ก็มีวัตถุประสงค์เพื่อให้มีตัวเลือกมาก ลดความน่าจะเป็นที่ผู้สอบที่เดาสุ่มสี่สุ่มห้าแล้วได้คำตอบที่ถูกต้อง<sup>2</sup> เมื่อมีตัวเลือกเยอะแล้ว การอ่านโจทย์หนึ่งข้อแล้วมีตัวเลือก 15 – 20 ตัว แล้วก็ไปข้อต่อไป ก็จะทำให้ผู้เข้าสอบต้องเสียเวลาค่อนข้างมาก จึงได้มีการสร้างโจทย์ที่มีหลายข้อ แล้วในกลุ่มโจทย์หลายข้อนี้ใช้ตัวเลือกกลุ่มเดียวกัน จึงเกิดเป็นข้อสอบจับคู่ขึ้น ข้อสอบในลักษณะนี้ได้รับความนิยมมากขึ้นเรื่อยๆ ในปัจจุบัน เนื่องด้วยมีการจัดสอบด้วยระบบคอมพิวเตอร์มากขึ้น การมีตัวเลือกหลายตัวเลือกบนหน้าจอคอมพิวเตอร์นั้นทำได้โดยสะดวก ไม่ต้องจัดทำกระดาษคำตอบที่มีตัวเลือก A,B,C,D,...,N ซึ่งมีค่าใช้จ่ายที่สูงมาก ตัวอย่างลักษณะการจับคู่ เช่น จับคู่ระหว่างชื่อโรค กับอาการและอาการแสดงของผู้ป่วย จับคู่ระหว่างเชื้อก่อโรค กับอาการและอาการแสดงของผู้ป่วย จับคู่ระหว่างชื่อยากับคุณสมบัติของยา เป็นต้น

### 2. ข้อสอบที่ผู้สอบมีปฏิสัมพันธ์กับคอมพิวเตอร์ในรูปแบบใหม่ๆ (Innovative computerized items)

ศักยภาพของคอมพิวเตอร์ที่ทำให้เกิดข้อสอบที่น่าสนใจเกิดจากปัจจัยที่สำคัญสองประการ คือ การแสดงผลที่มีคุณภาพสูง และ ความสามารถในการรับข้อมูลจากผู้ใช้ ในแง่ของการแสดงผลนั้นนอกเหนือไปจากแค่ตัวอักษร คอมพิวเตอร์ สามารถนำเสนอสื่อได้ทั้งภาพนิ่ง ภาพเคลื่อนไหว หรือเสียง ทำให้ผู้ออกข้อสอบสามารถใส่สื่อที่แสดงการประยุกต์ใช้ความรู้ได้เสมือนในชีวิตจริงได้มากขึ้น เช่น ภาพผู้ป่วย ภาพชิ้นเนื้อทางพยาธิวิทยา ภาพถ่ายรังสีวิทยา แผนภูมิลักษณะทางกายวิภาค เสียงการเต้นของหัวใจ วิดิทัศน์ผู้ป่วยที่มีการเคลื่อนไหวผิดปกติ วิดิทัศน์กระบวนการทำงานของทีมแพทย์ เป็นต้น

นอกจากนี้แล้วศักยภาพของคอมพิวเตอร์ที่สามารถรับข้อมูลจากผู้ใช้ได้มากกว่าแค่การเลือกตัวเลือกในกระดาษคำตอบ ทำให้ผู้จัดสอบสามารถสร้างสรรค์ข้อสอบในรูปแบบใหม่ๆ ที่สามารถวัดทักษะการคิดวิเคราะห์ และการแก้ปัญหาที่ใกล้เคียงกับชีวิตจริงได้มากขึ้น ตัวอย่างเช่น การใช้ mouse ชี้ (point) หรือจับวัตถุบนจอภาพเคลื่อนที่ไปวางที่อื่น (drag and drop) เป็นต้น ซึ่งมีการประยุกต์ใช้สร้างข้อสอบได้หลายรูปแบบ เช่น การเลื่อนลูกศรไปชี้ตำแหน่งที่มีพยาธิสภาพ การสลับคำ หรือข้อความเพื่อจัดลำดับขั้นตอนการปฏิบัติงาน การเลื่อนคำหรือวลีที่เป็นชื่อโรคไปจับคู่กับอาการของโรคนั้นๆ การเลื่อนชื่อเฉพาะไปวางตรงตำแหน่งที่ถูกต้องทางกายวิภาคของอวัยวะ เป็นต้น

### 3. ข้อสอบวัดความเข้ากันของข้อมูล (Script concordance items)

ข้อสอบชนิดนี้สามารถจัดเข้ากลุ่มการทดสอบแบบสร้างสถานการณ์แบบเขียน (written simulation)<sup>3</sup> เพื่อวัดความสามารถของผู้เข้าสอบในการคิดทางคลินิกอย่างมีเหตุผล (clinical reasoning) โดยมีพื้นฐานมาจากทฤษฎีต้นแบบ (script theory)<sup>4</sup> ซึ่งเชื่อว่าในการแก้ปัญหาทางคลินิกนั้นแพทย์จะดึงเอาความรู้ ความเข้าใจเกี่ยวกับตัวโรคต้นแบบ (script) ที่เก็บไว้ในความทรงจำออกมา แล้วค่อยๆ เทียบว่ามีรายละเอียดใดเหมือนหรือต่างไปจากผู้ป่วยที่เผชิญอยู่ตรงหน้า แล้วจึงตัดสินใจว่าผู้ป่วยน่าจะเป็นหรือไม่เป็นโรคที่สงสัย (hypothesis) โดยทั่วไปแล้วโจทย์ลักษณะนี้จะเริ่มจากการแสดงสถานการณ์ของผู้ป่วยที่มีความไม่แน่นอนในบางแง่มุม หลังจากนั้นผู้สอบจะได้รับข้อมูลเพิ่มเติม แล้วถูกถามว่าความน่าจะเป็นหรือความเหมาะสมในการวินิจฉัยโรคที่สงสัยดังกล่าวนั้น เพิ่มขึ้นหรือลดลงหลังจากที่ได้รับข้อมูลดังกล่าว<sup>5, 6</sup> ตัวอย่างดังแสดงในรูปที่ 1

โจทย์ : หญิงอายุ 50 ปี มีอาการปวดท้องใต้ชายโครงขวา 1 วัน

การวินิจฉัย	ข้อมูลที่ได้จากประวัติ ตรวจร่างกาย	ความน่าจะเป็นของการวินิจฉัยโรค
Hepatitis	Deep jaundice	-2 -1 0 +1 +2
Acute cholecystitis	Murphy sign positive	-2 -1 0 +1 +2
Acute pancreatitis	Periumbilical ecchymosis	-2 -1 0 +1 +2
Liver abscess	Stool occult blood +	-2 -1 0 +1 +2

Note: +2 almost certain

+1 somewhat more probable

0 neutral

-1 less probable

-2 almost rule out

รูปที่ 1 ตัวอย่างข้อสอบวัดความเข้ากันของข้อมูล (script concordance items) ในแง่การวินิจฉัย

ในการแก้ปัญหาโจทย์ลักษณะนี้ ผู้สอบจะต้องอ่านสถานการณ์ที่จัดให้ก่อน หลังจากนั้นผู้สอบจะได้รับชื่อโรคที่น่าจะเป็นการวินิจฉัย พร้อมกับให้ข้อมูลจากการซักประวัติหรือตรวจร่างกายเพิ่มเติม ผู้สอบต้องดึงเอาความเข้าใจเกี่ยวกับโรคที่ตั้งเป็นสมมติฐานออกมาจากความทรงจำ แล้วไตร่ตรองว่าข้อมูลเพิ่มเติมที่ให้มานั้น

ช่วยสนับสนุนสมมติฐานหรือไม่ หากช่วยสนับสนุนก็เลือกตัวเลขเป็นเลข +1 หรือ +2 หากคัดค้านการวินิจฉัยเลือก -1 หรือ -2 แต่หากไม่ส่งผลต่อการปรับเปลี่ยนสมมติฐาน ให้เลือก 0

นอกจากนี้ข้อสอบวัดการเข้ากันของข้อมูลยังสามารถประเมินการตัดสินใจส่งตรวจเพิ่มเติมทางห้องปฏิบัติการได้ด้วย ดังตัวอย่างในรูปที่ 2 ซึ่งเริ่มจากการได้รับสถานการณ์ แล้วได้ข้อมูลเพิ่มเติมจากประวัติหรือตรวจร่างกายแล้ว ผู้สอบต้องดึงเอาความเข้าใจว่าข้อมูลที่ได้มานั้นทำให้เกิดถึงโรคอะไร แล้วตัดสินใจว่าการส่งตรวจเพิ่มเติมที่โจทย์ให้มานั้นเหมาะสมหรือไม่

โจทย์: หญิง 35 ปี มีอาการปวดท้องน้อยข้างขวา 1 วัน

การส่งตรวจเพิ่มเติม	ข้อมูลที่ได้จากประวัติ ตรวจร่างกาย	ความเหมาะสมของการส่งตรวจเพิ่มเติม
Urine pregnancy test	ขาดประจำเดือนมา 2 เดือน	-2 -1 0 +1 +2
Urine analysis	อาการปวดร้าวไปหน้าขาขวาข้างใน	-2 -1 0 +1 +2
Chest x-ray PA upright	อาการปวดท้องเริ่มจากสะดือ แล้วย้ายไปปวดที่ท้องน้อย	-2 -1 0 +1 +2
CT abdomen	Tender with guarding at McBurney point	-2 -1 0 +1 +2

Note: +2 เหมาะสมมาก

+1 เหมาะสมพอควร

0 ตรวจหรือไม่ตรงก็ได้

-1 ไม่เหมาะสม

-2 ไม่เหมาะสมอย่างยิ่ง

รูปที่ 2 ตัวอย่างข้อสอบวัดความเข้ากันของข้อมูล (script concordance items) ในแง่การตรวจเพิ่มเติม

#### 4. ข้อสอบอัตนัยประยุกต์แบบแก้ปัญหาสำคัญ (Key feature problems)

หนึ่งในรูปแบบข้อสอบที่ใช้วัดการคิดวิเคราะห์และตัดสินใจทางคลินิกที่เป็นที่นิยมกันคือข้อสอบอัตนัยประยุกต์ (Modified Essay Question, MEQ) ซึ่งเป็นข้อสอบที่เริ่มจากการให้สถานการณ์ของผู้ป่วย แล้วมีโจทย์ถามให้ผู้สอบตอบคำถามเกี่ยวกับการแก้ปัญหาผู้ป่วยในสถานการณ์นั้น (เช่น สอบถามประวัติเพิ่ม ตรวจร่างกายเพิ่ม ส่งตรวจทางห้องปฏิบัติการเพิ่ม สั่งการรักษา) เมื่อผู้สอบตอบคำถามแล้วจะมีการเปิดเผยข้อมูล



ของผู้ป่วยเพิ่มขึ้นทีละน้อย และมีโจทย์ถามคำถามเพิ่มเติมเป็นลำดับ โดยผู้สอบไม่มีโอกาสย้อนกลับไปแก้ไขคำตอบของตนเองที่ได้ตอบไปในขั้นตอนก่อนหน้านี้<sup>7, 8</sup>

การใช้ข้อสอบอัตนัยประยุกต์มักประสบปัญหาคือ ข้อสอบส่วนใหญ่มุ่งเน้นวัดความครบถ้วนสมบูรณ์ของคำตอบมากกว่าการตัดสินใจแก้ปัญหา ทำให้ข้อสอบแต่ละข้อต้องใช้เวลาทำนาน และทำให้ผู้สอบได้แก้ปัญหาผู้ป่วยจำนวนน้อย มักส่งผลให้ได้ความเที่ยงของคะแนนสอบที่ต่ำ<sup>7, 9, 10</sup> ปัญหาสำคัญของการสอบด้วยสถานการณ์ผู้ป่วยจำนวนน้อยคือทักษะในการแก้ปัญหาทางคลินิกมีความจำเพาะต่อบริบทผู้ป่วยแต่ละราย (case specificity)<sup>11, 12</sup> การที่ผู้เข้าสอบแก้ปัญหาผู้ป่วยเจ็บหน้าอกได้ดี ไม่สามารถอนุมานได้ว่าผู้สอบคนดังกล่าวจะแก้ปัญหาผู้ป่วยปวดท้องได้ดีด้วย ดังนั้นจึงมีการพัฒนารูปแบบข้อสอบประเภทการแก้ปัญหาสำคัญ (Key Features Problem, KFP) ขึ้น

หลักการสำคัญของข้อสอบอัตนัยประยุกต์แบบแก้ปัญหาสำคัญนี้คือในการแก้ปัญหาผู้ป่วยแต่ละรายมีประเด็นปัญหาที่เป็นหัวใจสำคัญเพียงไม่กี่ประเด็นเท่านั้น ซึ่งประเด็นปัญหาเหล่านี้เรียกว่าปัญหาสำคัญ (Key features)<sup>13</sup> ซึ่งในผู้ป่วยแต่ละรายจะมีปัญหาสำคัญในประเด็นที่ต่างกันไปในข้อสอบอัตนัยประยุกต์แบบแก้ปัญหาสำคัญจะมุ่งเน้นตั้งโจทย์ถามเฉพาะประเด็นสำคัญเหล่านี้เท่านั้น ไม่จำเป็นต้องถามกระบวนการดูแลผู้ป่วยตั้งแต่ต้นจนจบในผู้ป่วยทุกราย จึงทำให้ข้อสอบลักษณะนี้ใช้เวลาสั้นในผู้ป่วยแต่ละราย นำไปสู่การประเมินได้หลายสถานการณ์มากขึ้น และมีความเที่ยงของคะแนนสอบที่สูง<sup>7, 9</sup>

##### 5. ข้อสอบอัตนัยที่มีปฏิสัมพันธ์แบบซับซ้อนกับคอมพิวเตอร์ (Complex computerized items)

ในการสอบรูปแบบต่างๆ ที่กล่าวมาข้างต้นแม้จะมีการนำคอมพิวเตอร์มาใช้บ้างแต่ก็จะเป็นการใช้ในรูปแบบที่ไม่ซับซ้อน การตอบสนองต่างๆ ของคอมพิวเตอร์ไม่ต้องอาศัยการประมวลผลมากนัก ผลลัพธ์คือสถานการณ์ผู้ป่วยที่ใช้สอบจะมีความไม่เหมือนจริงอยู่บ้าง ตัวอย่างเช่น ข้อสอบอัตนัยประยุกต์ซึ่งถูกออกแบบให้วัดการแก้ปัญหาผู้ป่วย ผู้เข้าสอบจะได้รับข้อสอบเป็นโจทย์ผู้ป่วยที่เหมือนกันทุกคน ไม่ว่าผู้เข้าสอบจะตอบคำถามอย่างไร โจทย์ข้อต่อไปก็ยังคงเดิมไม่เปลี่ยนแปลง เช่น โจทย์ถามว่าจะส่งตรวจทางห้องปฏิบัติการเพิ่มเติมอะไรบ้าง ผู้เข้าสอบตอบอะไรก็ตาม ในโจทย์ขั้นตอนต่อไปก็จะแสดงการตรวจทางห้องปฏิบัติการที่เหมาะสมกับผู้ป่วยรายนั้นเสมอ แต่ในชีวิตจริงเมื่อแพทย์สั่งการตรวจทางห้องปฏิบัติการที่ไม่เหมาะสม ผลการตรวจก็จะไม่ช่วยในการวินิจฉัย และแพทย์ก็ต้องคิดทบทวนใหม่ว่าจะส่งตรวจอื่นเพิ่มเติมหรือไม่ด้วย ซึ่งก็จะเสียเวลา และเสียโอกาสในการให้การรักษาที่รวดเร็วในผู้ป่วยบางราย

ในการประเมินทักษะการตัดสินใจแก้ปัญหาทางคลินิกนั้น หากจะทำให้เสมือนจริงคอมพิวเตอร์ควรจะต้องมีการประมวลผลที่มากขึ้นและตอบสนองต่อการตัดสินใจต่างๆ ของผู้เข้าสอบที่ต่างกัน ด้วยผลลัพธ์ที่ต่างกัน ซึ่งจะส่งผลให้แม้ผู้เข้าสอบจะเริ่มต้นด้วยสถานการณ์ผู้ป่วยเหมือนกัน แต่ด้วยกระบวนการคิด การ

ตัดสินใจ และการสั่งการรักษาที่ต่างกัน ผู้เข้าสอบแต่ละคนก็ต้องแก้ปัญหาในขั้นตอนต่อไปที่ต่างกัน และผลการรักษาที่ได้ก็จะต่างกันไปตามแนวทางของผู้เข้าสอบแต่ละคน การพัฒนาระบบคอมพิวเตอร์ที่สามารถตอบสนองต่อการตัดสินใจ การสั่งการรักษาที่หลากหลายของผู้เข้าสอบได้ในลักษณะนี้ต้องใช้ทรัพยากรสูง และใช้เวลาในการพัฒนานานมาก แต่ก็ได้มีการจัดสอบในลักษณะนี้บ้างในต่างประเทศ<sup>14</sup>

### เอกสารอ้างอิง

1. Haladyna TM. *Developing and validating multiple-choice test items*, 3rd ed. Mahwah, NJ: Lawrence Erlbaum Associates; 2004.
2. Case SM, Swanson DB. Extended-matching items: A practical alternative to free-response questions. *Teach Learn Med*. 1993/01/01 1993;5(2):107-115.
3. Van Der Vleuten CPM. The assessment of professional competence: Developments, research and practical implications. *Adv Health Sci Educ Theory Pract*. 1996;1(1):41-67.
4. Charlin B, Tardif J, Boshuizen HP. Scripts and medical diagnostic knowledge: theory and applications for clinical reasoning instruction and research. *Acad Med*. Feb 2000;75(2):182-190.
5. Charlin B, Roy L, Brailovsky C, Goulet F, van der Vleuten C. The Script Concordance test: a tool to assess the reflective clinician. *Teach Learn Med*. Fall 2000;12(4):189-195.
6. Humbert AJ, Johnson MT, Miech E, Friedberg F, Grackin JA, Seidman PA. Assessment of clinical reasoning: A Script Concordance test designed for pre-clinical medical students. *Med Teach*. 2011;33(6):472-477.
7. เชิดศักดิ์ ไชรมณีรัตน์. การสร้างข้อสอบอัตนัยประยุกต์. *เวชบัณฑิตศิริราช*. 2558;8(1):47 - 57.
8. Stratford P, Pierce-Fenn H. Modified essay question. *Phys Ther*. 1985;65(7):1075 - 1079.
9. 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 - 110.
10. Farmer EA, Page G. A practical guide to assessing clinical decision-making skills using the key features approach. *Med Educ*. 2005;39:1188 - 1194.
11. Eva KW. On the generality of specificity. *Med Educ*. 2003;37(7):587-588.

12. Perkins DN, Salomon G. Are cognitive skills context-bound? *Educ Researcher*. 1989;18:16-25.
13. 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.
14. USMLE Step 3: Content description and general information, Available from [http://www.usmle.org/pdfs/step-3/2017content\\_Step3.pdf](http://www.usmle.org/pdfs/step-3/2017content_Step3.pdf). Accessed April 2018.



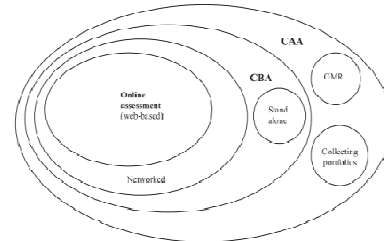
24 Apr 2018

หัวข้อ : IT in Assessment

## Information Technology in Assessment

เชิดศักดิ์ ไอร่มเจริญรัตน์  
ภาควิชาศึกษาศาสตร์ คณะแพทยศาสตร์ศิริราชพยาบาล  
มหาวิทยาลัย มหิดล

## Computer-aided assessment (CAA)



### Outline

- Test registration
- Test administration
- Test scoring and analysis

### Computerized Test Registration

- Advantages
  - Convenient
    - Examinees: No travel, No queue, Easy scheduling
    - Organizer: Decrease workload
  - Fairness
  - Completeness of information
  - Saving cost: paper, manpower, space
- Disadvantages
  - Data security
  - Computer errors/ crash

### Computerized Test Delivery

- Advantages
  - Multimedia presentation
  - Control of testing time
  - Reduce printing cost (paper, ink, personnel)
  - Eliminate printing error (missing pages, incorrect page order, unclear picture or text, printing at page border)
  - Improve test security
  - More data for testing research
  - Adaptive test: improve test efficiency

### Computerized Test Delivery

- Disadvantages
  - Students' adaptation to a new test format
    - Perception: Resistance to change
    - Computer literacy: skills in using computers
    - Eye strain from staring at monitor
  - Expensive start up cost
  - Require an appropriate system to maintain test item security
  - Potential for computer errors (system crash, power failure)

### Computers in the Assessment of Medical Competencies at Faculty of Medicine Siriraj Hospital

- MEQ: Modified Essay Question
- MCQ: Multiple-Choice Question
- OSCE: Objective Structured Clinical Examination

### Computerized MEQ

#### Computerized administration

- Automated advancement of the cases
- Multimedia presentation of cases, with student responses in writing on paper
- Fairness among examinees: Everyone sees the same image of the same size and resolution at the same time.
- Reduced paper use
- Prevent cheating

### Computer-based MCQ

- 1) CAT: Computerized adaptive test
- 2) CFT: Computerized fixed test
- 3) LOFT: Linear-on-the-fly test

### CAT

- Start a test with an item with moderate difficulty
- Examinees who answer the first item incorrectly will be given an easier item.
- Examinees who answer the first item correctly will be given a harder item.
- Goal: Efficient test

### Some Examples of CAT

- A. ABP: American Board of Pathology exam
- B. ASVAB: Armed Services Vocational Aptitude Test Battery
- C. GMAT: Graduate Management Admission Test
- D. GRE: Graduate Record Examination
- E. TOEFL: Test of English as Foreign Language

### GRE

- Verbal reasoning section x 2: CAT (30 min/section)
- Quantitative reasoning section x 2: CAT (35 min/section)
- Analytical writing section
  - Two essays: keyboard typing responses
    - Issue task 30 min
    - Argument task 30 min

## CFT

- Administration of a fixed-length, fixed-form computerized exam without any type of adaptive item selection
- Goal: cost saving, item innovations, computer scoring

## USMLE

- CBT was used in the USMLE in 1999
- Step 1
  - 8-hr computerized MCQ: 7 blocks of 40 items plus one hour break: some items include audio and/or video
- Step 2-CK
  - 9-hr computerized MCQ: 8 blocks of 39-40 items plus one hour break

## LOFT

- Construct a unique fixed-length test for each examinee
- A unique test form is constructed for each examinee according to content and difficulty criteria, regardless of examinee's ability level
- Goal: Item security, cost saving, item innovations, computer scoring

## Computers in an OSCE

- Timing and signaling system
- Standardized patient database

## Test Scoring and Analysis

- Computer-based test => import data from database into item analysis program
- Paper-based test => use Optical Mark Reader (OMR) to scan the answer sheet and transfer the Excel spreadsheet into item analysis program

## Test Statistics

โปรแกรมตรวจเฉลยข้อสอบ  
รูป 2.0

การสอบ : 010 121 (Basic Business)

วันที่ : 22 ธันวาคม 2555

จำนวนข้อสอบ : 100

จำนวนผู้เข้าสอบ : 744

Offical score	Raw score	Percentage of students answer item correctly	D-Index	Number of students answer item correctly	Item-to-total correlation coefficient (IRT)

### Test Statistics (cont.)

<b>SCORE STATISTICS</b>			
Mean = 68.152	S.D. = 11.915		
Mode = 65	(freq = 14)		
Max = 84	Min = 28		
<b>DIFFICULTY INDEX (p value)</b>			
Average p value = 0.566	Max p = 0.590	Min p = 0.010	
<b>DISCRIMINATION INDEX (D) (r value)</b>			
Average (D) = 0.344	Max D = 0.680	Min D = -0.180	
<b>RELIABILITY COEFFICIENT (RE) = 0.817</b> (Kuder-Richardson formula 20)			
<b>STANDARD ERROR OF MEASUREMENT (SEM) = 4.620</b> (S.D. x SQRT (RE))			

### Item Analysis and Option Analysis Faculty of Medicine Siriraj Hospital Mahidol University

<b>No. 1</b>		p Value : 0.34		r <sub>pbi</sub> : 0.23	
A	B	C	* D	E	
r <sub>pbi</sub> %	r <sub>pbi</sub> %	r <sub>pbi</sub> %	r <sub>pbi</sub> %	r <sub>pbi</sub> %	r <sub>pbi</sub> %
0.02	0.98	-0.16	0.95	-0.17	0.57
2.23	63.91	-0.67	13.26		
<b>No. 2</b>		p Value : 0.34		r <sub>pbi</sub> : 0.19	
A	B	C	D	* E	
r <sub>pbi</sub> %	r <sub>pbi</sub> %	r <sub>pbi</sub> %	r <sub>pbi</sub> %	r <sub>pbi</sub> %	r <sub>pbi</sub> %
0.31	4.76	-0.02	25.40	-0.19	10.79
-0.99	24.76	0.78	33.97		
<b>No. 3</b>		p Value : 0.58		r <sub>pbi</sub> : 0.25	
A	D	* C	E		
r <sub>pbi</sub> %	r <sub>pbi</sub> %	r <sub>pbi</sub> %	r <sub>pbi</sub> %	r <sub>pbi</sub> %	r <sub>pbi</sub> %
-0.03	8.89	-0.26	23.17	0.35	55.87
0.85	3.17	0.96	5.88		
<b>No. 4</b>		p Value : 0.50		r <sub>pbi</sub> : 0.33	
A	* B	C	D	E	
r <sub>pbi</sub> %	r <sub>pbi</sub> %	r <sub>pbi</sub> %	r <sub>pbi</sub> %	r <sub>pbi</sub> %	r <sub>pbi</sub> %
-0.15	1.30	0.33	50.48	0.15	4.13
0.18	90.88	0.13	33.02		
<b>No. 5</b>		p Value : 0.24		r <sub>pbi</sub> : 0.66	
A	B	C	* D	E	
r <sub>pbi</sub> %	r <sub>pbi</sub> %	r <sub>pbi</sub> %	r <sub>pbi</sub> %	r <sub>pbi</sub> %	r <sub>pbi</sub> %
-0.00	3.49	-0.30	53.02	0.95	-2.38
0.09	28.81	0.02	7.62		
<b>No. 6</b>		p Value : 0.53		r <sub>pbi</sub> : 0.20	
A	B	* C	D	E	
r <sub>pbi</sub> %	r <sub>pbi</sub> %	r <sub>pbi</sub> %	r <sub>pbi</sub> %	r <sub>pbi</sub> %	r <sub>pbi</sub> %
-0.10	23.17	-0.11	3.91	0.20	53.33
0.02	5.40	-0.02	14.20		

### Summary

- Test registration
- Test administration
- Test scoring and analysis



24 Apr 2018

## หัวข้อ : Advanced item analytic techniques

## Advanced Item Analytic Techniques

เชิดศักดิ์ ไอรณรัตน์

ภาควิชาคัลยศาสตร์ คณะแพทยศาสตร์ศิริราชพยาบาล

มหาวิทยาลัย มหิดล

## Item Analysis

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

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

## Basic Item Analysis

- MCQ item analysis using classical test theory
  - Item analysis
    - Item difficulty
    - Item discrimination
    - Distractor functionality
  - Test analysis
    - Reliability
    - Average difficulty, discrimination
    - Score distribution

## Outline

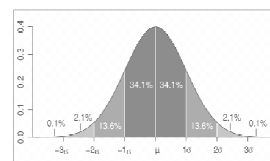
- Basic principles of Classical test theory and item response theory
- Classical test theory analysis
- Item response theory analysis

## Measurement Models

- |  |  |
|--|--|
| <ul style="list-style-type: none"> <li>• Classical Test Theory (CTT)</li> </ul> <p>Looking at a test item as an inseparable component of a test. The meaning of a score from one item can only be interpreted with the whole test.</p> | <ul style="list-style-type: none"> <li>• Item Response Theory (IRT)</li> </ul> <p>Looking at each test item individually. A test item has its own characteristic that can be interpreted separately from the test.</p> |
|--|--|

## Traditional Concept of Testing

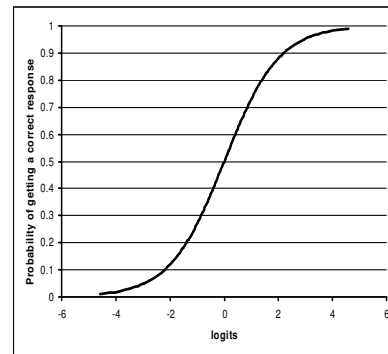
- Classical Test Theory (True score theory)
  - Each student has a true score, a hypothetical value representing a score free of error.
  - If we test a student repeatedly, the average of the obtained scores would approximate the true score, with a standard deviation of SEM.



### Psychometric Models

- Classical test theory (CTT)
  - $X = T + e$
- Item response theory (IRT)
  - One-parameter logistic model
  - Two-parameter logistic model
  - Three-parameter logistic model

### IRT Model



### Conceptual Differences

- IRT and CTT employ many different concepts.

### Concept 1

- Standard error of measurement
  - CTT: The same SEM applies to all scores.
  - IRT: SEM differs across scores.
    - Extreme values generally have larger SEMs than values around mean scores.

### Concept 2

- Test length and reliability
  - CTT: Longer tests are more reliable than shorter tests.
  - IRT: Shorter tests can be more reliable than longer tests.

### Concept 3

- Comparing scores across tests
  - CTT: Comparing test scores across multiple test forms is optimal when test forms are parallel (equality of test means, variances, and covariances with external variables).
  - IRT: Comparing test scores across multiple test forms is optimal when both tests have varying item difficulty levels.

### Concept 4

- Item statistics (difficulty, discrimination)
  - CTT: Accurate estimates of item statistics depend on having a representative sample of examinees.
  - IRT: Accurate estimates of item statistics can be obtained from an unrepresentative sample of examinees.

### Concept 5

- Meaning of test scores
  - CTT: Examinees' scores obtain their meaning by comparing them to a norm group.
  - IRT: Examinees' scores obtain their meaning by comparing them to task complexity or difficulty of items.

### Concept 6

- Scaling
  - CTT: Test scores are on an ordinal scale.
  - IRT: Test scores are on an interval scale.

### Concept 7

- Combining test scores
  - CTT: Combining raw test scores can lead to unbalanced total scores (i.e., some portion of the test can have more weight than others).
  - IRT: Combining test scores always produce balanced total scores.

### Classical Test Theory Analysis

- MEQ and OSCE
  - Item difficulty
  - Item discrimination
  - Reliability

### Item Difficulty

- Score of an item is on an ordinal scale instead of a dichotomous scale
  - Item difficulty
    - P-value => Percentage score
  - Item discrimination
    - Point biserial correlation => Pearson correlation

### Item Difficulty

- Score percentage
  - An easy item: High value
  - A difficult item: Low value
  - Which part is difficult?
    - Subscale analysis of each part

### Item Discrimination

- Pearson correlation  
Excel = Pearson (Range 1, Range 2)

### Internal Consistency Reliability

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

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

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

### Generalizability Theory (GT)

- Multi-faceted assessment
  - Assessment of MEQ scores
    - Sources of error
      - Students
      - Raters
      - Items
      - Interaction of these sources

### Generalizability

290 students x 10 raters x 40 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 student  $P$  given by rater  $R$  on item  $I$
- $\mu$  = Grand mean of the population
- $V_P, V_R, V_I$  = Effect of a student  $P$ , rater  $R$ , item  $I$
- $V_{PI}$  = Effect of student  $P$  crossed with item  $I$
- $V_{PR}$  = Effect of student  $P$  crossed with rater  $R$
- $V_{RI}$  = Effect of rater  $R$  crossed with item  $I$
- $V_{PRI}$  = Effect of student  $P$  crossed with rater  $R$  and crossed with item  $I$

### Generalizability

- The nature of decision: Absolute
- Absolute error variance:  $\sigma_\Delta^2$

$$\sigma_\Delta^2 = \sigma_R^2 + \sigma_I^2 + \sigma_{PI}^2 + \sigma_{PR}^2 + \sigma_{RI}^2 + \sigma_{PRI}^2$$

- Variance component study revealed:

$$\begin{matrix} \sigma_R^2 = 0.024 & \sigma_I^2 = 0.028 \\ \sigma_{PI}^2 = 0.000 & \sigma_{PR}^2 = 0.134 \\ \sigma_{RI}^2 = 0.021 & \sigma_{PRI}^2 = 0.091 \end{matrix}$$

### Factors Influencing MEQ Scores

- Examinees
- Items
- Scorers

### Multi-Faceted Assessment

#### Multi-Faceted Rasch Measurement Model

$$P_{nij}(X | \theta) = \frac{e^{\sum(B_n - C_j - D_i - E_{ik})}}{\sum e^{\sum(B_n - C_j - D_i - E_{ik})}}$$

- $P$  = Probability of person  $n$  being rated by rater  $j$  on item  $i$  with rating category  $k$
- $B_n$  = Ability level of person  $n$
- $C_j$  = Severity level of rater  $j$
- $D_i$  = Difficulty level of item  $i$
- $E_{ik}$  = Difficulty of rating  $k$  relative to  $(k-1)$  for item  $i$

### 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

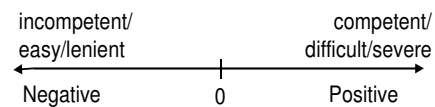
#### Restriction of range

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

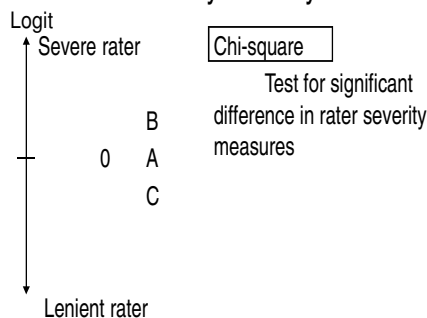
### Measurement Units

All facets (examinee competency, item difficulty, and rater severity) are measured on the same interval scale

Logits = log-odds of probability of getting high ratings over low ratings



### Leniency/Severity



### Fit Statistics

#### Infit mean-square statistics

- Information-weighted sum of squared standardized residuals
- Shows the size of measurement variability
- Expected values = 1.0
- Values less than 1.0 indicate too predictable observations
- Values greater than 1.0 indicate unpredictability

### Rating Consistency

Infit mean-square	Interpretation
>2.0	Too many unpredictable ratings, indicating significant rater inconsistency
1.5 - 2.0	Presence of some unpredictable ratings
0.5 - 1.5	Appropriate amount of variation in ratings, productive for measurement
< 0.5	Too little variation in ratings, suggesting restriction of range

### Multi-Faceted Rasch Model

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

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

$B_n$  = Clinical skills competence of a student  $n$

$C_j$  = Severity level of a rater  $j$

$D_i$  = Difficulty level of a skill  $i$

$F_m$  = Difficulty level of a case  $m$

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

24 Apr 2018

หัวข้อ : Setting a passing standard

## การตั้งเกณฑ์ผ่านการสอบ

นพ.เชตศักดิ์ ไอรณณรัตน์

ภาควิชาศัลยศาสตร์

คณะแพทยศาสตร์ศิริราชพยาบาล

## Twelve Steps for Test Development

1. Overall plan
2. Content definition
3. Test specification
4. Item development
5. Test assembly
6. Test production
7. Test administration
8. Scoring
9. Passing score
10. Reporting results
11. Item banking
12. Technical report

Downing SM, Haladyna TM. Handbook of test development. Mahwah, NJ: Lawrence Erlbaum Associates 2006.

## Standard

- A score that is set to be a boundary between those who perform well enough on the test (pass) from those who do not (fail).
- Standard = cutpoint

## Objectives

- เมื่อสิ้นสุดการบรรยายแล้ว ผู้เข้าอบรมสามารถ
  - บอกถึงความสำคัญของการตั้งเกณฑ์ผ่านได้ถูกต้อง
  - บอกถึงขั้นตอนของการตั้งเกณฑ์ผ่านได้ถูกต้อง
  - ยกตัวอย่างวิธีการตั้งเกณฑ์ผ่านได้อย่างน้อยสามวิธี
  - จัดทำเกณฑ์ผ่านการสอบ MCQ ด้วยวิธีการ modified Angoff method ในการสอบที่ตนเองเกี่ยวข้องได้อย่างเหมาะสม

## Outline

- Basic concepts
- Steps in setting standards
  - The type of standard
  - The method
  - Selecting judges
  - Standard setting meeting
  - Calculate the standards
  - Checking the standards

## Basic Concepts

- A standard is an answer to the question, "How much is enough?"
- The classification of examinees into two groups can result in two types of wrong decisions
  - False positive: Passing an examinee who should fail the exam
  - False negative: Failing an examinee who should pass the exam

## Judgment

1. Made by qualified judges
2. Meaningful to the persons who are making the decision
3. Made in a way that takes into account the purpose of the test

## 1. Types of Standards

- Absolute standard
- Relative standard

### Absolute Standard

- The standard is fixed, based on specific criteria of performance, but may undergo periodic re-evaluation of the standard
- Strengths
  - A standard is known in advance
  - A stable performance level is required to pass the examination => content-related standard
  - Provide clear feedback to examinees
  - Nobody has to fail the exam if their knowledge/skills is adequate for the purpose of the exam.
  - Promote a collaborative learning environment.

### Relative Standard

- The standard is set in reference to the group of examinees. The resulting standard may be reasonable providing a representative heterogeneous group.
- Strengths
  - The failure rate is stable, which in some way is easy for curriculum management

## 2. Methods for Setting Standards

1. Test-centered methods
2. Examinee-centered methods
3. Compromised methods

### Test-Centered Methods

- The judges set standards by reviewing the test items and provide judgments regarding the “just adequate” level of performance on these items.
  - Angoff's method
  - Nedelsky's method
  - Ebel's method



### Modified Angoff's Method

- The judgment
  - The probability that a borderline examinee would answer the test item correctly
- The passing score
  - The sum of all the probability of correct answers for all items on the exam

### Modified Angoff's Method (2)

Item	Probability
1	0.8
2	0.6
3	0.4
4	0.5
5	0.5
Passing score	2.8

### Nedelsky's Method

- The judgment
  - How many options a borderline examinee can eliminate from choosing in an item
- The passing score
  - The probability of correct answer for an item =  $1/(\text{the number of options not eliminated})$
  - The passing score of the test = the sum of all the probability of correct answers of all items on the test

### Nedelsky's Method (2)

Item	A	B	C	D	E	Not eliminated	Probability
1			X	X	X	2	$1/2 = 0.50$
2	X	X				3	$1/3 = 0.33$
3	X					4	$1/4 = 0.25$
4	X		X	X		2	$1/2 = 0.50$
5	X				X	3	$1/3 = 0.33$
Passing score							1.91

### Ebel's Method

- The judgment
  - What is the level of difficulty of an item?
    - Easy/Medium/difficult
  - What is the level of importance of that content in clinical practice?
    - Essential/Important/Acceptable/Questionable
  - The probability that a borderline examinee will answer an item in each category correctly
- The passing score
  - The sum of all the probability of correct answers for all items on the exam

### Ebel's Method (2)

	Easy	Medium	Difficult
Essential	0.95	0.85	0.80
Important	0.90	0.75	0.60
Acceptable	0.85	0.60	0.40
Questionable	0.55	0.45	0.35

### Ebel's Method (3)

Item	Difficulty	Importance	Probability
1	Easy	Essential	0.95
2	Easy	Importance	0.90
3	Difficult	Essential	0.80
4	Difficult	Acceptable	0.40
5	Medium	Acceptable	0.60
Passing score			3.65

### 2.Methods for Setting Standards

1. Test-centered methods
2. Examinee-centered methods
3. Compromised methods

#### Examinee-Centered Methods

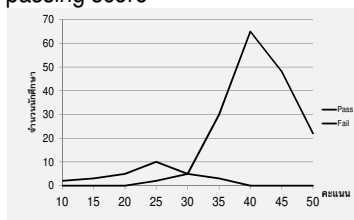
- The judges set a standard by reviewing the overall performance of examinees and determine who should pass and who should fail. The scores of examinees are reviewed and the passing score is set based on these judgments
  - Borderline-group method
  - Contrasting-groups method

#### Borderline-Group Method

- The judgment
  - Identify examinees who are “borderline”
- The passing score
  - The median score of this “borderline group”

#### Contrasting-Groups Method

- The judgment
  - Identify examinees who should “pass” and those who should “fail”
- The passing score



#### Compromised Method

- Combining relative and absolute standard setting methods
  - Hofstee method

### Hofstee Method

- The judgment
  - Minimum failure rate
  - Maximum failure rate
  - Minimum passing score
  - Maximum passing score
- The passing score
  - The intersection of test scores curve with diagonal line drawn from upper left to lower right corner

### 3. Selecting Judges

- The number of judges
- The qualification of judges

### 4. Standard Setting Meeting

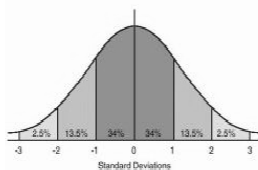
- Discussion of the purpose of the test, the characteristics of examinees, and the nature of competence.
- Explanation of the method and practice before the real standard setting procedure.

### 5. Calculating Standard

- Outliers
- Errors of the cutpoint

### Do we have to care about error?

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



### SEM

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

SD = standard deviation

r = internal consistency reliability

↑SD (more spread of score): higher SEM

↑r (more accurate measures): smaller SEM

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

## 6. Checking Standard

- Stakeholders' acceptance of the results
- Relationship with other markers of competence
- Prediction of future performance

## Summary

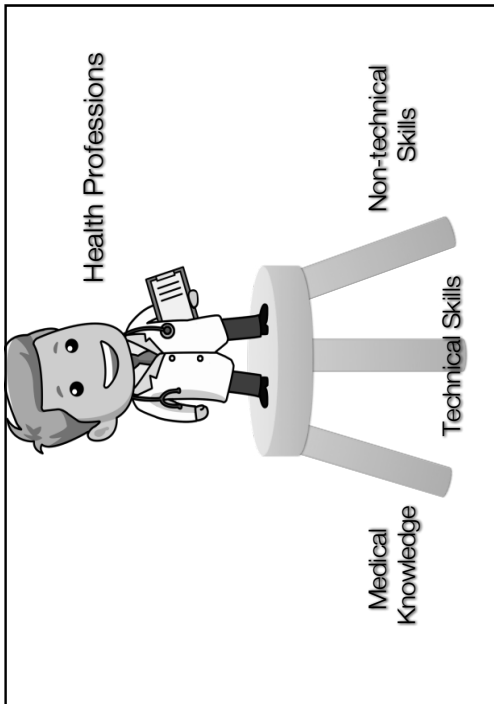
- Steps in setting up a standard
  1. Deciding on the type of standard
  2. Deciding on the method for setting standards
  3. Selecting judges
  4. Holding the standard setting meeting
  5. Calculating the standards
  6. Checking the standards after test

24 Apr 2018

หัวข้อ : Entrust able Professional Activities (EPA)

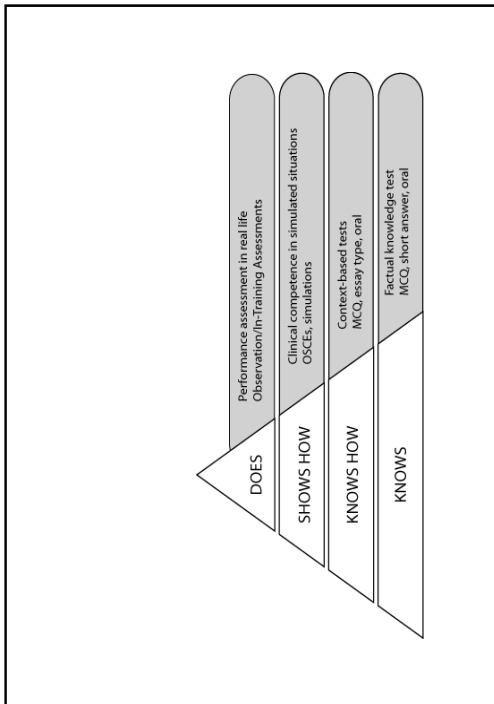
### Assessment during clinical clerkship

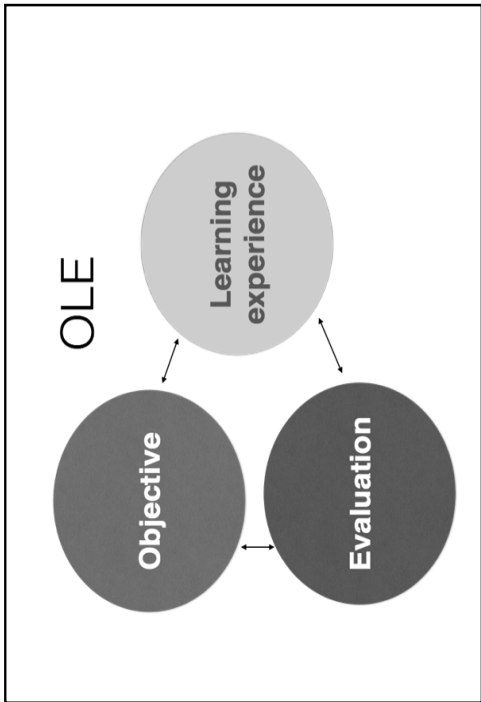
- Different settings from preclinical years
- Education & assessment need to
  - Include a broader range of Miller's pyramid
  - Support the development and assessment of integrated skills



### Entrustable Professional Activities (EPA)

Kasana Raksamani, MD, MHPE  
 Department of Anesthesiology  
 Faculty of Medicine Siriraj Hospital



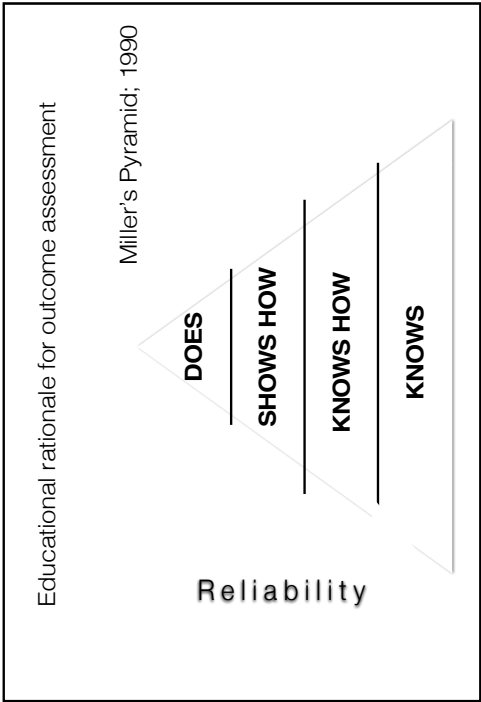


**Entrustable Professional Activities (EPA)**

- Shift from 'fixed time / flexible standards' to 'flexible standards / flexible time'
- A unit of professional practice (task) that can be entrusted to a sufficiently competent learner

**Problem of CBME**

- Suboptimal competency description
- Inadequate assessment instrument
- Data collection



## The matter of TRUST

- › Competence
- › Truthfulness / honesty
- › Conscientiously / reliability
- › Know their own limitation



## EPA

- › Do not aim to discriminate trainees
- › Value unique quality of trainees
- › Discard grade and scale
- › Primary focus: when to trust trainee (less supervision)
- › Use multiple source of information over time

## Entrustment scale = supervision scale

1. Not ready for entrustment
2. Ready for direct supervision
3. Ready for indirect supervision
4. Ready for unsupervised practice
5. Ready to supervise

**Table 2. Levels of proficiency for an entrustable professional activity<sup>25,26</sup>**

I	Resident has knowledge and some skill, but is not allowed to perform the EPA independently; mostly observes the EPA being performed by their supervisor
II	Resident may act under proactive, ongoing, full supervision. Supervisor present in the same room
III	Resident may act under reactive supervision, that is on request. Supervisor readily available
IV	Resident may act relatively independently from supervisor, under postponed or backstage supervision
V	Resident may act as a supervisor and instructor

The consecutive levels of proficiency and supervision (I to V) are described. Level IV indicates the proficiency level at which a task may be entrusted to a resident. EPA, entrustable professional activity.

**Table 3 Expected level of supervision by training stage**

Professional activities	Expected level of supervision				
	Training year 1	Training year 2	Training year 3	Training year 4	Posttraining subspecialization
EPA1	III	IV	V	V	V
EPA2	III	IV	IV	IV	V
EPA3	II	III	IV	V	V
EPA4	II	III	IV	IV	V
EPA5	I	II	III	III	V

- ▶ **Perform a pre-operative anesthesia assessment**
  - **Focused history and physical**
  - **Review of patient chart and interpretation of relevant investigations**
    - EPA1: Gather a history and perform a physical exam
    - EPA3: Recommend and interpret common diagnostic and screening tests
    - canMEDS 2015: Medical expert2
- ▶ **Create a differential diagnosis for common intraoperative complications including:**
  - Hypoxia
  - Hyper/hypotension
  - Hypercarbia
  - Tachycardia
  - EPA2: Prioritize a differential diagnosis following a clinical encounter.
  - canMEDS 2015: Medical expert 2.2, 5.1
- ▶ **Discuss and write post-operative recovery room orders including:**
  - **Orders for pain control and prevention of nausea and vomiting.**
  - **Orders for DVT prophylaxis following regional anesthesia**
  - EPA4: Enter and discuss orders and prescriptions
  - canMEDS 2015: Communicator 5.2
- ▶ **Utilize the electronic anesthetic record to document and retrieve patient information.**
  - EPA5: Document a clinical encounter in the patient record
  - canMEDS 2015: Communicator 5



## RIP OUT

# Nuts and Bolts of Entrustable Professional Activities

OLLE TEN CATE, PhD

## The Challenge

The entrustable professional activity (EPA) concept allows faculty to make competency-based decisions on the level of supervision required by trainees. Competency-based education targets standardized levels of proficiency to guarantee that all learners have a sufficient level of proficiency at the completion of training.<sup>1-6</sup> Collectively, the competencies (ACGME or CanMEDS) constitute a framework that describes the qualities of professionals. Such a framework provides generalized descriptions to guide learners, their supervisors, and institutions in teaching and assessment. However, these frameworks must translate to the world of medical practice. EPAs were conceived to facilitate this translation, addressing the concern that competency frameworks would otherwise be too theoretical to be useful for training and assessment in daily practice.

## What Is Known

Trust is a central concept for safe and effective health care. Patients must trust their physicians, and health care providers must trust each other in a highly interdependent health care system. In teaching settings, supervisors decide when and for what tasks they entrust trainees to assume clinical responsibilities. Building on this concept, EPAs are units of professional practice, defined as tasks or responsibilities to be entrusted to the unsupervised execution by a trainee once he or she has attained sufficient specific competence. EPAs are independently executable, observable, and measurable in their process and outcome, and therefore, suitable for entrustment decisions. Sequencing EPAs of increasing difficulty, risk, or sophistication can serve as a backbone for graduate medical education.<sup>6</sup>

## How Do EPAs Differ From Competencies?

- EPAs are not an alternative for competencies, but a means to translate competencies into clinical practice.
- Competencies are descriptors of physicians, EPAs are descriptors of work.
- EPAs usually require multiple competencies in an integrative, holistic nature. TABLE 1 shows how different EPAs require proficiency in several competency domains.

Olle ten Cate, PhD, is Professor of Medical Education and Director of the Center for Research & Development of Education at the University Medical Center Utrecht, the Netherlands.

Corresponding author: Th J (Olle) ten Cate, PhD, PO Box 85500, 3508 GA Utrecht, the Netherlands, t.j.tencate@umcutrecht.nl

DOI: <http://dx.doi.org/10.4300/JGME-D-12-00380.1>

## What Is Included in a Full EPA Description?

An EPA must be described at a sufficient level of detail to set trainee expectations and guide supervisor's assessment and entrustment decisions (see TABLE 2 for guidelines).

## How Do EPAs Relate to Milestones?

Milestones, as defined by the ACGME, are stages in the development of specific competencies. Milestones may link to a supervisor's EPA decisions (eg, direct proactive supervision versus distant supervision). The Pediatrics Milestone Project provides examples of how milestones can be linked to entrustment decisions.<sup>7,8</sup>

## What Do Entrustment Decisions Require?

Entrustment decisions involve clinical skills and abilities as well as more general facets of competence, such as understanding one's own limitations and knowing when to ask for help. Making entrustment decisions for unsupervised practice requires observed proficiency, usually on multiple occasions.

In practice, entrustment decisions are affected by 4 groups of variables: (1) attributes of the trainee (tired, confident, level of training); (2) attributes of the supervisors (eg, lenient or strict); (3) context (eg, time of the day, facilities available); and (4) the nature of the EPA (rare, complex versus common, easy). Entrustment decisions can be further distinguished as ad hoc (eg, happening during a night shift) or structural (establishing the recognition that a trainee may do this activity at a specific level of supervision from now on). In the clinical context, many ad hoc entrustment decisions happen every day. Structural entrustment decisions formally acknowledge that a trainee has passed a threshold that allows for decreased supervision. The certificate awarded at such occasions has been called a *statement of awarded responsibility* (STAR) and should be carefully documented.<sup>2</sup>

Linking an EPA with a competency framework emphasizes essential competency domains when observing a trainee executing the EPA.

## How You Can Start TODAY

Decide how many EPAs are useful for training.

While there can be many EPAs that serve to make ad hoc entrustment decisions, EPAs that lead to structural entrustment decisions (ie, certification or STARS) should involve broad-based responsibilities and be limited in number. For a graduate medical education program, no more than 20 to 30 EPAs are recommended.

RIP OUT

**TABLE 1** EXAMPLES OF EPAs RELATED TO THEIR MOST IMPORTANT ACGME COMPETENCY DOMAINS

Illustrative EPAs	ACGME Competencies					
	MK	PC	ISC	P	PBLI	SBP
Performing an appendectomy	•	•				
Executing a patient handover	•	•	•			•
Designing a therapy protocol	•				•	
Chairing a multidisciplinary meeting		•	•	•		•
Requesting organ donation			•	•		
Chronic disease management		•	•	•		•

Abbreviation: EPAs, entrustable professional activities; ACGME, Accreditation Council for Graduate Medical Education; MK, Medical Knowledge; PC, Patient Care; ISC, Interpersonal Skills and Communication; P, Professionalism; PBLI, Practice-Based Learning and Improvement and SBP, Systems-Based Practice.

**Use of EPAs in Assessing Trainees**

EPAs can be the focus of assessment. The key question is: Can we trust this trainee to execute this EPA? The answer may be translated to 5 levels of supervision for the EPA:

1. Observation but no execution, even with direct supervision
2. Execution with direct, proactive supervision
3. Execution with reactive supervision, ie, on request and quickly available

4. Supervision at a distance and/or post hoc
5. Supervision provided by the trainee to more junior colleagues

**What You Can Do LONG TERM**

- Review the specialty requirements and milestones, and work with your professional organization and local colleagues to identify EPAs.
- Detail the EPAs, following TABLE 2.
- Prepare faculty to provide EPA-based assessments.
- Use structural entrustment decisions as a “license” for trainees to execute EPAs with distant supervision.

**Resources**

- 1 ten Cate O. Entrustability of professional activities and competency-based training. *Med Educ.* 2005;39(12):1176–1177.
- 2 ten Cate O, Scheele F. Competency-based postgraduate training: can we bridge the gap between theory and clinical practice? *Acad Med.* 2007;82(6):542–547.
- 3 Mulder H, ten Cate O, Daalder R, Berkvens J. Building a competency-based workplace curriculum around entrustable professional activities: the case of physician assistant training. *Med Teach.* 2010;32(10):e453–e459.
- 4 ten Cate O, Young JQ. The patient handover as an entrustable professional activity: adding meaning in teaching and practice. *BMJ Qual Saf.* 2012. 2012 21: i9–i12. doi: 10.1136/bmjqs-2012-001213.
- 5 Chang A, Bowen JL, Buranosky RA, Frankel RM, Ghosh N, Rosenblum MJ, et al. Transforming primary care training-patient-centered medical home entrustable professional activities for internal medicine residents [published online ahead of print September 21, 2012]. *J Gen Int Med.* DOI: 10.1007/s11606-012-2193-3
- 6 Nasca TJ. The Next Accreditation System, June 2012. <http://www.acgme-nasca.org/assets/pdf/Nasca%20NAS%20June%202012%20Presentation%20Slide%20Show.pdf>. Accessed October 21, 2012.
- 7 Hicks PJ, Schumacher DJ, Benson BJ, Burke AE, Englander R, Guralnick S, et al. The pediatrics milestones: conceptual framework, guiding principles, and approach to development. *J Grad Med Educ.* 2010;2(3):410–418.
- 8 Pediatrics Milestone Project. [http://www.acgme.org/acgmeweb/Portals/0/PFAAssets/ProgramResources/320\\_PedsMilestonesProject.pdf](http://www.acgme.org/acgmeweb/Portals/0/PFAAssets/ProgramResources/320_PedsMilestonesProject.pdf). Accessed October 14, 2012.

**TABLE 2** GUIDELINES FOR FULL ENTRUSTABLE PROFESSIONAL ACTIVITIES DESCRIPTIONS

1. Title	Make it short; avoid words related to proficiency or skill. Ask yourself: Can a trainee be scheduled to do this? Can an entrustment decision for unsupervised practice for this EPA be made and documented?
2. Description	To enhance universal clarity, include everything necessary to specify the following: What is included? What limitations apply? Limit the description to the actual activity. Avoid justifications of why the EPA is important, or references to knowledge and skills.
3. Required Knowledge, Skills, and Attitudes (KSAs)	Which competency domains apply? Which subcompetencies apply? Include only the most relevant ones. These links may serve to build observation and assessment methods.
4. Required KSAs	Which KSAs are necessary to execute the EPA? Formulate this in a way to set expectations. Refer to resources that reflect necessary or helpful standards (books, a skills course, etc).
5. Information to assess progress	Consider observations, products, monitoring of knowledge and skill, multisource feedback.
6. When is unsupervised practice expected?	Estimate when full entrustment for unsupervised practice is expected, acknowledging the flexible nature of this. Expectations of entrustment moments can shape an individual workplace curriculum.
7. Basis for formal entrustment decisions	How many times must the EPA be executed proficiently for unsupervised practice? Who will judge this? What does formal entrustment look like (documented, publicly announced)?

## เอกสารประกอบการอบรม



25 Apr 2018



25 Apr 2018

หัวข้อ : Technical skills assessment

# Clinical review

## Objective assessment of technical skills in surgery

Krishna Moorthy, Yaron Munz, Sudip K Sarker, Ara Darzi

In the past few years, considerable developments have been made in the objective assessment of technical proficiency of surgeons. Technical skills should be assessed during training, and various methods have been developed for this purpose

Department of  
Surgical Oncology  
and Technology,  
Imperial College,  
St Mary's Hospital,  
London W2 1NY  
Krishna Moorthy  
*clinical research fellow*  
Yaron Munz  
*clinical research fellow*  
Sudip Sarker  
*clinical research fellow*  
Ara Darzi  
*professor of surgery*

Correspondence to:  
K Moorthy  
kmoorthy@  
imperial.ac.uk

BMJ 2003;327:1032-7

Surgical competence entails a combination of knowledge, technical skills, decision making, communication skills, and leadership skills. Of these, dexterity or technical proficiency is considered to be of paramount importance among surgical trainees. The assessment of technical skills during training has been considered to be a form of quality assurance for the future.<sup>1</sup> Typically surgical learning is based on an apprenticeship model. In this model the assessment of technical proficiency is the responsibility of the trainers. However, their assessment is largely subjective.<sup>2</sup> Objective assessment is essential because deficiencies in training and performance are difficult to correct without objective feedback.<sup>3</sup>

The introduction of the Calman system in the United Kingdom, the implementation of the European Working Time Directive, and the financial pressures to increase productivity<sup>4</sup> have reduced the opportunity to learn surgical skills in the operating theatre. Studies have shown that these changes have resulted in nearly halving the surgical case load that trainees are exposed to.<sup>5</sup> Surgical proficiency must therefore be acquired in less time, with the risk that some surgeons may not be sufficiently skilled at the completion of training.<sup>6</sup> This and increasing attention of the public and media on the performance of doctors have given rise to an interest in the development of robust methods of assessment of technical skills.<sup>7</sup> We review the research in this field in the past decade. Our objectives are to explore all the available methods, establish their validity and reliability, and examine the possibility of using these methods on the basis of the available evidence.

### Methods

We collected information for this review from our own experience, from discussions with other experts in this field, and from Medline searches by using the search terms "assessment," "technical skills," "psychomotor skills," "competence," "surgery," "simulations," "dexterity," and "virtual reality." We cross referenced some of the information from other articles and proceedings and abstracts of papers presented at conferences.

### Summary points

The assessment of technical skills is currently subjective and unreliable

Objective feedback of technical skills is crucial to the structured learning of surgical skills

Methods of assessment such as examinations, log books, and non-criteria based direct observation of procedures lack validity and reliability

Validated methods such as checklists, global rating scales, and dexterity analysis systems are suitable for the objective formative feedback of technical skills during training

Virtual reality systems have the potential to be used for assessment in the future

Further research is needed before these methods can be used for summative assessment and revalidation of surgeons

The surgical community could follow the example of other high reliability organisations such as aviation and design training programmes, where continuous assessment is a part of training

### Current methods of assessment in surgery and their limitations

Any assessment method should be feasible, valid, and reliable (box 1).<sup>2</sup> Currently five methods are available for assessing technical skills that are valid and reliable to varying degrees (table 1).<sup>2</sup> It is evident that some of the methods of assessment currently in use have poor validity and reliability. Examinations such as the fellowship and membership of the Royal College of Surgeons (FRCS and MRCS), conducted jointly by the royal colleges, focus mainly on the knowledge and clinical abilities of the trainee and do not assess a trainee's technical ability. One study showed that no relation existed between the American Board of Surgery in Training Exam (ABSITE) score and technical skill.<sup>8</sup> All



**Table 1** Assessment of technical skills—validity and reliability

Method of assessment	Reliability	Validity
Procedure lists with logs	Not applicable	Poor
Direct observation	Poor	Modest
Direct observation with criteria	High	High
Animal models with criteria	High	Proportional to realism
Videotapes	High	Proportional to realism

trainees in the United Kingdom are required to maintain a log of the procedures performed by them, which is submitted at the time of examinations, at job interviews, and during annual assessments. However, it has been found that log books are indicative merely of procedural performance and not a reflection of operative ability and therefore lack content validity.<sup>1, 2</sup> Time taken for a procedure does not assess the quality of performance and is an unreliable measure when used during real procedures, owing to the influence of various other factors.

The assessment of technical skills by observation, as currently occurs in the operating room, is subjective. As the assessment is global and not based on specific criteria it is unreliable. As it is influenced by the subjectivity of the observer it would possess poor test-retest reliability and also be affected by poor interobserver reliability as even experienced senior surgeons have a high degree of disagreement while rating the skills of a trainee.<sup>2</sup>

Morbidity and mortality data, often used as surrogate markers of operative performance, are influenced by patients' characteristics, and it is believed that they do not truly reflect surgical competence.<sup>10</sup>

## Objective methods of assessing technical skills

### Checklists and global scores

The availability of set criteria against which technical skills can be assessed makes the assessment process more objective, valid, and reliable (box 2). It has been said that checklists turn examiners into observers, rather than interpreters, of behaviour, thereby removing the subjectivity of the evaluation process.<sup>11</sup> The wide acceptance of the objective structured clinical examination (OSCE) led a group in Toronto to develop a similar concept for the assessment of technical skills.<sup>9</sup> The objective structured assessment of technical skills (OSATS) consists of six stations where residents and trainees perform procedures on live animal or bench models in fixed time periods.<sup>12</sup> Performance during the performance of tasks is assessed by using checklists specific to the operation or task (table 2) and a global rating scale (table 3). The global scale consists of seven generic components of operative skill that are marked on a 5 point Likert scale, with the middle and the extreme points anchored by explicit descriptors<sup>12</sup> to help in the criterion referenced assessment of performance. By using both formats of assessment Regehr et al have shown that checklists do not add any additional value to the assessment process and that their reliability is lower than that for the global rating scale.<sup>11</sup>

Both live animal operating and bench models have been used for the OSATS assessment. No differences became obvious in the performance of trainees in both

### Box 1: Principles of assessment

#### Validity

*Construct validity* is the extent to which a test measures the trait that it purports to measure. One inference of construct validity is the extent to which a test discriminates between various levels of expertise

*Content validity* is the extent to which the domain that is being measured is measured by the assessment tool—for example, while trying to assess technical skills we may actually be testing knowledge

*Concurrent validity* is the extent to which the results of the assessment tool correlate with the gold standard for that domain

*Face validity* is the extent to which the examination resembles real life situations

*Predictive validity* is the ability of the examination to predict future performance

#### Reliability

Reliability is a measure of a test to generate similar results when applied at two different points.<sup>2</sup> When assessments are performed by more than one observer another type of reliability test is applicable that is referred to as inter-rater reliability, which measures the extent of agreement between two or more observers<sup>9</sup>

formats of the examination.<sup>9</sup> The interstation reliability for the six stations was also found to be high.<sup>12</sup>

The only drawback to the performance of the OSATS assessment are the resources and time involved in getting several staff surgeons to observe the performance of trainees. Retrospective video watching of the performance may be a way forward. It does not entail the presence of multiple faculty raters and also adds to the element of objectivity by making the assessment blinded. By using this method of assessment Datta et al showed the construct validity of the global rating scale with an inter-rater reliability of 0.81.<sup>13</sup>

The use of both checklists and global rating scales entail the presence of multiple faculty raters or extensive video watching. Systems therefore need to be developed that can produce an assessment of technical skills in real time, with little need for several observers. Dexterity analysis systems have the potential to address this issue.

### Dexterity analysis systems

#### Imperial College surgical assessment device

This is a commercially available electromagnetic tracking system (Isotrak II, Polhemus, United States), which

### Box 2: Methods of assessment in surgery

#### Current

- Examinations
- Operative log books
- Time taken for a procedure
- Direct observation and assessment by trainers
- Morbidity and mortality data

#### Recent developments

- Checklists
- Global rating scales, such as OSATS (objective structured assessment of technical skills)
- Dexterity analysis systems, such as ICSAD (Imperial College surgical assessment device), ADEPT (advanced Dundee endoscopic psychomotor trainer)
- Virtual reality simulators
- Analysis of the final product on bench models
- Error scoring systems

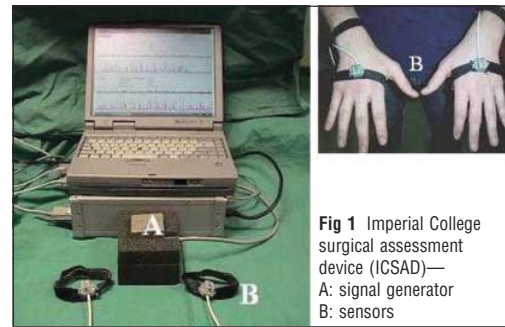
Clinical review

**Table 2** Checklists

Small bowel anastomosis:  
Interrupted end to end single layer anastomosis. Score one point for each correctly performed action

Tape No: \_\_\_\_\_ Item No: \_\_\_\_\_  
Assessor: \_\_\_\_\_ (initials only)

Procedural step	Correctly performed	Incorrectly performed
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		
11		
12		
13		
14		
15		
16		
17		
18		
19		
20		
21		
22		
23		



**Fig 1** Imperial College surgical assessment device (ICSAD)—  
A: signal generator  
B: sensors



**Fig 2** Typical trace from Imperial College surgical assessment device

consists of an electromagnetic field generator and two sensors that are attached to the dorsum of the surgeon's hands at standardised positions (fig 1). Bespoke software is used for converting the positional data generated by the sensors to dexterity measures such as the number and speed of hand movements, the distance travelled by the hands and the time taken for the task. Recently, our group has developed new software that allows us to stream video files along with the dexterity data in order to zoom into certain key steps of a procedure (fig 2).

Studies have shown the construct validity of the Imperial College surgical assessment device with respect to a range of laparoscopic and open surgical tasks.<sup>14</sup> The correlation between dexterity and previous

laparoscopic experience on a simple task in a box trainer<sup>15</sup> and on more complex tasks such as laparoscopic cholecystectomy on a porcine model is strong.<sup>16</sup> Experienced and skilled laparoscopic surgeons are more economical in terms of the number of movements and more accurate in terms of target localisation and therefore use much shorter paths.

**Other motion analysis systems**

Motion tracking can be based on electromagnetic, mechanical, or optical systems. The advanced Dundee endoscopic psychomotor trainer (ADEPT) was originally designed as a tool for the selection of trainees for endoscopic surgery, based on the ability of psychomotor tests to predict innate ability to perform relevant tasks. Studies have shown the validity and reliability of the

**Table 3** Global rating scale

Variable	Rating				
	1	2	3	4	5
Respect for tissue	Often used unnecessary force on tissue or caused damage by inappropriate use of instruments		Careful handling of tissue but occasionally caused inadvertent damage		Consistently handled tissues appropriately, with minimal damage
Time and motion	Many unnecessary moves		Efficient time and motion, but some unnecessary moves		Economy of movement and maximum efficiency
Instrument handling	Repeatedly makes tentative or awkward moves with instruments		Competent use of instruments, although occasionally appeared stiff or awkward		Fluid moves with instruments and no awkwardness
Knowledge of instruments	Frequently asked for the wrong instrument or used an inappropriate instrument		Knew the names of most instruments and used appropriate instrument for the task		Obviously familiar with the instruments required and their names
Use of assistants	Consistently placed assistants poorly or failed to use assistants		Good use of assistants most of the time		Strategically used assistant to the best advantage at all times
Flow of operation and forward planning	Frequently stopped operating or needed to discuss next move		Demonstrated ability for forward planning with steady progression of operative procedure		Obviously planned course of operation with effortless flow from one move to the next
Knowledge of specific procedure	Deficient knowledge. Needed specific instruction at most operative steps		Knew all important aspects of the operation		Demonstrated familiarity with all aspects of the operation

trainer.<sup>17</sup> Optical motion tracking systems consist of infrared cameras surrounded by infrared light emitting diodes. The infrared light is reflected off sensors that are placed on the limb of a surgeon.<sup>18</sup> Software is used to extrapolate the positional data of the markers to data on movement analysis. The disadvantages of such optical systems are that they suffer from disturbances to the line of vision. If the camera and signal generators become obscured from the markers the resulting loss in link leads to lost data. Signal overlap also prevents the use of markers on both limbs, making these systems restrictive.

### Virtual reality

Virtual reality is defined as a collection of technologies that allow people to interact efficiently with three dimensional computerised databases in real time by using their natural senses and skills.<sup>19</sup> Surgical virtual reality systems allow interaction to occur through an interface, such as a laparoscopic frame with modified laparoscopic instruments (fig 3).

The minimally invasive surgical trainer-virtual reality (MIST-VR) system was one of the first virtual reality laparoscopic simulators developed as a task trainer (fig 3). The system was developed as the result of collaboration between surgeons and psychologists who performed a task analysis of laparoscopic cholecystectomy. This resulted in a toolkit of skills needed to perform the procedure successfully.<sup>19</sup> These were then replicated in the virtual domain by producing three dimensional images of shapes, which users can manipulate.

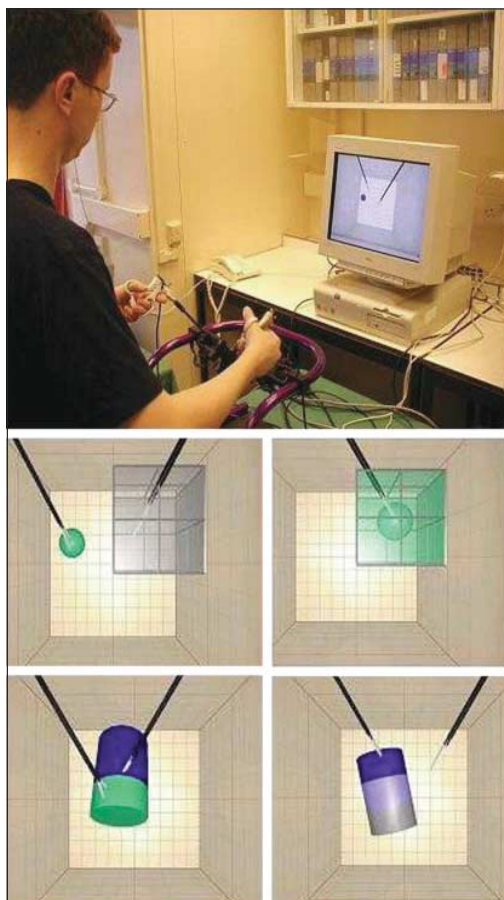


Fig 3 Minimally invasive surgical trainer-virtual reality (MIST-VR)

As virtual reality simulators are computer based systems they generate output data, or what is commonly referred to as metrics. In an international workshop a group of experts reviewed all methods of assessment and suggested parameters that should constitute output metrics for the assessment of technical skills.<sup>20</sup> Included were variables such as economy of movement, length of path, and instrument errors. The MIST-VR system has been validated extensively for the assessment of basic laparoscopic skills.<sup>21</sup> As these systems are currently low fidelity task trainers they are effective only in the assessment of basic skills. In the future, however, there is a possibility that higher fidelity virtual reality systems may be used as procedural trainers and for the assessment of procedural skill. One of the main advantages of virtual reality systems, in comparison to dexterity analysis systems, is that they provide real time feedback about skill based errors.

### Analysis of the final product

As outcomes after surgery are often difficult to ascribe solely to surgical technique, and as adverse outcomes from poor technique may not be apparent for many years—for example, recurrence after cancer resections—some researchers have suggested the idea of using outcome measures on bench models. Szalay et al assessed the quality of the final product after performance of six different bench model tasks.<sup>22</sup> The investigators found that the method possessed construct validity. They also found a correlation between OSATS and the final product assessment, which implies that analysis of the final product may overcome some of the problems involved with live ratings.

Datta et al assessed the leak rates and cross sectional area of the lumen after performance of a vascular anastomosis on a bench model and found a significant correlation between these outcome measures and surgical dexterity.<sup>23</sup> Hanna et al studied the quality of knots performed laparoscopically by using a tensiometer and derived a quality score for knots as an index of knot reliability.<sup>24</sup> The main advantage of these outcome measures is that they address the limitations of live and video based assessments and can be combined with dexterity data to derive proficiency scores, making assessment much more objective.

### Discussion

The previous section has highlighted the various methods that have been developed over the past decade for the objective assessment of technical skills in surgery. Educational bodies such as the Joint Committee for Higher Surgical Training (JCHST) in general surgery in the United Kingdom have appreciated the need to increase the emphasis laid on the assessment of technical skills during training. Hence the committee has recommended the use of an assessment of operative competence, with consultant surgeons assessing their trainees for procedures performed during a fixed training period. The assessment uses five overall global ratings (box 3) to rate the trainee's ability to carry out procedures.

However, a crucial issue that needs to be addressed regarding the assessment environment is whether assessments should be carried out on simulations with adequate face validity or performed during real proce-



Clinical review

**Box 3: Assessment of operative competence**

- U: Unknown (not assessed) during the training period
- A: Competent to perform the procedure unsupervised (can deal with complications)
- B: Does not usually require supervision but may need help occasionally
- C: Able to perform the procedure under supervision
- D: Unable to perform the entire procedure under supervision

dures. One drawback of the latter approach is that it is impossible to ensure standardisation because all patients are different, and trainers can be more preoccupied with patients' safety and timely completion of the procedure than with concentrating on the details of operative skill and technique. Further research is required regarding the feasibility and reliability of assessment during real procedures.

Animal laboratories are a common and popular method of learning skills in North America and Europe, but their use is banned in the United Kingdom. The use of synthetic models or simulations for the acquisition of basic technical skills is becoming increasingly acknowledged. Research groups such as ours have shown the validity of synthetic models for both training and assessment. Such simulations, while being crucial for learning, can also be used to assess skills simultaneously. In fact, assessment and training are synergistic. Without objective, valid, and reliable assessment training programmes cannot ensure the learning of skill, tackle deficiencies in training, and implement remedial measures. Figure 4 shows a recommended format for a cycle of training and assessment that could ensure that progression from one level to the higher one is based on robust criteria. This is based on the model of continuous training and assessment of pilots. Table 4 shows a suggested panel of tasks on bench models and virtual reality simulators according to the level of training. Such a paradigm shift in surgical learning will also ensure that trainees get objective feedback throughout their training programmes.

**Other applications of technical skills assessment**

In addition to the assessment of skills during training, measures described in this article can be used to show the efficacy of training courses in teaching the participants psychomotor skills. This would also give participants in the course an opportunity to gain an insight into the skills that they have learnt and allow training

centres to strive to improve the quality of teaching, ensure standardisation, and change course formats according to the performance of participants. The benefit of being able to measure surgical skill also presents the surgical community with the opportunity to show objectively the effect of training interventions, such as the use of virtual reality simulators, and to study the effect on technical performance of adverse environmental conditions, such as sleep deprivation, and distractions, such as noise in the operating theatre.

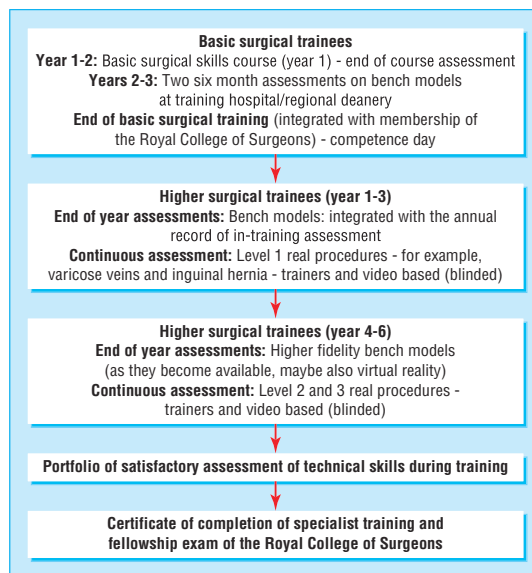


Fig 4 Recommended format for a cycle of training and assessment

The objective assessment of skills inevitably raises the issue of summative assessment and revalidation. Our group has previously alluded to the creation of a competence day for junior surgical trainees as an examination format by using OSATS and ICSAD (table 4). However, certain challenges (box 4) will have to be addressed before this becomes feasible for senior surgical trainees and consultant surgeons.<sup>25</sup>

**Conclusion**

We have attempted to highlight the considerable progress made in the past decade in the objective assessment of technical skills. The surgical community can choose from a wide range of methods of assessment initially during training to make feedback

Table 4 Suggested panel of tasks

Level of training	Tasks and methods of assessment
Basic surgical trainees (junior residents)	Knot tying (ICSAD)
	Suturing (simple, mattress, and precision)—ICSAD
	Excision of skin lesions—for example, sebaceous cyst (OSATS)
	Small bowel enterotomy (OSATS)
	Basic laparoscopic skills (virtual reality simulators)
Higher surgical trainees, years 1-3 (middle level residents)	Small bowel anastomosis (ICSAD, OSATS)
	Vein patch insertion (OSATS)
	Saphenous vein dissection and ligation (OSATS)
	Basic laparoscopic skills (virtual reality simulators)
	Basic endoscopy skills (virtual reality simulators)

ICSAD=Imperial College surgical assessment tool.  
OSATS=objective structured assessment of technical skills.

**Box 4: Research currently in progress**

- Feasibility, validity, and reliability of objective assessment during real procedures using ICSAD and video based assessment
- Studies to explore a link between technical skills and outcome for patients
- Studies addressing the predictive validity of objective assessment methods
- Correlation between performance on simulations and real procedures
- Longitudinal studies using a large cohort of trainees followed over a period of time to evaluate the link between training and assessment
- The establishment of databases for different tasks and procedures to establish performance criteria in order to make progress during training criteria based
- Use objective measures of surgical skills to demonstrate transfer of skills from virtual reality to real procedures
- Assessment of surgical competence in realistic environments such as a simulated operating theatre for assessment of communication, decision making, and leadership, and for training and assessment of crisis management

more objective, to base progression on criteria, and to help poorly performing trainees take remedial action. There are, however, still some issues that need to be addressed (box 4).

In addition to being of crucial importance for training, technical skills assessment is also driven by the need for the surgical community to ensure surgical care of the highest quality and reduce any potential errors resulting from poor technical performance. Owing to our inability to measure surgical skills objectively, so far it has been difficult to show a link between technical performance and outcome for patients. Future research should try to explore the link between technical skills assessed objectively and postoperative measures such as complication and recurrence rates and postoperative pain.

It must, however, be emphasised that technical skills are only a part of surgeon's competence, and the assessment of technical skills needs to be integrated with

cognitive and behavioural characteristics such as team skills and decision making in order to develop methods that assess surgical competence comprehensively.

Contributors: All authors participated in the preparation of the manuscript and in the revision process. AD is the guarantor.

Funding: BUPA Foundation.

Competing interests: None declared.

- 1 Cuschieri A, Francis N, Crosby J, Hanna GB. What do master surgeons think of surgical competence and revalidation? *Am J Surg* 2001;182:110-6.
  - 2 Reznick RK. Teaching and testing technical skills. *Am J Surg* 1993;165:358-61.
  - 3 Kopta JA. An approach to the evaluation of operative skills. *Surgery* 1971;70:297-303.
  - 4 Bridges M, Diamond DL. The financial impact of teaching surgical residents in the operating room. *Am J Surg* 1999;177:28-32.
  - 5 Ross DG, Harris CA, Jones DJ. A comparison of operative experience for basic surgical trainees in 1992 and 2000. *Br J Surg* 2002;89(suppl 1):60.
  - 6 Skidmore FD. Junior surgeons are becoming deskilled as result of Calman proposals. *BMJ* 1997;314:1281.
  - 7 Darzi A, Smith S, Taffinder N. Assessing operative skill. Needs to become more objective. *BMJ* 1999;318:887-8.
  - 8 Scott DJ, Valentine RJ, Bergen PC, Rege RV, Laycock R, Tesfay ST, et al. Evaluating surgical competency with the American board of surgery in-training examination, skill testing, and intraoperative assessment. *Surgery* 2000;128:613-22.
  - 9 Martin JA, Regehr G, Reznick R, MacRae H, Murnaghan J, Hutchison C, et al. Objective structured assessment of technical skill (OSATS) for surgical residents. *Br J Surg* 1997;84:273-8.
  - 10 Bridgewater B, Grayson AD, Jackson M, Brooks N, Grotte GJ, Keenan DJ, et al. Surgeon specific mortality in adult cardiac surgery: comparison between crude and risk stratified data. *BMJ* 2003;327:13-7.
  - 11 Regehr G, MacRae H, Reznick RK, Szalay D. Comparing the psychometric properties of checklists and global rating scales for assessing performance on an OSCE-format examination. *Acad Med* 1998;73:993-7.
  - 12 Reznick R, Regehr G, MacRae H, Martin J, McCulloch W. Testing technical skill via an innovative "bench station" examination. *Am J Surg* 1997;173:226-30.
  - 13 Datta V, Chang A, Mackay S, Darzi A. The relationship between motion analysis and surgical technical assessments. *Am J Surg* 2002;184:70-3.
  - 14 Datta V, Mackay S, Mandalia M, Darzi A. The use of electromagnetic motion tracking analysis to objectively measure open surgical skill in the laboratory-based model. *J Am Coll Surg* 2001;193:479-85.
  - 15 Taffinder N, Smith S, Mair J, Russell R, Darzi A. Can a computer measure surgical precision? Reliability, validity and feasibility of the ICSAD. *Surg Endosc* 1999;13(suppl 1):81.
  - 16 Smith SG, Torkington J, Brown TJ, Taffinder NJ, Darzi A. Motion analysis. *Surg Endosc* 2002;16:640-5.
  - 17 Francis NK, Hanna GB, Cuschieri A. The performance of master surgeons on the advanced Dundee endoscopic psychomotor tester: contrast validity study. *Arch Surg* 2002;137:841-4.
  - 18 Emam TA, Hanna GB, Kimber C, Cuschieri A. Differences between experts and trainees in the motion pattern of the dominant upper limb during intracorporeal endoscopic knotting. *Dig Surg* 2000;17:120-3.
  - 19 McCloy R, Stone R. Science, medicine, and the future. Virtual reality in surgery. *BMJ* 2001;323:912-5.
  - 20 Satava RM, Cuschieri A, Hamdorf J. Metrics for objective assessment. *Surg Endosc* 2003;17:220-6.
  - 21 Taffinder N, Sutton C, Fishwick RJ, McManus IC, Darzi A. Validation of virtual reality to teach and assess psychomotor skills in laparoscopic surgery: results from randomised controlled studies using the MIST-VR laparoscopic simulator. *Stud Health Technol Inform* 1998;50:124-30.
  - 22 Szalay D, MacRae H, Regehr G, Reznick R. Using operative outcome to assess technical skill. *Am J Surg* 2000;180:234-7.
  - 23 Datta V, Mandalia M, Mackay S, Chang A, Cheshire N, Darzi A. Relationship between skill and outcome in the laboratory-based model. *Surgery* 2002;131:318-23.
  - 24 Hanna GB, Frank TG, Cuschieri A. Objective assessment of endoscopic knot quality. *Am J Surg* 1997;174:410-3.
  - 25 Darzi A, Datta V, Mackay S. The challenge of objective assessment of technical skill. *Am J Surg* 2001;181:484-6.
- (Accepted 4 September 2003)

**Additional educational resources****Reviews**

Darzi A, Datta V, Mackay S. The challenge of objective assessment of surgical skill. *Am J Surg* 2001;181:484-6.

Grantcharov TP, Bardram L, Funch-Jensen P, Rosenberg J. Assessment of technical surgical skill. *Eur J Surg* 2002;168:139-44.

Moorthy K, Munz Y, Dosis A, Bello F, Darzi A. Motion analysis in the training and assessment of laparoscopic surgery. *Min Invas Ther Allied Technol* 2003;12: 37-42.

**Websites**

www.jchst.org—Joint Committee of Higher Surgical Training, for further details of the assessment of operative competence for higher surgical trainees in the United Kingdom

www.surgicaleducation.com—Association of Surgical Education, dedicated to promoting the art and science of education in surgery

**Endpiece****Hard people**

Writers, like teeth, are divided into incisors and grinders.


Walter Bagehot (1826-77), English economist, social scientist, and journalist in *The first Edinburgh reviewers*, 1879

Fred Charatan, retired geriatric physician, Florida

25 Apr 2018

หัวข้อ : Communication skills assessment

## Communication skills assessment



พ.ญ.กมลทิพย์ เลิศชัยสถาพร  
รศ.ดร.นพ.เชิดศักดิ์ ไอรรมณีรัตน์

### Goals

- After this session, participants will be able to
  - Explain key concepts and methods of communication skills assessment
  - Give examples of tools for communication skills assessment
  - Explain key concepts of preparation in communication skills assessment

### Outline

- Basic consideration in communication skills assessment
- Commonly used assessment tools
- Practice using the instruments
- Key concepts of preparation in communication skills assessment

### How to assess communication skills

- Choosing the appropriate assessment level
- Clinical context-based and integrated with other clinical skills
- Clinical relevant simulation & Clinical practice
  - Basic and applied knowledge - written assessment
  - Performance in quasi-real situations - OSCE
  - Performance in everyday practice - Work-based (e.g. mini-CEX)
  - Videotaped encounter-based

General principles to consider when designing a clinical communication assessment program  
Patient Education and Counseling 2017

### How to construct an assessment program

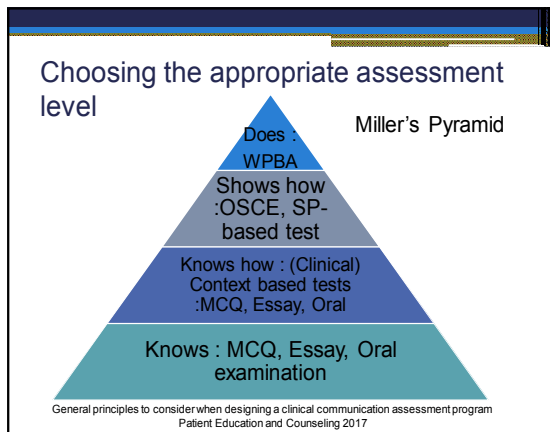
- Formative vs. Summative
- Multisource
- Longitudinal
- External feedback
- Self-assessment

General principles to consider when designing a clinical communication assessment program  
Patient Education and Counseling 2017

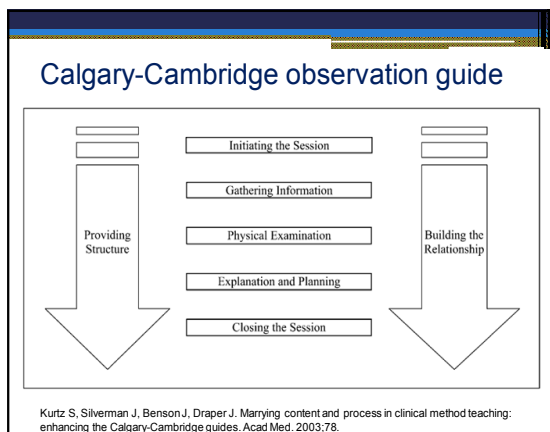
### How to choose an assessment tool

- According to learning goals and levels of competence (Miller's pyramid)
- Criteria of reliability, validity, educational impact, credibility, feasibility

General principles to consider when designing a clinical communication assessment program  
Patient Education and Counseling 2017



- ### Commonly used assessment tools
- Faculty
    - Global Consultation Rating Scale (GCRS)
    - Gap-Kalamazoo Communication Skills Assessment Form (GKCSAF)
  - Standardized or real patients
    - Revised UIC Communication and Interpersonal Skills (RUCIS)



- ### Global Consultation Rating Scale (GCRS)
- Rating by examiner, 12 domains
1. Initiating the session
  2. Problem Identification
  3. Problem Exploration
  4. Patient's Perspective
  5. Non-verbal Communication
  6. Developing Rapport
  7. Providing Structure
  8. Providing correct information
  9. Aiding accurate recall and understanding
  10. Incorporating the patient's perspective
  11. Planning and shared decision-making
  12. Closure
- Burt J et al. Assessing communication quality of consultations in primary care : initial reliability of the Global Consultation Rating Scale. BMJ Open 2014

- ### Gap-Kalamazoo Communication Skills Assessment Form (GKCSAF)
- Rating by examiner, 5-point Likert scale
- 9 components**
- Relationship building
  - Discussion opening
  - Information gathering
  - Understanding patient perspective
  - Information sharing
  - Agreement
  - Closure
  - Demonstrates Empathy
  - Communicates Accurate Information
- Peterson EB, et al. The reliability of a modified Kalamazoo Consensus Statement Checklist for assessing the communication skills of multidisciplinary clinicians in the simulated environment. Patient Educ Couns. 2014;96(3):411-418.

- ### RUCIS
- Rating by SP
1. Friendly communication
  2. Respectful treatment
  3. Listening
  4. Honest communication
  5. Interest in patient as a person
  6. Discussion of options
  7. Encourage questions
  8. Clear explanation
  9. Physical examination
  10. Appropriate vocabulary
  11. Sensitive subject matters
  12. Receptiveness to feedback
  13. Overall impression
- Iramanerat C, et al. Evaluating the effectiveness of rating instruments for a communication skills assessment of medical residents. Adv Health Sci Educ Theory Pract. 2009;14(4):575-594.

**กิจกรรม**

- ให้ผู้เข้าร่วมอบรม สังเกตการแจ้งข่าวร้ายและให้คะแนนการแจ้งข่าวร้าย
- ให้ผู้เข้าอบรมอภิปรายความคิดเห็นจากการทดลองใช้เครื่องมือในการให้คะแนนทักษะการสื่อสาร
- ให้ทำเนืถึง ข้อดี 1 อย่างและข้อควรพัฒนา 1 อย่างของเครื่องมือในการให้คะแนนทักษะการสื่อสารที่ท่านได้รับมอบหมาย

VDO




<https://drive.google.com/open?id=18CdkCUjeEk-wQf48sMLePrivuNFILQKI>

**กิจกรรม**

- ให้ผู้เข้าร่วมอบรม สังเกตการแจ้งข่าวร้ายและให้คะแนนการแจ้งข่าวร้าย
- ให้ผู้เข้าอบรมอภิปรายความคิดเห็นจากการทดลองใช้เครื่องมือในการให้คะแนนทักษะการสื่อสาร
- ให้ทำเนืถึง ข้อดี 1 อย่างและข้อควรพัฒนา 1 อย่างของเครื่องมือในการให้คะแนนทักษะการสื่อสารที่ท่านได้รับมอบหมาย


**“Brain storming”**



**Summary**

- Basic consideration in communication skills assessment
- Commonly used assessment tools
  - Global Consultation Rating Scale (GCRS)
  - Gap-Kalamazoo Form
  - RUCIS scale
- Practice using the instruments
- Key concepts of preparation in communication skills assessment

FACULTY OF MEDICINE SIRIRAJ HOSPITAL



**Thank You !**



# Appendix 1 Global Consultation Rating Scale

Calgary Cambridge- Global Consultation Scale (CC-GCS)	Good (2)	Adequate (1)	Not done/poor (0)	Not applicab
<b>Initiating the session</b>				
Greets patient				
Introduces self and nature of interview				
Demonstrates interest and respect, attends to patient's physical comfort				
Uses appropriate opening question				
<i>Overall Score for Initiating the Session</i>				
<b>Gathering Information</b>				
Listens attentively, minimising interruption and leaving space for patient				
Encourages patient to tell the story of the problem(s) from when first started to the present				
Checks and screens for further problems and negotiates agenda				
<i>Overall Score for Problem Identification</i>				
Uses open and closed questions, appropriately moving from open to closed				
Facilitates patient's responses verbally and non-verbally e.g. silence, repetition, paraphrasing				
Picks up and responds to verbal and non-verbal cues (body language, speech, facial expression)				
Clarifies statements which are vague or need amplification				
Periodically summarises & invites patient to correct interpretation or provide further information.				
Uses clear, easily understood language, avoids jargon				
<i>Overall Score for Problem Exploration</i>				
Actively determines patient's perspective (ideas, concerns, expectations, feelings, effects on life)				
Appropriately and sensitively responds to and further explores patient's perspective				
<i>Overall Score for Patient's Perspective</i>				
<b>Building the relationship</b>				
Demonstrates appropriate non-verbal behaviour e.g. eye contact, posture, position, movement, facial				
<i>Overall Score for Non-verbal Communication</i>				
Acknowledges patient's views and feelings; is not judgmental				
Uses empathy to communicate appreciation of the patient's feelings or predicament				
Provides support: expresses concern, understanding, willingness to help				
<i>Overall Score for Developing Rapport</i>				
<b>Providing Structure</b>				
Progresses from one section to another using signposting; includes rationale for next section				
Structures interview in logical sequence, attends to timing, keeps interview on task				
<i>Overall Score for Providing Structure</i>				
<b>Providing the correct amount/type of info for the individual patient</b>				
Chunks and checks, using patient's response to guide next steps				
Assesses the patient's starting point (good if carefully tailors explanation)				
Discovers what other information would help patient, seeks and addresses patient's info needs				
<i>Overall Score for providing correct amount and type of information</i>				
<b>Aiding accurate recall and understanding</b>				
Organises explanation (good if uses signposting/summarising)				
Checks patient's understanding (good if asks patient to restate information given)				
Uses clear language, avoids jargon and confusing language				
<i>Overall Score for aiding accurate recall and understanding</i>				
<b>Achieving a shared understanding: incorporating the patient's perspective</b>				
Relates explanations to patient's illness framework				
Encourages patient to contribute reactions, feelings and own ideas (good if responds well)				
Picks up and responds to patient's non-verbal and covert verbal cues				
<i>Overall Score for incorporating the patient's perspective</i>				
<b>Planning: shared decision making</b>				
Explores management options with patient				
Involves patient in decision making (good if establishes level of involvement patient wishes)				
Appropriately negotiates mutually acceptable action plan				
<i>Overall Score for planning and shared decision-making</i>				
<b>Closure</b>				
Contracts with patient re next steps				
Safety nets				
Summarises session briefly and clarifies plan of care				
Final check that patient agrees and is comfortable with plan				
<i>Overall Score for closure</i>				

**Gap-Kalamazoo Communication Skills Assessment Form\*** – Faculty/Peer Assessment

Date:	Your Name:	Your Title:
-------	------------	-------------

Title of Case:	Title of Conversation:
----------------	------------------------

**How well did the participant(s) do the following (please select one):**

	1 Poor	2 Fair	3 Good	4 Very Good	5 Excellent
<b>A: Builds a relationship (includes the following):</b> <ul style="list-style-type: none"> <li>• Greets and shows interest in the patient’s family</li> <li>• Uses words that show care and concern throughout the interview</li> <li>• Uses tone, pace, eye contact, and posture that show care and concern</li> <li>• Responds explicitly to patient and family statements about ideas and feelings</li> </ul>					
<b>B: Opens the discussion (includes the following):</b> <ul style="list-style-type: none"> <li>• Allows patient and family to complete opening statement without interruption</li> <li>• Asks “is there anything else?” to elicit full set of concerns</li> <li>• Explains and/or negotiates an agenda for the visit</li> </ul>					
<b>C: Gathers information (includes the following):</b> <ul style="list-style-type: none"> <li>• Addresses patient and family statements using open-ended questions</li> <li>• Clarifies details as necessary with more specific or “yes/no” questions</li> <li>• Summarizes and gives family opportunity to correct or add information</li> <li>• Transitions effectively to additional questions</li> </ul>					
<b>D: Understands the patient’s and families perspective (includes the following):</b> <ul style="list-style-type: none"> <li>• Asks about life events, circumstances, other people that might affect health</li> <li>• Elicits patient’s and family’s beliefs, concerns, and expectations about illness and treatment</li> </ul>					
<b>E: Shares information (includes the following):</b> <ul style="list-style-type: none"> <li>• Assesses patient’s/family’s understanding of problems and desire for more info</li> <li>• Explains using words that family can understand</li> <li>• Asks if family has any more questions</li> </ul>					
<b>F: Reaches agreement (includes the following):</b> <ul style="list-style-type: none"> <li>• Includes family in choices and decisions to the extent they desire</li> <li>• Checks for mutual understanding of diagnostic and/or treatment plans</li> <li>• Asks about acceptability of diagnostic and/or treatment plans</li> <li>• Identifies additional resources as appropriate</li> </ul>					
<b>G: Provides closure (includes the following):</b> <ul style="list-style-type: none"> <li>• Asks if patient and family have questions, concerns or other issues</li> <li>• Summarizes</li> <li>• Clarifies future time when progress will again be discussed</li> <li>• Provides appropriate contact information if interim questions arise</li> <li>• Acknowledges patient and family, and closes interview</li> </ul>					
<b>H. Demonstrates Empathy (includes the following):</b> <ul style="list-style-type: none"> <li>• Clinician’s demeanor is appropriate to the nature of the conversations</li> <li>• Shows compassion and concerns</li> <li>• Identifies/labels/validates patient’s and family’s emotional responses</li> <li>• Responds appropriately to patients and family’s emotional cues</li> </ul>					
<b>I: Communicates accurate information (includes the following):</b> <ul style="list-style-type: none"> <li>• Accurately conveys the relative seriousness of the patient’s condition</li> <li>• Takes other participating clinician’s input into account</li> <li>• Clearly conveys expected disease course</li> <li>• Clearly presents and explains options for future care</li> <li>• Gives enough clear information to empower decision making</li> </ul>					

\*Adapted from: Essential Elements: The Communication Checklist, © 2001 Kalamazoo Consensus Statement Group, and from: Rider EA. Interpersonal and Communication Skills. In: Rider EA, Nawotniak RH. *A Practical Guide to Teaching and Assessing the ACGME Core Competencies, 2nd edition.* Marblehead, MA: HCPro, Inc., 2010. © 2010 HCPro, Inc. Used with permission. Contacts: Elizabeth Rider, MSW, MD - elizabeth\_rider@hms.harvard.edu (member, Kalamazoo Consensus Statement Group) and Aaron Calhoun, MD - aaron.calhoun@louisville.edu (PERCS

**What did the participant(s) do best? (Please pick three choices)**

- 
- Builds a Relationship
  - Opens the Discussion
  - Gathers Information
  - Understands the Patient's and Family's Perspective
  - Shares Information
  - Reaches Agreement
  - Provides Closure
  - Demonstrates Empathy
  - Communicates Accurate Information
- 

**Why did you choose those particular answers?****In which domains could the participant(s) improve? (Please pick three choices)**

- 
- Builds a Relationship
  - Opens the Discussion
  - Gathers Information
  - Understands the Patient's and Family's Perspective
  - Shares Information
  - Reaches Agreement
  - Provides Closure
  - Demonstrates Empathy
  - Communicates Accurate Information
- 

**What could have been done better?**

**\*Adapted from:** Essential Elements: The Communication Checklist, © 2001 Kalamazoo Consensus Statement Group, and from: Rider EA. Interpersonal and Communication Skills. In: Rider EA, Nawotniak RH. *A Practical Guide to Teaching and Assessing the ACGME Core Competencies, 2nd edition*. Marblehead, MA: HCPro, Inc., 2010. © 2010 HCPro, Inc. Used with permission. Contacts: Elizabeth Rider, MSW, MD - elizabeth\_rider@hms.harvard.edu (member, Kalamazoo Consensus Statement Group) and Aaron Calhoun, MD - aaron.calhoun@louisville.edu (PERCS Program)



## Revised UIC Communication and Interpersonal Skills Scale

Please choose the option that best describes how you feel toward the resident's communication skills. Some items also have a 'not applicable' option. Select this option when the context of the case does not allow you to observe that aspect of the resident's performance.

1. Friendly communication	<input type="checkbox"/> You <u>did not greet me</u> , or <u>greeted me perfunctorily</u> , or communicated with me <u>rudely</u> during the encounter. <input type="checkbox"/> Your greeting and/or behavior during the encounter was generally <u>polite but impersonal or distant</u> . <input type="checkbox"/> You greeted me warmly and communicated with me in a <u>friendly, personal manner</u> throughout the encounter. <input type="checkbox"/> Your greeting and overall communication were friendly and compassionate. Your tone of voice was appropriate for the situation. Overall, you <u>created an exceptionally warm and friendly environment</u> that made me <u>feel comfortable</u> to tell you all of my problems.
2. Respectful treatment	<input type="checkbox"/> You showed an <u>obvious sign of disrespect</u> during the encounter. You <u>treated me as an inferior</u> . <input type="checkbox"/> You did not show disrespect to me. However, I observed some <u>signs of condescending behavior</u> . Although I believe it was <u>unintentional</u> , it made me feel that I was not at the same level with you. <input type="checkbox"/> You gave <u>several indications of respecting me</u> . If there was a physical exam, this includes draping me appropriately. <input type="checkbox"/> You were exceptionally respectful throughout the encounter. Your <u>verbal and nonverbal</u> communication showed <u>respect for my privacy, my opinions, my rights, and my socioeconomic status</u> .
3. Listening to my story	<input type="checkbox"/> You <u>rarely gave me any opportunity to tell my story</u> or <u>frequently interrupted me</u> while I was talking, not allowing me to finish what I said. Sometimes I felt you were not paying attention (for example, you asked for information that I already provided). <input type="checkbox"/> You let me tell my story without interruption, or only <u>interrupted appropriately</u> and respectfully. You seemed to pay attention to my story and <u>responded to what I said</u> appropriately. <input type="checkbox"/> You allowed me to tell my story without interruption, responded appropriately to what I said, and <u>asked thoughtful</u>

	<p><u>questions</u> to encourage me to tell more of my story.</p> <p><input type="checkbox"/> You were an exceptional listener. You encouraged me to tell my story and checked your understanding by <u>restating important points</u>.</p>
4. Honest communication	<p><input type="checkbox"/> You <u>did not seem truthful and frank</u>. I felt that there might be something that you were trying to hide from me.</p> <p><input type="checkbox"/> You <u>did not seem to hide any critical information</u> from me.</p> <p><input type="checkbox"/> You explained the facts of the situation <u>without trivializing negative information or possibilities</u> (e.g., side effects, complications, failure rates).</p> <p><input type="checkbox"/> You were exceptionally frank and honest. You <u>fully explained the positive and negative aspects</u> of my condition. You openly <u>acknowledged your own lack of knowledge or uncertainty</u>, and things you would have to consult with others. When appropriate, you also suggested I seek <u>a second opinion</u>.</p> <p><input type="checkbox"/> <b>Not applicable.</b> There was no information for the resident to provide.</p>
5. Interest in me as a person.	<p><input type="checkbox"/> You never showed interest in me as a person. You <u>only focused on the disease</u> or medical issue.</p> <p><input type="checkbox"/> In addition to talking about my medical issue, you spent some time <u>getting to know me as a person</u>.</p> <p><input type="checkbox"/> You spent some time exploring <u>how my medical issue affects my personal or social life</u>.</p> <p><input type="checkbox"/> You were exceptionally interested in me as a person. You not only explored how my medical problem affects my personal and social life, but also <u>showed your willingness to help me</u> address those challenges.</p>
6. Discussion of options/plans	<p><input type="checkbox"/> You <u>did not explain any options or plans</u>, you just told me what you would do without asking for my opinion.</p> <p><input type="checkbox"/> You explained options to me, but <u>did not involve me in decision making</u>. If you <u>solicited my opinion</u>, you just <u>ignored it</u>. You <u>made all the decisions for me</u> based on your medical opinion.</p> <p><input type="checkbox"/> You discussed options with me, made recommendations, <u>solicited my opinion</u> regarding the options/plans, and <u>incorporated my opinion into your medical planning</u>.</p>

	<input type="checkbox"/> You not only solicited my input, but also <u>explored the reasons for my choice and showed your understanding and respect for my decisions</u> by negotiating a mutually agreeable plan.
	<input type="checkbox"/> <b>Not applicable.</b> There were no decisions to be made in this case.
7. Encouraging my questions	<input type="checkbox"/> You <u>did not solicit questions</u> , or frequently <u>avoided my questions</u> , or did not provide helpful answers.
	<input type="checkbox"/> You sometimes asked if I had questions, but <u>seldom waited</u> at least 5 seconds to allow me to formulate questions. You <u>addressed my questions briefly</u> without avoiding them.
	<input type="checkbox"/> You <u>actively encouraged me to ask questions</u> , <u>paused to allow me to formulate them</u> , and provided <u>clear and sufficient answers</u> to all of my questions.
	<input type="checkbox"/> You actively encouraged me to ask questions several times during the encounter, with <u>sufficient wait time</u> . You spent significant time and effort to answer my questions clearly and <u>confirmed that I understood the answer</u> and that my concerns were addressed.
8. Providing clear explanations	<input type="checkbox"/> You <u>rarely explained things</u> to me; you <u>did not help me better understand my situation</u> .
	<input type="checkbox"/> You gave me only <u>brief explanations</u> of my situation; you did not help me understand what would happen next.
	<input type="checkbox"/> You gave me a <u>full and understandable explanation</u> of my situation, pertinent findings, and important next steps.
	<input type="checkbox"/> You gave me a full explanation of my situation, your thinking about it and your recommendation, and <u>probed my understanding</u> by letting me summarize pertinent information.
	<input type="checkbox"/> <b>Not applicable.</b> There was nothing to be explained in this case.
9. Physical examination	<input type="checkbox"/> You <u>never or rarely warned me about what you were going to do</u> with my body. You also never or <u>rarely explained what you found</u> from the physical examination.
	<input type="checkbox"/> You <u>did not warn me</u> about what you were going to do with my body, OR <u>did not explain to me pertinent findings</u> (both negative and positive) from your physical examination.
	<input type="checkbox"/> You <u>told me what you were going to do to my body</u> AND <u>described what you found</u> .

	<p><input type="checkbox"/> You helped me understand clearly what you were going to do to my body. You also provided <u>clear explanation of what you found</u> from the physical examination and <u>the implications of your findings</u> for my situation.</p>
<p>10. Appropriate vocabulary</p>	<p><input type="checkbox"/> <b>Not applicable.</b> There was no physical examination in this case.</p>
<p>11. Sensitive subject matters (e.g., sexual history, tobacco/alcohol/drug use, religious/cultural issues, giving bad news, or difficult emotional states)</p>	<p><input type="checkbox"/> You used vocabulary that was too simple or too complex for me, or <u>frequently used medical terms without explaining them</u> to me. Sometimes I could not understand what you told me without asking for explanations of terms you used.</p>
	<p><input type="checkbox"/> Your vocabulary was generally appropriate but you <u>sometimes inadvertently used medical terms without explaining them</u> to me.</p>
	<p><input type="checkbox"/> Your vocabulary was appropriate and if needed you provided <u>brief explanations of any medical terms you used</u> without need for prompting.</p>
	<p><input type="checkbox"/> Your vocabulary was appropriate and you <u>always provided clear and full explanation of relevant medical terms</u> you used. In addition, you helped me <u>better my understanding</u> of my condition with the medical terms you explained to me.</p>
	<p><input type="checkbox"/> You <u>never warned me</u> before approaching sensitive subject matters. You seemed judgmental and clearly <u>expressed your disapproval of my positions or feelings</u>, making me feel uncomfortable about discussing these subjects or feelings with you.</p>
	<p><input type="checkbox"/> You were careful and nonjudgmental in discussing sensitive subject matters. However, you <u>did not express understanding</u> of my feelings and <u>did not provide much emotional support</u>.</p>
	<p><input type="checkbox"/> You were sensitive about discussing difficult subjects and were respectful of my feelings. I never sensed that you were judgmental or disapproving of my positions or feelings on these subjects. You <u>showed empathic understanding</u> of my position or feelings and provided appropriate <u>emotional support</u>.</p>
	<p><input type="checkbox"/> You were unusually empathic, sensitive and respectful of me and of my feelings and provided exceptional emotional support. In addition, you <u>verbally reflected these back to me</u> (e.g., “You sound sad”) to show your understanding.</p>
	<p><input type="checkbox"/> <b>Not applicable.</b> There were no sensitive subject matters in this case.</p>

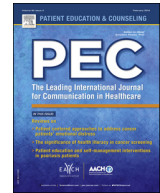
12. Receptiveness to feedback	<input type="checkbox"/> You <u>did not seem open to my feedback</u> about your performance. You <u>responded defensively</u> or dismissively to many of my comments.
	<input type="checkbox"/> You listened to my feedback agreeably but passively. You <u>did not actively participate</u> during the feedback session.
	<input type="checkbox"/> You were able to <u>describe some of your own effective and ineffective behaviors</u> , were attentive to my comments, and had an <u>open discussion with me about alternative behaviors</u> .
	<input type="checkbox"/> You <u>actively solicited additional feedback</u> and <u>showed signs of integrating my feedback</u> into your behavioral repertoire. For example, you tried to role-play the communication techniques I suggested.
	<input type="checkbox"/> <b>Not applicable.</b> I provided no feedback.
13. Do I want to see you again as my personal physician?	<input type="checkbox"/> I did not feel comfortable in communicating with you at all. <u>I would rather see a different physician.</u>
	<input type="checkbox"/> I think <u>you were okay in general and might come see you again.</u>
	<input type="checkbox"/> I was impressed by the way you communicated with me. <u>I would like to see you again.</u>
	<input type="checkbox"/> I was very impressed with you. I think you are <u>one of the best physicians I have ever seen.</u> I would feel very comfortable discussing any medical problems with you, and would recommend you to my friends.





Contents lists available at ScienceDirect

## Patient Education and Counseling

journal homepage: [www.elsevier.com/locate/pateducou](http://www.elsevier.com/locate/pateducou)

## Discussion

## General principles to consider when designing a clinical communication assessment program



Claudia Kiessling<sup>a</sup>, Zoi Tsimtsiou<sup>b,\*</sup>, Geurt Essers<sup>c</sup>, Marc van Nuland<sup>d</sup>, Tor Anvik<sup>e</sup>,  
 Maria M. Bujnowska-Fedak<sup>f</sup>, Richard Hovey<sup>g</sup>, Ragnar Joakimsen<sup>e</sup>, Noëlle Junod Perron<sup>h</sup>,  
 Marcy Rosenbaum<sup>i</sup>, Jonathan Silverman<sup>j</sup>

<sup>a</sup> Department Assessment, Brandenburg Medical School, Neuruppin, Germany<sup>b</sup> Department of Hygiene, School of Medicine, Aristotle University of Thessaloniki, University Campus, 54124, Thessaloniki, Greece<sup>c</sup> Department of Public Health and Primary Care, Leiden University Medical Centre, Leiden, The Netherlands<sup>d</sup> Department of Public Health and Primary Care, University of Leuven, Leuven, Belgium<sup>e</sup> UiT The Arctic University of Norway, Tromsø, Norway<sup>f</sup> Department of Family Medicine, Wroclaw Medical University, Wroclaw, Poland<sup>g</sup> Division of Oral Health & Society, Montreal, Canada<sup>h</sup> Unit of Development and Research in Medical Education, Geneva Faculty of Medicine, University of Geneva, Geneva, Switzerland<sup>i</sup> Department of Family Medicine and Office of Consultation and Research in Medical Education, University of Iowa Carver College of Medicine, Iowa City, USA<sup>j</sup> School of Clinical Medicine, University of Cambridge, Cambridge, UK

## ARTICLE INFO

## Article history:

Received 11 June 2016

Received in revised form 26 February 2017

Accepted 25 March 2017

## Keywords:

Assessment

Communication skills

Formative

Summative

Feedback

Clinical communication

## ABSTRACT

**Objectives:** Assessment of clinical communication helps teachers in healthcare education determine whether their learners have acquired sufficient skills to meet the demands of clinical practice. The aim of this paper is to give input to educators when planning how to incorporate assessment into clinical communication teaching by building on the authors' experience and current literature.

**Methods:** A summary of the relevant literature within healthcare education is discussed, focusing on what and where to assess, how to implement assessment and how to choose appropriate methodology.

**Results:** Establishing a coherent approach to teaching, training, and assessment, including assessing communication in the clinical context, is discussed. Key features of how to implement assessment are presented including: establishing a system with both formative and summative approaches, providing feedback that enhances learning and establishing a multi-source and longitudinal assessment program.

**Conclusions:** The implementation of a reliable, valid, credible, feasible assessment method with specific educational relevance is essential for clinical communication teaching.

**Practice implications:** All assessment methods have strengths and limitations. Since assessment drives learning, assessment should be aligned with the purpose of the teaching program. Combining the use of different assessment formats, multiple observations, and independent measurements in different settings is advised.

© 2017 Elsevier B.V. All rights reserved.

## 1. Introduction

Numerous studies have shown the importance of effective clinical communication in health care. Communication is now taught in most medical and allied health professional schools [1,2]. Communication assessment enables teachers to determine whether their students are fit for later professional life and have acquired sufficient skills to be able to meet the demands of clinical

reality (assessment of learning) [3]. It is also recognized that assessment is important for learners, as it drives their learning (assessment for learning) [4]. For learners, assessment helps to identify their learning needs. Moreover, assessment legitimizes the subject: if communication skills are not proportionally assessed, learners may assume that they are not as important as other topics. The content and format of assessment tools used will influence students' learning behavior, implicitly and explicitly. The way they are assessed will send out a strong message to learners about what teachers consider to be effective communication in clinical practice.

\* Corresponding author.

E-mail address: [zoitsimtsiou@yahoo.gr](mailto:zoitsimtsiou@yahoo.gr) (Z. Tsimtsiou).



However, many educators in the field of clinical communication struggle with implementing a feasible assessment program. Like teaching itself, assessing learners needs specific expertise, time and money, which is often not readily available among communication educators, especially in countries that are starting to implement communication training.

As members of the Teaching committee of EACH, the authors have experienced a strong need for advice among communication teachers when it comes to assessment. This discussion paper aims to give input to educators that are planning to incorporate assessment into clinical communication teaching. This paper is not a new systematic review regarding communication assessment instruments. However, it adds to the existing systematic reviews [5-7] by bringing together the authors' expertise and experiences with selected literature from the field of clinical communication teaching, medical education and assessment. Thus, we aim to identify and present important aspects that are useful to consider when starting or improving a communication assessment program. In the next paragraphs, we will discuss what to assess, how to choose the appropriate assessment level and how to construct an assessment program for clinical communication.

## 2. What and how to assess clinical communication?

Our core message would be: "assess what you teach and train". In our view, the content and form of assessment reflect the purpose and desired outcomes of the teaching program [8]. A successful clinical communication teaching program, therefore, is based on the application of sound theoretical principles and scientific evidence of effective communication [9]. According to Thomas et al. in their recommendations for curriculum development, educational objectives derive from these underlying theoretical principles and provide clear descriptions of the outcome expected from learners as a result of a course [10]. Educational strategies and teaching methods which support learner-centered environments have shown to be effective to encourage cumulative learning and self-reflection. "Learner-centered" implies that education is driven by learner needs. Adult learners prefer contextual learning (e.g. solving work-related problems in simulated scenarios) in small groups and building new content on prior knowledge [11,12]. Situated learning is one theory fostering these fundamental principles [13,14].

Assessment is based on the same theoretical principles and educational objectives [15]. The assessment methods mirror the instruction methods and are selected to measure students' achievements according to the educational objectives and their level of competence (Fig. 1). This process of planning an assessment program can be supported by a careful blueprinting process, in which the content of the assessment is congruent with the conceptual frameworks and educational objectives found in the curriculum [16]. For example, if the educational objective was "gather information from a patient", this could be taught with work-related problems in small groups using simulate patients. Scenarios should be based on authentic patients' cases representing a realistic range of health problems from the community. An assessment should also include authentic and relevant patients' problems, using the methodology of Objective Structured Clinical Examination (OSCE) or workplace-based assessment with real patients. Blueprinting is the process of designing the assessment program in such a way that a balance of aspects are covered, such as knowledge domains, levels of expertise, as well as various diseases, organ systems, patient characteristics, and settings of care. Blueprinting encourages sampling from across the curriculum and guarantees that assessment is seen as an integral part of the communication curriculum as a whole [17].

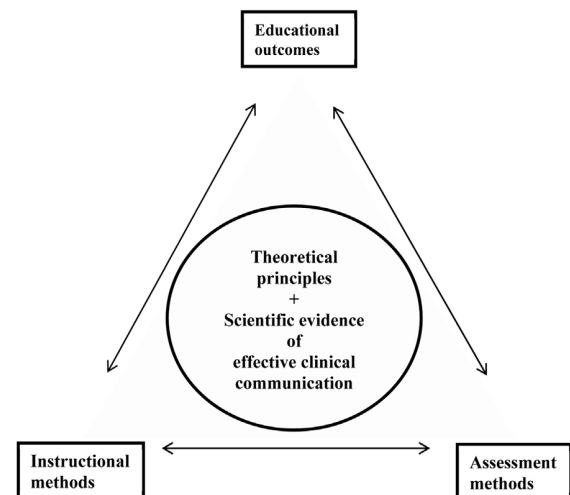


Fig. 1. Global framework on assessment of clinical communication.

### 2.1. Attitudes and "multidimensional constructs"

The most difficult challenge in education is assessing learners' attitudes and other "multidimensional constructs", like empathy or patient-centeredness. Using the example of clinical empathy, Stepien and Baernstein describe four dimensions: the emotive dimension, the moral dimension, the cognitive dimension, and the behavioral dimension [18]. In comparison to technical skills, like venepuncture, multidimensional constructs are much more difficult to measure.

Hemmerdinger and colleagues recommend classifying instruments measuring a construct like empathy from three different viewpoints: self-ratings (first person assessment), patient-ratings (second person assessment), and observer ratings (third person assessment) [7]. Each of these ratings may have a place in formative and summative assessment, depending on the purpose of the assessment. A typical first person assessment would be a questionnaire asking learners to estimate their own ability to communicate with other or express empathy (e.g. Jefferson Scale of Physician Empathy, JSPE) [19]. A typical second person assessment would be a questionnaire asking patients to express their satisfaction with the provider's communication or to measure the extend of empathy expressed by the provider (e.g. Consultation and relational empathy, CARE) [20]. First and second person instruments may have the potential to enhance self-reflection in communication training and in formative assessment. Third person instruments typically focus on the behavioral dimension, for example demonstration of verbal or non-verbal clinical empathy in terms of patient-centered approach. They are commonly used in OSCEs and other observation-based settings to assess behaviorally measurable skills rather than intention [21].

### 2.2. Assessment in context

Clinical communication is content and context bound [22]. In order to prepare learners in the healthcare professions for later professional reality effectively, learning communication should take place in contexts closely resembling clinical practice [23]. This means that teaching communication skills and their assessment needs to be integrated in the clinical context, either in clinical practice or in clinically relevant simulations [9,24]. This principle derives from socio-cultural learning theories, particularly the theory of situated learning. Learning is a function of the activity, context and culture in which it occurs [25]. Even in the early years of medical training, when students have less patient contact, it is



helpful to use clinical examples and simulations to teach basic communication. For many students, it makes little intuitive sense to isolate communication from later professional practice. The same principles can be applied to assessment which means to embed assessment scenarios into a meaningful context, combined with and integrated into the assessment of other clinical skills, knowledge and clinical reasoning. In this way, learners are assessed approaching a clinically realistic task in which communication is one of the important components needed to achieve the goals of evidence-based clinical practice.

### 2.3. Establishing a program of formative and summative assessment

Assessment can have different purposes. On the one hand, formative assessment will detect strengths and areas for improvement of students, and will offer structure for guiding future learning, provide reassurance and motivation. It is often integrated into teaching, timely, informal, and has low stakes. Summative assessment on the other hand, usually is intended to reach decisions about competence and readiness for moving to the next stage in the curriculum. In high stakes examinations, summative assessment acts as a barrier to protect patients and the public by identifying incompetent current or future health care providers [26]. Both formative and summative aspects of assessment are useful to be taken into consideration when planning a comprehensive assessment system.

In formative assessment, educational impact for learning may be more important, whereas in summative assessment, the tools may need to meet higher psychometric requirements. Reproducibility of decisions based upon the test results is essential when we draw decisions about professional advancement.

In summative assessments, the pass/fail cut-off point needs to be defined [27], based on an established (transparent) standard setting procedure which aims to fairly judge examinee's performance. Different procedures have been described [28,29]. Friedman Ben-David offers a categorization between norm-referenced and criterion-referenced standards. Norm-referenced standards are based on external representative samples (norm groups). An example for norm-referenced standards would be an admission test. The aim is to select a number of the best applicants that is equivalent to a predefined percentage. Criterion-referenced standards link the standard to the content of the competence level under consideration [27]. Most standard setting procedures used for assessing clinical communication are criterion-referenced standards (e.g. Angoff, Ebel, methods based on borderline groups). Here, you decide that each student that will pass the test shows a predefined level of competence. The choice of procedure depends on the assessment method and tool to be used, the resources available and the consequences of misclassifying examinees as having passed or failed [22,27,30-37].

## 3. Choosing the appropriate assessment level

Numerous different assessment tools have been described in the literature, which are mainly applicable to the assessment of communication skills within an educational program [2]. There are different ways of categorizing assessment strategies and

instruments. A simple way is to divide them according to the level of competence they aim to assess: basic and applied knowledge, performance in standardized quasi-real situations, and performance or action in the work-place, as presented by Miller (Table 1) [38]. There are some general criteria to consider when exploring "the utility of an assessment" [17]:

- Reliability (the degree to which the measurement is reproducible and consistent),
- Validity (whether the assessment measures what it claims to measure),
- Educational impact (impact on future learning and practice),
- Credibility (to students and faculty),
- Feasibility and costs (to individual trainee, the institution, the society).

These criteria will be discussed while introducing different assessment methods according to the above-mentioned level of competence.

### 3.1. Knowledge and written assessment to assess basic and applied knowledge

Multiple methods of written assessment of communication knowledge have been developed. Questions can be open-ended or multiple choice (which means selecting the best possible answer from a list); context-rich (e.g. with a clinical vignette) or context-poor; media-enhanced (e.g. video, audio); paper-and-pencil-based or computer-based. Written assessment can be administered in a short time with high standardization. With clear construction and grading guidelines, different types of written assessments tend to reach a high reliability per hour of testing [22]. Written assessment allows testing of different types of cognitive abilities: knowledge and understanding of facts, processes and concepts in communication. It does not allow for the testing of skills, although it may be able to predict the performance of skills [39,40]. Nevertheless, there seems to be a place for written assessment in the field of communication, especially in the beginning of training and following observation of prepared videos [41]. Advancements in written assessment methods include the script concordance test [42], key feature assessment [43], situational judgment test [44], and reflective portfolio assessment [45].

### 3.2. Performance in clinical simulations: performance in standardized quasi-real situations

A highly established assessment method using clinical simulation is the OSCE [21,46,47]. Its implementation into medical education has become a world-wide success. Students complete a number of stations with standardized patients trained to consistently repeat typical clinical situations. Students perform medical interviews, physical examinations and clinical procedures. Raters – standardized patients, clinicians or faculty members – use either a checklist or global rating scale to evaluate students' performance. In order to achieve acceptable reliability, it is important to consider the number of cases, stations, examiners and patients [48]. For example, with an overall testing time of three

**Table 1**  
Matching assessment strategies to the increasing level of competence, including some examples (according to Miller).

Competence level	Appropriate assessment method
Knows	Factual tests: MCQ, Essay type, Oral examination
Knows how	(Clinical) Context based tests: MCQ, Essay type, Oral examination
Shows how	Simulated performance assessment: OSCE, SP-based test
Does	Workplace based performance assessment: Video review, Direct observation, Masked SP consultations

to four hours and a minimum of ten stations, OSCEs are reliable enough for high stakes examinations, e.g. federal licensing examinations [22,49,50]. However, there is still a need for further educational research focusing on communication, as this seems harder to assess reliably than other clinical skills [48].

There is an ongoing discussion about the use of communication skills checklists versus global ratings for the assessment of clinical competence, both in OSCEs and in workplace assessment. Important aspects for consideration include, among others, reliability (e.g. between different raters, between different OSCE stations), validity (e.g. reducing complex competencies to easily measurable but trivial patterns of behaviors), and feasibility (e.g. amount of time for rater training). At the moment, there seems to be a slight preference for global ratings concerning validity [51–54]. For the field of communication, Setyonugroho et al. stated in a systematic review of reliability and validity of checklists, that the most striking finding was “a demonstrated absence of consensus in rubrics used to assess communication skills in undergraduate medical education worldwide” [55]. Despite this, it remains important that the assessment instrument matches the educational objectives and teaching models for maximal educational impact [15].

Many observation-based assessments focus on dyadic provider-patient encounters. Simulation of other real-world tasks has become increasingly important such as triadic consultations, collaborative clinical reasoning and teamwork [26]. One way to assess these complex clinical tasks are computer simulations, and high fidelity simulators, providing feedback to learners [49,56–58]. Therefore, new developments in the field of simulation-based education also offer innovative research opportunities for training and assessing communication in realistic health care settings.

### 3.3. Direct observation and assessment of every-day performance in the work-place

Observations by supervisors, peers, co-workers, or other health professions are commonly used to evaluate learners' performance with patients [59]. These observations can be unstructured or structured e.g. by using a checklist or rating scale to direct the observation towards specific tasks. The mini-CEX has become a helpful instrument to support structured feedback for postgraduate training in medicine [59,60]. It may turn unstructured observations into structured assessment moments providing feedback to learners [27]. If these observations are collected over periods of time, they can be turned from “single event measures” into “global performance measures”. Such examples are portfolios and 360° or multi-source feedback. The formative character of multisource feedback from direct observation of practice situations may be valuable for future learning [61].

Patients' ratings may add to the value of formative assessment [62]. The patients' perspective can serve as an indicator of the quality of doctor/provider-patient communication and provide added value to the assessment performed by faculties and peers. Real or standardized patients can be valuable in the assessment of diagnostic reasoning, treatment decision, and communication [63,64]. However, patients and doctors may not necessarily agree about the quality of the interaction, and may have contradictory goals. A satisfied patient can still have an incompetent doctor. In the era of patient-centered medicine, patients' views definitely matter and routine collection of patient feedback in assessing service quality is increasingly adopted. A number of validated patient rating instruments has already been published and seem to be feasible for assessment purposes in medical education [65–67]. However, more evidence is needed on clinical communication assessment, including patients as raters [15,26,68–71].

### 3.4. Ratings based on videotaped encounters

Assessment of clinical communication in the workplace and of simulated encounters can be done by rating videotaped consultations [62]. It allows learners to review and analyze their own behavior, and third persons (including peers) can be invited to assess the performance. Similar to simulation-based assessments, the health provider's communication can be measured by rating scales and checklists. If an adequate case mix in practice can be achieved, a sample of eight consultations can suffice for a valid and reliable assessment [72].

## 4. How to construct an assessment program for clinical communication?

### 4.1. Establishing a multi-source and longitudinal assessment program

All assessment methods have strengths and weaknesses [46]. For example, written tests are cost-effective because you can assess an unlimited number of students with one test. However, written assessment methods only test knowledge and not performance or behavior. Observation-based assessments are on the other hand powerful in assessing clinical skills, but their realization is costly and time-consuming. Therefore, it is important to align assessment methods to the level of expertise to be addressed and to the resources available. “Each single assessment is a biopsy, and a series of biopsies will provide a more complete, more accurate picture” [52]. Clinical communication is a complex skill. Its utilization is based on factual and procedural knowledge which allows experts, in combination with their experience, to apply and adapt clinical communication in various settings with various patients. Therefore, it is important to use a variety of different assessment formats, multiple observations, independent measurements in different settings and preferably using multiple assessors. This principle mimics the principle of triangulation in qualitative research [36,73,74].

### 4.2. Providing feedback to enhance learning

There is a growing awareness of the connection between assessment, feedback and continuous learning [8,75]. Assessment, especially formative assessment, is most helpful to learners when accompanied by concrete specific feedback. Moreover, although summative assessment concentrates on pass-fail decisions, it is also useful to provide feedback synchronously to learners, to increase the educational impact of the assessment [51]. Constructive feedback from teachers and peers will help learners understand and accept the criteria that teachers apply in the assessment of their communication [76–78].

### 4.3. How to select assessment tools

The purpose of an assessment influences the choice of the instrument: if the assessment is summative, requirements of validity and reliability are high. A selection of different assessment tools are on the tEACH website [79]. Most of these tools are directed at assessing communication in an entire patient-provider consultation in real or simulated settings. This mirrors the current state of assessment with a strong focus on direct observation of real and simulated encounters. Some tools have been developed for specific professions, or for assessing specific topics or parts of the consultation process (e.g. patient's perspective and health beliefs, clinical reasoning or shared decision making), highlighting the context specificity of communication assessment.

**Table 2**  
Steps to consider when developing a communication assessment program.

Questions	Steps
What to assess	Assessment blueprinting based on the global framework and including <ul style="list-style-type: none"> <li>- knowledge domains</li> <li>- levels of expertise</li> <li>- various diseases, organ systems</li> <li>- patient characteristics</li> <li>- settings of care</li> </ul>
How to assess Choosing the appropriate assessment level	Clinical context-based and integrated with other clinical skills <ul style="list-style-type: none"> <li>- Clinical relevant simulation</li> <li>- Clinical practice</li> <li>1. Basic and applied knowledge               <ul style="list-style-type: none"> <li>- written assessment (consider type of question, context, media)</li> </ul> </li> <li>2. Performance in quasi-real situations               <ul style="list-style-type: none"> <li>- OSCE</li> </ul> </li> <li>3. Performance in everyday practice               <ul style="list-style-type: none"> <li>- Work-based (e.g. mini-CEX)</li> </ul> </li> <li>4. Videotaped encounter-based</li> </ul>
How to construct an assessment program	Two formats: <ul style="list-style-type: none"> <li>(a) Formative</li> <li>(b) Summative</li> </ul> Multisource Longitudinal External feedback and self-assessment
How to choose an assessment tool	According to learning goals and levels of competence (Miller's pyramid) Criteria of reliability, validity, educational impact, credibility, feasibility

## 5. Conclusion and practice implications

### 5.1. Conclusion

Since assessment drives learning, the implementation of a reliable, valid, credible, and feasible assessment method with educational impact is a vital component of clinical communication curricula. Establishing a coherent entity of teaching, training and assessment, and assessing communication within the clinical context form the basis for a coherent competency development. In terms of how to implement assessment, establishing a system with both formative and summative assessment, providing feedback that enhances learning and establishing a multi-source and longitudinal assessment program are presented. The last decades have brought a large array of assessment tools that can be used for assessing communication. Table 2 summarizes the steps to consider when developing a communication assessment program.

### 5.2. Practice implications

The ideal communication assessment tool does not exist. All have advantages and disadvantages and need to be matched carefully to the educational and professional context. Implementing assessment in the curriculum will need careful change management to guarantee sustainability and continuous improvement, and will in particular require support from colleagues, superiors and teachers for the chosen assessment approaches.

### Funding

No funding was received.

### Conflict of interest

The authors declare no conflict of interest.

### Acknowledgements

We sincerely thank Dr. GötzFabry (Freiburg, Germany), Anja Härtl (Munich, Germany), Dr. Robert L. Hulsman (Amsterdam, The Netherlands), and Tanja Pander (Munich, Germany) for critically reviewing the paper. We would also like to thank Katarzyna Jankowska (Poland) and Rute Meneses (Portugal), ex-members of the Assessment group of t-EACH for their contribution in reviewing the initial paper.

### References

- [1] J.R. Frank, J. Sherbino, The Draft CanMEDS 2015 Physician Competency Framework – Series IV, TRCoPaSo, Ottawa, 2015.
- [2] J. Heyrman, EURACT Educational Agenda, (2005) . (Assessed 10 March 2016) <http://www.euract.org/pdf/agenda.pdf>.
- [3] R.L. Hulsman, The art of assessment of medical communication skills, *Pat. Educ. Couns.* 83 (2011) 143–144.
- [4] L.W. Schuwirth, C.P. van der Vleuten, General overview of the theories used in assessment: AMEE Guide No. 57, *Med. Teach.* 33 (2011) 783–797.
- [5] H. Boon, M. Stewart, Patient-physician communication assessment instruments: 1986 to 1996 in review, *Pat. Educ. Couns.* 35 (1998) 161–176.
- [6] J.M. Schirmer, L. Mauksch, F. Lang, M.K. Marvel, K. Zoppi, R.M. Epstein, D. Brock, M. Pryzbylski, Assessing communication competence: a review of current tools, *Fam. Med.* 37 (2005) 184–192.
- [7] J.M. Hemmerdinger, S.D. Stoddart, R.J. Lilford, A systematic review of tests of empathy in medicine, *BMC Med. Educ.* 7 (2007) 24.
- [8] J.J. Norcini, B. Anderson, V. Bollala, V. Burch, M.J. Costa, R. Duvivier, R. Galbraith, R. Hays, A. Kent, V. Perrott, T. Roberts, Criteria for good assessment: consensus statement and recommendations from the Ottawa 2010 Conference, *Med. Teach.* 33 (2011) 206–214.
- [9] J. Silverman, Teaching clinical communication: a mainstream activity or just a minority sport? *Pat. Educ. Couns.* 76 (2009) 361–367.

- [10] P.A. Thomas, D.E. Kern, M.T. Hughes, B.Y. Chen, Curriculum Development for Medical Education: A Six-Step Approach, The Johns Hopkins University Press, Baltimore, 2016.
- [11] S. Rollnick, P. Kinnersley, C. Butler, Context-bound communication skills training: development of a new method, *Med. Educ.* 36 (2002) 377–383.
- [12] L.A. Baig, C. Violato, R.A. Crutcher, Assessing clinical communication skills in physicians: are the skills context specific or generalizable? *BMC Med. Educ.* 9 (2009) 22.
- [13] J. Lave, E. Wenger, *Situated Learning: Legitimate Peripheral Participation*, Cambridge University Press, Cambridge, 1991.
- [14] L. Vygotsky, *Mind in Society: Development of Higher Psychological Processes*, Cambridge University Press, Cambridge, 1978.
- [15] F.D. Duffy, G.H. Gordon, G. Whelan, K. Cole-Kelly, R. Frankel, Assessing competence in communication and interpersonal skills: the Kalamazoo II report, *Acad. Med.* 79 (2004) 495–507.
- [16] T.J. Wilkinson, W.B. Wade, L.D. Knock, A blueprint to assess professionalism: results of a systematic review, *Acad. Med.* 84 (2009) 551–558.
- [17] L. Coombes, M. Roberts, D. Zahra, S. Burr, Twelve tips for assessment psychometrics, *Med. Teach.* 38 (2016) 250–254.
- [18] K.A. Stepien, A. Baernstein, Educating for empathy. A review, *J. Gen. Intern. Med.* 21 (2006) 524–530.
- [19] M. Hojat, J.S. Gonnella, K. Maxwell, *Jefferson Scales of Empathy (JSE). Professional Manual & User's Guide*, Jefferson Medical College, Philadelphia, 2009.
- [20] M. Wirtz, M. Boecker, T. Forkmann, M. Neumann, Evaluation of the "Consultation and Relational Empathy" (CARE) measure by means of Rasch – analysis at the example of cancer patients, *Patient Educ. Couns.* 82 (2011) 298–306.
- [21] B. Hodges, M. Hanson, N. McNaughton, G. Regehr, Creating, monitoring, and improving a psychiatry OSCE: a guide for faculty, *Acad. Psychiatry* 26 (2002) 134–161.
- [22] V. Wass, C.P.M. van der Vleuten, J. Shatzer, R. Jones, Assessment of clinical competence, *Lancet* 357 (2001) 945–949.
- [23] L.W.T. Schuwirth, C.P.M. van der Vleuten, Changing education changing assessment, changing research? *Med. Educ.* 38 (2004) 805–812.
- [24] S. Kurtz, J. Silverman, J. Benson, J. Draper, Marrying content and process in clinical method teaching: enhancing the Calgary–Cambridge guides, *Acad. Med.* 78 (2003) 802–809.
- [25] K.V. Mann, Theoretical perspectives in medical education: past experience and future possibilities, *Med. Educ.* 45 (2011) 60–68.
- [26] R.M. Epstein, Assessment in medical education, *N. Engl. J. Med.* 356 (2007) 387–396.
- [27] M. Friedman Ben David, AMEE guide No. 18: standard setting in student assessment, *Med. Teach.* 22 (2000) 120–130.
- [28] S.M. Downing, A. Tekian, R. Yudkowsky, Procedures for establishing defensible absolute passing scores on performance examinations in health professions education, *Teach. Learn. Med.* 18 (2006) 50–57.
- [29] K. Boursicot, T. Roberts, Setting standards in a professional higher education course: defining the concept of the minimally competent student in performance based assessment at the level of graduation from medical school, *High. Educ. Q.* 60 (2006) 74–90.
- [30] J.C. De Haes, F.J. Oort, R.L. Hulsmann, Summative assessment of medical students' communication skills and professional attitudes through observation in clinical practice, *Med. Teach.* 27 (2005) 583–589.
- [31] C.H. Shah, D. Parmar, R. Parmar, Study of standard setting in constructed response type written examination, *Int. J. Med. Sci. Public Health* 3 (2014) 1046–1050.
- [32] S.M. Downing, A. Tekian, R. Yudkowsky, Procedures for establishing defensible absolute passing scores on performance examinations in health professions education, *Teach. Learn. Med.* 18 (2006) 50–57.
- [33] A. Kramer, A. Muijtjens, K. Jansen, H. Dusman, L. Tan, C. van der Vleuten, Comparison of a rational and an empirical standard setting procedure for an OSCE. Objective structured clinical examinations, *Med. Educ.* 37 (2003) 132–139.
- [34] J.J. Norcini, Setting standards on educational tests, *Med. Educ.* 37 (2003) 464–469.
- [35] T. Wood, S. Humphrey-Murto, G. Norman, Standard setting in a small scale OSCE: a comparison of the modified borderline-group method and the borderline regression method, *Adv. Health Sci. Educ. Theory Pract.* 11 (2006) 115–122.
- [36] D.W. McKinley, J.J. Norcini, How to set standards on performance-based examinations: AMEE Guide No. 85, *Med. Teach.* 36 (2014) 97–110.
- [37] N. Yousuf, C. Violato, R.W. Zuberi, Standard setting methods for pass/fail decisions on high-stakes objective structured clinical examinations: a validity study, *Teach. Learn. Med.* 27 (2015) 280–291.
- [38] G.E. Miller, The assessment of clinical skills/competence/performance, *Acad. Med.* 65 (1990) S63–S67.
- [39] J. van Dalen, E. Kerkhofs, G.M. Verwijnen, B.W. van Knippenberg-van den Berg, H.A. van den Hout, A.J. Scherpbier, C.P. van der Vleuten, Predicting communication skills with a paper-and-pencil test, *Med. Educ.* 36 (2002) 148–153.
- [40] P. Ram, C. van der Vleuten, J.J. Rethans, B. Schouten, S. Hobma, R. Grol, Assessment in general practice: the predictive value of written- knowledge tests and a multiple-station examination for actual medical performance in daily practice, *Med. Educ.* 33 (1999) 197–203.
- [41] G.M. Humphris, S. Kaney, The objective structured video exam for assessment of communication skills, *Med. Educ.* 34 (2000) 939–945.
- [42] B. Charlin, C.P. Van der Vleuten, Standardized assessment of reasoning in contexts of uncertainty. The script concordance test, *Eval. Health Prof.* 27 (2004) 304–319.
- [43] G. Page, G. Bordage, T. Allen, Developing key-feature problems and examinations to assess clinical decision-making skills, *Acad. Med.* 70 (1995) 194–201.
- [44] J. Strahan, G.J. Fogarty, M.A. Machin, Predicting performance on a situational judgement test: the role of communication skills, listening skills, and expertise, *Proceedings of the 40 Annual Conference of the Australian Psychological Society* (2005) 323–327.
- [45] C.E. Rees, C.E. Sheard, The reliability of assessment criteria for undergraduate medical students' communication skills portfolios: the Nottingham experience, *Med. Educ.* 38 (2004) 138–144.
- [46] L.W. Schuwirth, C.P. van der Vleuten, ABC of learning and teaching in medicine. Written assessment, *Br. Med. J.* 326 (2003) 643–645.
- [47] D. Newble, Techniques for measuring clinical competence: objective structured clinical examinations, *Med. Educ.* 38 (2004) 199–203.
- [48] M.T. Brannick, H.T. Erol-Korkmaz, M. Prewett, A systematic review of the reliability of objective structured clinical examination scores, *Med. Educ.* 45 (2011) 1181–1189.
- [49] C.P. van der Vleuten, L.W. Schuwirth, Assessing professional competence: from methods to programmes, *Med. Educ.* 39 (2005) 309–317.
- [50] C. Berendonk, C. Schirlo, G. Balestra, R. Bonvin, S. Feller, P. Huber, E. Jünger, M. Monti, K. Schnabel, C. Beyeler, S. Guttormsen, S. Huwendiek, The new final clinical skills examination in human medicine in Switzerland: essential steps of exam development, implementation and evaluation, and central insights from the perspective of the National Working Group, *GMS Z. Med. Ausbild.* 32 (2015), doi:http://dx.doi.org/10.3205/zma000982 Doc40.
- [51] M. van Nuland, W. van den Noortgate, C. van der Vleuten, J. Goedhuys, Optimizing the utility of communication OSCEs: omit station-specific checklists and provide students with narrative feedback, *Patient Educ. Couns.* 88 (2012) 106–112.
- [52] C.P. van der Vleuten, L.W. Schuwirth, F. Scheele, E.W. Driessen, B. Hodges, The assessment of professional competence: building blocks for theory development, *Best Pract. Res. Clin. Obstet. Gynaecol.* 24 (2010) 703–719.
- [53] G. Regehr, H. MacRae, R.K. Reznick, D. Szalay, Comparing the psychometric properties of checklists and global rating scales for assessing performance on an OSCE-format examination, *Acad. Med.* 73 (1998) 993–997.
- [54] B. Hodges, G. Regehr, N. McNaughton, R. Tiberius, M. Hanson, OSCE checklists do not capture increasing levels of expertise, *Acad. Med.* 74 (1999) 1129–1134.
- [55] W. Setyonugroho, K.M. Kennedy, T.J. Kropmans, Reliability and validity of OSCE checklists used to assess the communication skills of undergraduate medical students: a systematic review, *Patient Educ. Couns.* 98 (2015) 1482–1491, doi:http://dx.doi.org/10.1016/j.pec.2015.06.004 pii: S0738-3991(15)00277-3.
- [56] S.B. Issenberg, W.C. McGaghie, E.R. Petrusa, D.L. Gordon, R.J. Scalese, Features and uses of high-fidelity medical simulations that lead to effective learning, *Med. Teach.* 27 (2005) 10–28.
- [57] P. Bradley, The history of simulation in medical education and possible future directions, *Med. Educ.* 40 (2006) 254–262.
- [58] R. Hovey, M. Dvorak, M. Hatlie, T. Burton, J. Padilla, S. Worsham, A. Morck, Patient safety: a consumer's perspective, *Qual. Health Res.* 21 (2011) 662–672.
- [59] J.R. Kogan, E.S. Holmboe, K.E. Hauer, Tools for direct observation and assessment of clinical skills of medical trainees: a systematic review, *JAMA* 302 (2009) 1316–1326.
- [60] J. Norcini, V. Burch, Workplace-based assessment as an educational tool: AMEE guide No. 31, *Med. Teach.* 29 (2007) 855–871.
- [61] G. Essers, A. Kramer, B. Andriess, C. van Weel, C. van der Vleuten, S. van Dulmen, Context factors in general practitioner-patient encounters and their impact on assessing communication skills – an exploratory study, *BMC Fam. Pract.* 14 (2013) 65.
- [62] M.E. Reinders, A.H. Blankenstein, H.W. van Marwijk, D.L. Knol, P. Ram, H.E. van der Horst, H.C. de Vet, C.P. van der Vleuten, Reliability of consultation skills assessments using standardised versus real patients, *Med. Educ.* 45 (2011) 578–584.
- [63] J.J. Rethans, S. Gorter, L. Bokken, L. Morrison, Unannounced standardised patients in real practice: a systematic literature review, *Med. Educ.* 41 (2007) 537–549.
- [64] L.A. Siminoff, H.L. Rogers, A.C. Wallera, S.H. Hayward, R.M. Epstein, F. BorrellCarro, G. Gliva-McConvey, D.R. Longof, The advantages and challenges of unannounced standardized patients methodology to assess healthcare communication, *Patient Educ. Couns.* 82 (2011) 318–324.
- [65] M.E. Reinders, A.H. Blankenstein, D.L. Knol, H.C.W. de Vet, H.W.J. van Marwijk, Validity aspects of the Patient Feedback questionnaire on Consultation Skills (PFC), a promising learning instrument in medical education, *Patient Educ. Couns.* 76 (2009) 2002–2006.
- [66] G. Makoul, E. Krupat, C.H. Chang, Measuring patient views of physician communication skills: development and testing of the Communication Assessment Tool, *Patient Educ. Couns.* 67 (2007) 333–342.
- [67] M. Greco, M. Cavanagh, A. Browlea, J. Mc Govern, The doctor's interpersonal skills questionnaire (DISQ): a validated instrument for use in GP training, *Educ. Gen. Pract.* 10 (1999) 256–264.
- [68] M.J.B. Govaerts, L.W. Schuwirth, C.P. Van der Vleuten, A.M.M. Muijtjens, Workplace-based assessment: effects of rater expertise, *Adv. Health Sci. Educ.* 16 (2011) 151–165.



- [69] P. Yeates, P. O'Neill, K. Mann, K. Eva, Seeing the same thing differently: mechanisms that contribute to assessor differences in directly-observed performance assessments, *Adv. Health Sci. Educ. Theory Pract.* 18 (2013) 325–341.
- [70] G. Essers, P. Dielessen, C. van Weel, C. van der Vleuten, S. van Dulmen, A. Kramer, How do trained raters take context factors into account when assessing GP trainee communication performance? An exploratory, qualitative study, *Adv. Health Sci. Educ. Theory Pract.* 20 (2015) 131–147.
- [71] R. Hovey, H. Massfeller, Exploring the relational aspects of patient and doctor communication, *J. Med. Pers.* 10 (2012) 81–86.
- [72] P. Ram, R. Grol, J.J. Rethans, B. Schouten, C. van der Vleuten, A. Kester, Assessment of general practitioner by video observation of communicative and medical performance in daily practice: issues of validity, reliability and feasibility, *Med. Educ.* 33 (1999) 447–454.
- [73] L.A. Baig, C. Violato, R.A. Crutcher, Assessing clinical communication skills in physicians: are the skills context specific or generalizable? *BMC Med. Educ.* 9 (2009) 22.
- [74] J.L. Turner, M.E. Dankowski, Objective structured clinical exams: a critical review, *Fam. Med.* 40 (2008) 574–578.
- [75] L. Konopasek, J. Norcini, E. Krupat, Focusing on the formative: building an assessment system aimed at student growth and development, *Acad. Med.* (2016).
- [76] F. Dochy, M. Segers, D. Sluijsmans, The use of self-, peer and co-assessment in higher education: a review, *Stud. High. Educ.* 24 (1999) 331–350.
- [77] S. Gielen, Peer assessment as a tool for learning (dutch) Dissertation. +S. +GIELEN.pdf (Accessed 6 August 2012)
- [78] R.L. Hulsman, J.F. Peters, M. Fabriek, Peer-assessment of medical communication skills: the impact of students' personality, academic and social reputation on behavioural assessment, *Patient Educ. Couns.* 92 (2013) 346–354.
- [79] Assessment group of t-EACH, Assessment tools. <http://www.each.eu/teaching/can-Tumerteach-offer/assess/assessment-tools/> (Assessed 15 April 2016).



# BMJ Open Assessing communication quality of consultations in primary care: initial reliability of the Global Consultation Rating Scale, based on the Calgary-Cambridge Guide to the Medical Interview

Jenni Burt,<sup>1</sup> Gary Abel,<sup>1</sup> Natasha Elmore,<sup>1</sup> John Campbell,<sup>2</sup> Martin Roland,<sup>1</sup> John Benson,<sup>3</sup> Jonathan Silverman<sup>4</sup>

**To cite:** Burt J, Abel G, Elmore N, *et al.* Assessing communication quality of consultations in primary care: initial reliability of the Global Consultation Rating Scale, based on the Calgary-Cambridge Guide to the Medical Interview. *BMJ Open* 2014;**4**: e004339. doi:10.1136/bmjopen-2013-004339

► Additional material is available. To view please visit the journal (<http://dx.doi.org/10.1136/bmjopen-2013-004339>).

Received 25 October 2013  
Revised 29 January 2014  
Accepted 13 February 2014



CrossMark

For numbered affiliations see end of article.

#### Correspondence to

Dr Jenni Burt;  
jab35@medschl.cam.ac.uk

## ABSTRACT

**Objectives:** To investigate initial reliability of the Global Consultation Rating Scale (GCRS: an instrument to assess the effectiveness of communication across an entire doctor–patient consultation, based on the Calgary-Cambridge guide to the medical interview), in simulated patient consultations.

**Design:** Multiple ratings of simulated general practitioner (GP)–patient consultations by trained GP evaluators.

**Setting:** UK primary care.

**Participants:** 21 GPs and six trained GP evaluators.

**Outcome measures:** GCRS score.

**Methods:** 6 GP raters used GCRS to rate randomly assigned video recordings of GP consultations with simulated patients. Each of the 42 consultations was rated separately by four raters. We considered whether a fixed difference between scores had the same meaning at all levels of performance. We then examined the reliability of GCRS using mixed linear regression models. We augmented our regression model to also examine whether there were systematic biases between the scores given by different raters and to look for possible order effects.

**Results:** Assessing the communication quality of individual consultations, GCRS achieved a reliability of 0.73 (95% CI 0.44 to 0.79) for two raters, 0.80 (0.54 to 0.85) for three and 0.85 (0.61 to 0.88) for four. We found an average difference of 1.65 (on a 0–10 scale) in the scores given by the least and most generous raters: adjusting for this evaluator bias increased reliability to 0.78 (0.53 to 0.83) for two raters; 0.85 (0.63 to 0.88) for three and 0.88 (0.69 to 0.91) for four. There were considerable order effects, with later consultations (after 15–20 ratings) receiving, on average, scores more than one point higher on a 0–10 scale.

**Conclusions:** GCRS shows good reliability with three raters assessing each consultation. We are currently developing the scale further by assessing a large sample of real-world consultations.

## Strengths and limitations of this study

- The Global Consultation Rating Scale (GCRS) is based on the widely used Calgary-Cambridge guide to the medical interview, and is designed to evaluate a practitioner's communication skills across an entire consultation, linking the identification of potential training needs to an established approach to teaching communication skills.
- We considered evaluator bias and order effects to obtain a more robust assessment of the reliability of GCRS to evaluate communication competence within a particular consultation.
- A particular limitation is that our findings are based on the use of simulated patient consultations. This had an impact on our ability to assess the performance of GCRS to evaluate communication competence of individual doctors, rather than particular consultations. A full evaluation of the performance of GCRS requires the assessment of real-world consultations and we are undertaking this at present.

## INTRODUCTION

During the past 30 years, an extensive research literature has defined the skills that enhance communication between doctor and patient. This evidence demonstrates the essential role that communication plays in high-quality healthcare by enabling more accurate, efficient and supportive interviews, by enhancing patient and professional experience and by improving health outcomes for patients. The use of specific communication skills has been shown to lead to improvements in symptom relief, in clinical outcomes and possibly in medicine adherence.<sup>1–6</sup> In light of these findings,

## Open Access



there has been increasing pressure from professional medical bodies to improve the training and evaluation of doctors in communication.<sup>7-13</sup>

In order to evaluate doctors' communication skills effectively, tools with solid theoretical grounding and good psychometric properties are required. Various rating scales exist to assess doctor-patient consultations, which vary widely in their setting, approach and in the published details of their psychometric properties.<sup>14 15</sup> Perhaps for these reasons, none have become standard to use within the National Health Service (NHS), in spite of National Institute for Health and Care Excellence (NICE) standards which require that "Patients experience effective interactions with staff who have demonstrated competency in relevant communication skills."<sup>16</sup> Recently, there has been a move towards domain, or global, marking schemes (awarding overall marks to groupings of items) rather than itemised checklists, the suggestion being that checklists may reward thoroughness rather than competence and work better for novices than for experts.<sup>17</sup> Global marking schemes may be more useful in postgraduate assessments, improving professional authenticity. We have, therefore, developed the Global Consultation Rating Scale (GCRS), based on the Calgary-Cambridge guide to the medical interview, to evaluate the communication effectiveness of an entire doctor-patient consultation, using the domain marking approach.

At present, there is a dearth of assessment tools that robustly measure the overall communication skills of an individual general practitioner (GP) in real-world practice. While a number of existing tools may be used to assess doctor-patient communication, their suitability to assess a doctor's overall communication skills in day-to-day practice irrespective of the content of the consultation is limited and they do not link specifically to educational material commonly used in the UK for subsequent communication skills development. GCRS differs from some alternative instruments, such as the MAAS-Global, in its aim of measuring communication skills only, irrespective of clinical content, to provide an assessment of doctors' generic communication skills and to thereby enable targeted communication teaching. For example, 4 of the 17 items in the MAAS-Global specifically assess medical content related to history, examination, diagnosis and management and other communication items are highly specific to particular content areas.<sup>18</sup> In comparison, the 12 global areas of GCRS include only communication process skills without content. Following the approach of the Calgary-Cambridge guide from which it is derived, GCRS takes the standpoint that, although the context of the interaction changes and the content of the communication varies, the process skills themselves remain the same and can be evaluated independently. This, together with domain rather than individual skill marking, enables the assessment of communication skills across a wide variety of consultations, especially helpful in real-world

consultations where communication checklists cannot be specific and tailored for each case.

The Calgary-Cambridge guide to the medical interview<sup>1 19-21</sup> was developed by Silverman, Kurtz and Draper to delineate effective physician-patient communication skills and to provide an evidence-based structure for their analysis and teaching. Within the UK, over half of UK medical schools now use the Calgary-Cambridge approach in their communication skills programmes.<sup>22</sup> It has been widely translated and is used in the USA, Canada and Europe. It has been used to teach communication in general practice and specialist environments, at undergraduate and postgraduate levels.

Specific tools have been developed from the guide for the assessment of medical students, practising paediatricians, dentists, pharmacists and veterinary practitioners, as well as for specific components of the consultation such as explanation and planning in OSCE style examinations.<sup>23-25</sup> Before now however, there has been no validated method of using the Calgary-Cambridge consultation guide to assess complete consultations between qualified doctors and patients. This type of assessment is particularly important in postgraduate and continuing medical education in which the observation of whole consultations from real practice provides increased validity. In addition, for personal development and annual appraisal, a reliable validated assessment tool which also enables a specific link to targeted teaching of communication skills is particularly relevant. Our intention with GCRS is to develop an instrument capable of credibly evaluating a doctor's communication competence, identifying potential areas for improvement which could then be addressed directly with linked, tailored education, using the Calgary-Cambridge guide.

The aim of this study was to investigate the initial reliability of GCRS in simulated patient consultations such as those which might be used in training, as a precursor to its use with real patient consultations where GPs are assessed on their performance. To assess reliability, we asked five specific questions. These are detailed below, together with the reasons for their investigation:

- A. Does a fixed difference between scores in GCRS have the same meaning at all levels of performance? If it does not, GCRS scores may not be useful for distinguishing between performance uniformly at all levels of performance, and could require transformation prior to analysis.
- B. What is the reliability of GCRS in assessing individual consultations (with different numbers of raters per consultation)? One of two core questions: how consistently does GCRS perform in evaluating communication skills within a particular consultation, and how many raters are required to obtain performance estimates we are confident distinguish better from worse consultations?
- C. What is the reliability of GCRS in assessing individual doctors' performance across a number of





consultations (with different numbers of raters and consultations per doctor)? The second core question: how many consultations, and how many raters, do we need to evaluate a particular doctors' consultation skills such that we can differentiate them from their peers?

- D. Are some raters more generous than others in their assessments of consultations? Wide variation between the scores assigned by raters can lead to reduced reliability. Understanding whether systematic biases are present helps to inform whether to adjust reliability estimates for these or not.
- E. Does the order in which a consultation is rated affect the score? Psychological experiments have shown that the order in which information is presented can influence the way in which that information is processed.<sup>26</sup> Sequential order biases may present themselves either as an overall increase or decrease in scores throughout a judging period; or as observable effects of implicit comparisons being made between the previous and current items being judged.<sup>27 28</sup> Thus, a GCRS rater may use norm-based rather than criterion-based referencing when assigning scores as they proceed through the consultations being evaluated.

## METHODS

Trained GP raters watched video recordings of consultations between volunteer GPs and simulated patients and completed GCRS for each. We used videos from a previous study investigating the way in which GPs discussed taking statins to prevent cardiovascular disease with simulated patients trained to play one of two roles. The two roles differed in the extent of the actor's assertiveness in asking questions about proposed management. Both roles displayed sufficient cardiovascular risk to be eligible for statins according to current NICE recommendations. Actors were experienced in playing the role of simulated patients. They were provided with a detailed written role description, including notes on their intended style of response to questions. Actors rehearsed their roles before undertaking videotaped simulations with participant GPs. GPs (n=23) selected for recruitment to the original study varied in age, gender, length of time since qualification and nature of practice (location, size and involvement with dispensing or training). They were recruited from four primary care trusts across the East of England (Cambridge, Luton, Bedford and Peterborough). Each GP conducted two consultations in their practice (one with each simulated patient), furnished with the results of appropriate medical investigations for the simulated patient. The purpose of the consultation was, from the perspective of GP and patient, to discuss the possibility of starting statin medication. This generated a total of 46 recorded consultations. For this study, we excluded videos from two GPs: one had since become a trained GP GCRS

evaluator, while the videos for the second were damaged (see online supplementary appendix 1 figure S1). This left 42 videoed consultations for assessment. All GPs gave their written consent for the re-use of their videos.

## Global Consultation Rating Scale

The GCRS covers 12 domains from 'initiating the session' to 'closure' (see online supplementary appendix 3 for the full scale). Guidance is given within the text of the scale as to the nature of the skills that are assessed within each individual domain, which is given a score as follows: Not applicable (not scored)

0. Not done/poor
1. Adequate
2. Good

The use of a three-point scale, while narrow, (1) enables a clear focus on identifying the likely need for targeted training in that area and (2) reflects the need for a simple and easy-to-use scale suitable for use while observing a consultation. A total consultation score between 0 and 24 is obtained by summing the scores from the 12 domains. In the case where a domain is considered to be not applicable, scores are renormalised to be out of 24, for example, a score of 12 out of 22 would become a score of 13.1 ( $=12 \times 24 / 22$ ) out of 24 (NB: this was not required in this study).

## GP raters

We recruited six GP raters experienced in teaching and assessing communication skills using the Calgary-Cambridge consultation guide within the School of Clinical Medicine, University of Cambridge. All attended a 2 h training session on the use of GCRS with JS, which included a specially created training video of consultations for evaluation. In training, particular attention was paid to the differences between 'good', 'adequate' and 'poor' communication behaviours, guided by the criterion referenced norms established by the Calgary-Cambridge guide. The aim was to establish a shared understanding of expected standards of behaviour across each domain.<sup>29</sup> Following training, each evaluator rated 28 videos. These were randomly assigned and provided in a random order for rating. Randomisation was performed with maximum cross over between raters to allow study of possible order effects (see online supplementary appendix for further details).

GP raters were requested to complete evaluations within 1 month of collecting the videos and were paid for their time. On receipt of ratings some missing domain scores were noted (19 of 2184, 0.87%). The five raters who had missed scores watched the corresponding videos again and filled in the missing sections only. Double data entry was conducted (NE, GA) for all ratings. For the four scores (0.20%), in which there was inconsistency, the original score sheets were consulted to obtain the correct score.

## Open Access



### Statistical analysis

The overall aim of this work was to estimate the statistical reliability of GCRS as a tool to assess consultations or doctors. Statistical reliability is an index of how well better performance can be distinguished from worse performance, and estimates how much of the variation in scores is due to true variation in performance rather than to noise due to different raters rating the same consultation differently. A reliability of 1 indicates that all the variation in measured scores is due to true variation in performance, that is, that scores are perfectly reliable. A reliability of 0 indicates that all the variation in measured scores is due to statistical noise. Between these two extremes, a reliability of 0.8 is generally considered the minimum required for most applications.<sup>30</sup>

#### Does a fixed difference between scores in GCRS have the same meaning at all levels of performance?

One of the key assumptions made when calculating reliability is that measurement errors are independent of the true values. When this is not true a single reliability value cannot apply to all scores. Another way of thinking of this is that we require a fixed difference between two scores (eg, a two point difference) to have the same distinguishing quality across the full range of scores. For this to be true, the variability in raters' scores of the same consultation must be the same at all levels of performance. We checked this by plotting the SD of ratings for each consultation against the mean score for that consultation (a variation on the standard Bland-Altman plot, allowing for more than two ratings per consultation). We found that the variance was not the same across all mean scores, implying that, for raw scores, a fixed difference does not have the same meaning at all levels of performance. We, therefore, sought a transformation to stabilise the variance across all mean scores. The transformed data were used for all further analysis.

#### What is the reliability of GCRS for assessing single consultations?

Our experimental setup allowed us to distinguish between three different sources of variance:

1. differing performance between doctors
2. differing performance of the same doctor between consultations, and
3. differing evaluator scores of the same consultation

In order to calculate the crude reliability, we fitted a three-level linear regression model to reflect this, with no fixed effects and with random intercepts for consultation and doctor (ie, rating nested within consultation further nested within doctor). From such a model we can estimate the reliability that would be achieved for assessing single consultations with different numbers of raters (see online supplementary appendix). The same analysis was performed on the scores for each of the individual domain of GCRS.

#### What is the reliability of GCRS in assessing individual doctors' performance across a number of consultations?

Using the same approach, we can also estimate the reliability of GCRS for assessing doctor's performance using different numbers of raters to assess each doctor, and using different numbers of consultations per doctor (see online supplementary appendix).

#### Are some raters more generous than others in their assessments of consultations?

In order to establish whether there were systematic biases between the scores given by different raters, we augmented the model described above with fixed effects for raters. If present, biases between raters will increase the variation in scores, and in turn reduce the reliability of scores. The systematic biases between raters could be accounted for, and we estimated adjusted reliabilities after doing so.

#### Does the order in which a consultation is rated affect the score?

Finally, to investigate possible order effects we included the order of rating in the above model. To account for non-linear effects we used a restricted cubic spline with three knots. We excluded data from one evaluator in this analysis because they had not rated the consultations in the order requested.

CI's on all estimates were calculated using bias corrected bootstrapping with 1000 repetitions and resampling at the doctor level.

The approach outlined above falls somewhere between classical reliability studies in which only one source of variance is identified (eg, inter-rater reliability) and a generalisability theory approach.<sup>31</sup> However, due to the limited data available we feel the approach taken is the most appropriate, and further it allows a more nuanced investigation of order effects considering non-linear functions.

Statistical analysis was conducted using Stata V.11.2.

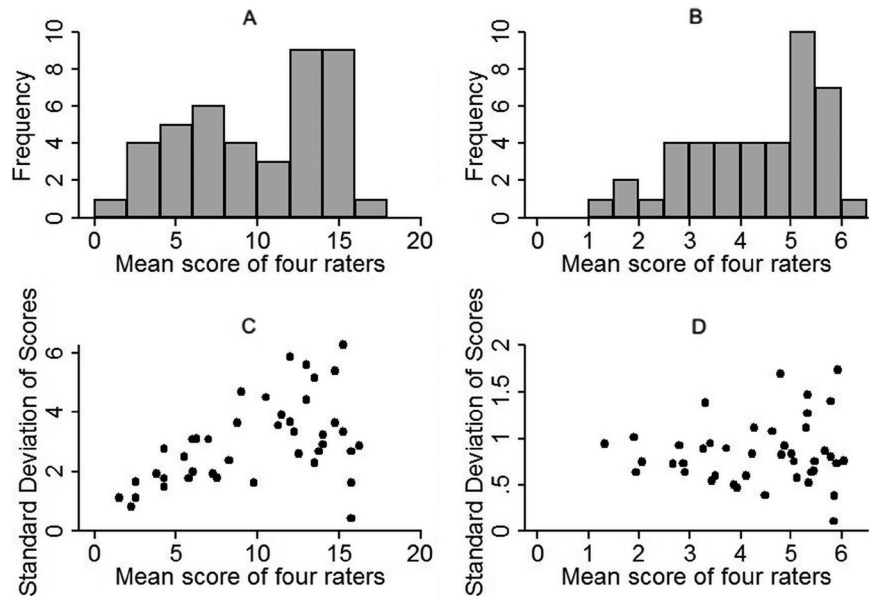
## RESULTS

The distribution of mean scores for the 42 consultations assessed (untransformed on a 0–24 scale) is shown in figure 1A. The highest mean consultation score was 16.25 of 24 and the lowest 1.5.

#### Does a fixed difference in GCRS have the same meaning at all levels of performance?

Figure 1C shows the Bland-Altman type plot for the untransformed data. There was a clear trend of increasing SD of scores for each consultation with increasing mean score. This implies that there was a higher degree of agreement between raters at low scores than at the moderate scores (10–14) which form the upper end of our data set. We found that a transformation based on the logit function performed reasonably well at stabilising the variance (see online supplementary appendix

**Figure 1** Histograms showing the distribution of mean consultation scores on the native (possible values 0 to 24) scale (A) and transformed (possible values 0 to 10) scale (B). Bland-Altman plot of consultation ratings shown on the native scale (C) and transformed scale (D).



for details and lookup table). The transformation has been constructed such that the transformed scores lie between 0 and 10. The distribution of the transformed scores is shown in figure 1B.

The resulting Bland-Altman plot of transformed data is shown in figure 1D in which there is little indication of a trend (note that the increase in spread of SDs is due to the possible values available and is not considered to be a major issue). All further results relate to the transformed data.

**What is the reliability of GCRS in assessing single consultations, and in assessing individual doctors' performance?**

The SDs for the three sources of variation estimated from the crude mixed model (with no adjustment for rater bias) are shown in table 1. The largest SD was that for between doctors, implying that this is where the largest variation is seen. The SD of scores of the same consultation by different raters was slightly smaller than that attributed to between doctors' performance. Finally, the estimates suggested that variation at the consultation level within individual doctors was essentially zero ( $SD=1.03 \times 10^{-9}$ ). This finding is likely to be a function of our dataset. We do not present any reliability estimates for rating doctors here, and outline the reasons for this

in the discussion. The reliability estimates for rating consultations for different numbers of raters are shown in table 2. In the crude model, the commonly used reliability thresholds of 0.7 (modest), 0.8 (acceptable) and 0.9 (excellent) were achieved using two, three and seven raters, respectively.<sup>30</sup> With four raters, as used in this study, we achieved a reliability of 0.85 (95% CI 0.61 to 0.88). Details of the distribution of scores and the reliabilities of individual domains are available in online supplementary appendix figure S2 and online supplementary appendix table S2. These indicate that four raters would be sufficient to provide a broad indication of domains where a doctor may have some performance issues.

**Are some raters more generous than others in their assessments of consultations?**

When we allowed for systematic bias between raters in our model we found that such bias was present (table 3). On an average, a difference of 1.65 (on the 0–10 scale for transformed data) was seen between the least and most generous raters. By adjusting for evaluator bias we increased reliability somewhat (table 2), and the number of raters needed to reach the 0.7, 0.8 and 0.9 thresholds became two, three and five, respectively.

**Table 1** SDs estimated for the three sources of variation from a crude model and one adjusting for systematic bias between raters

Source of variation	SD	
	Crude model	Model adjusted for evaluator bias
Between doctors	1.21 (0.87, 1.38)	1.18 (0.87, 1.33)
Within doctors and between consultations	$1.03 \times 10^{-9}$ ( $7.25 \times 10^{-13}$ , $1.95 \times 10^{-9}$ )	0.14 (0.00, 0.15)
Within consultations and between raters	1.03 (0.96, 1.16)	0.88 (0.82, 1.01)

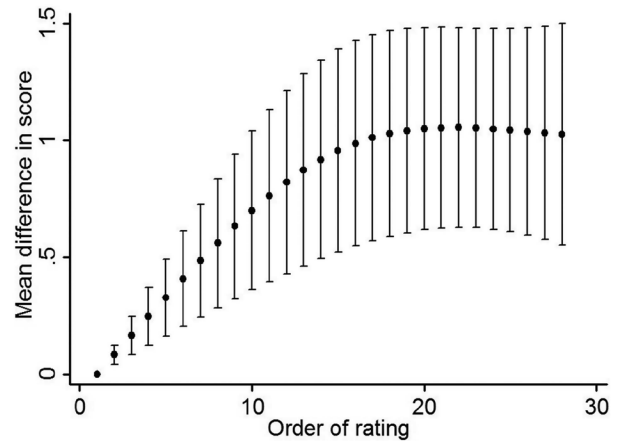
Open Access



**Table 2** Crude and adjusted reliability for evaluating consultations for different numbers of raters using GCRS (transformed 0–10 data)

Number of raters	Crude reliability* (95% CI)	Reliability adjusted for evaluator bias* (95% CI)
1	0.58 (0.28 to 0.66)	0.65 (0.36 to 0.71)
2	0.73 (0.44 to 0.79)	0.78 (0.53 to 0.83)
3	0.80 (0.54 to 0.85)	0.85 (0.63 to 0.88)
4	0.85 (0.61 to 0.88)	0.88 (0.69 to 0.91)
5	0.87 (0.66 to 0.91)	0.90 (0.74 to 0.93)
6	0.89 (0.70 to 0.92)	0.92 (0.77 to 0.94)
7	0.91 (0.73 to 0.93)	0.93 (0.80 to 0.95)
8	0.92 (0.76 to 0.94)	0.94 (0.82 to 0.95)
9	0.93 (0.78 to 0.95)	0.94 (0.84 to 0.96)
10	0.93 (0.80 to 0.95)	0.95 (0.85 to 0.96)

\*Calculated from the estimated SDs shown in table 1. GCRS, Global Consultation Rating Scale.



**Figure 2** The effect of order of rating on transformed scores compared with the first rating performed. Dots indicate point estimates and bars show 95% CIs.

**Does the order in which a consultation is rated affect the score?**

Finally, we found evidence of considerable order effects, with raters giving higher scores, on average, as they progressed through the rating of consultations (figure 2). It appears that raters’ scoring levelled out after performing around 15–20 ratings. Later consultations received, on average, scores more than one point higher on the 0–10 scale.

**DISCUSSION**

GCRS shows good reliability (>0.8) with three raters assessing each consultation, and modest reliability (>0.7) with two raters. Overall, consultations received low-to-moderate scores. This reflects previous findings with simulated patients, where it has been seen that participating doctors only attain about 40–60% of the guidelines or standards used for evaluation.<sup>32</sup> GCRS is designed to assess overall communication effectiveness of the entire doctor–patient consultation, encapsulating the quality of the interaction from the opening moments, through the gathering of information, provision of information, achieving shared understanding and shared decision-making, through to closure. It is a performance-based assessment (assessing what doctors

actually do in professional practice) rather than a competence-based assessment (assessing what doctors can do in controlled representations of professional practice).<sup>33</sup> It is additionally a criterion-referenced measure; GCRS training course highlights the importance of assessing performance against the ‘gold standard’ outlined in the Calgary-Cambridge guide.

While GCRS was devised as a global assessment, doctors may be interested in knowing their performance in particular domains in order to most efficiently target training. For individual GCRS domains, reliability was broadly acceptable with four raters. Low reliability for two particular domains—non-verbal communication and closure—may be attributable to small between-consultation variance rather than to raters disagreeing with each other on these areas. There are two possible explanations: either that raters find it difficult to distinguish differences in doctors’ behaviours on these items (reflecting inadequate training for raters in how to assess these domains, or challenges in capturing, eg, non-verbal behaviour) or that doctors perform comparably across consultations and compared with each other on these two domains, prompting raters to award consistently similar scores.

We found that a fixed difference between scores in GCRS did not have the same meaning at all levels of performance: untransformed scores (on a scale of 0 to 24) showed a higher degree of agreement between raters at low scores than at moderate scores. For this reason, analyses were performed on transformed scores. This has implications for the most suitable score to feedback to participants if, for example, GCRS is to be used in a training situation. Transformed scores may be intuitively more difficult for participants to understand, and we need to undertake further work on the acceptability of using transformed scores in assessments of an individual doctors’ performance, and how best to calculate and present transformed scores for doctors and trainers.

**Table 3** Estimated biases between raters using GCRS (transformed 0–10 data)

Evaluator	Mean difference (95% CI)
1	Reference
2	–0.25 (–0.57 to 0.13)
3	–0.68 (–1.20 to –0.18)
4	0.97 (0.66 to 1.33)
5	–0.25 (–0.76 to 0.31)
6	0.49 (0.04 to 0.96)

GCRS, Global Consultation Rating Scale.



While we found good reliability of GCRS in assessing the communication quality of individual consultations, comparison with existing instruments is difficult due to limited published psychometric data on assessing consultation (rather than doctor) quality. Interconsultation doctor reliability has been evaluated using the Four Habits Coding Scheme over 13 consultations (reliability of 0.72 with two raters),<sup>34</sup> and using the Liv-MAAS over nine consultations (reliability of 0.78 with three raters).<sup>35</sup> Evaluating the reliability of GCRS for assessing performance of individual doctors using different numbers of consultations will require more consultations per doctor, probably with greater subject variety, than we had in our dataset. We hope that further work on GCRS will enable us to estimate this in future.

We found consistent differences in scores assigned to consultations by the most and least generous raters. The Hawk/Dove phenomenon is well documented across a wide range of performance assessments, and can be addressed through training, through the use of more than one rater and through the use of post hoc statistical techniques.<sup>36</sup> All of these were employed in this study, and our finding of such variation highlights the importance of using pre-evaluation and postevaluation approaches in monitoring and acting upon differences between raters.<sup>37</sup>

We found evidence of considerable order effects. The use of multiple raters rating consultations in random order will tend to reduce order effects: sometimes a consultation will be rated early by an evaluator, and sometimes late; thus different orders for different raters average out. We have not been able to find other examples of the examination of this in GP consultation evaluation, but as previously stated, the influence of the sequential presentation of information on subsequent assessments of this information is a well-known phenomenon in the psychological literature.<sup>26</sup> Again, this is something which requires further work to assess how GCRS will perform in training situations.

The current study has a number of limitations. We included only a small number of GPs whose consultations had been recorded, derived from an earlier study, and only two similar scenarios per GP. These standardised scenarios do not reflect real-world consultations of variable nature and content, and we believe these are the reasons why we find little variation between consultations of the same doctor. We could not, therefore, assess how raters responded to different contexts: this is the focus of our next stage of work.

There are various sources of possible bias we did not examine due to sample size limitations. For example, contrast effect bias may be important in influencing rater behaviour, where, for example, viewing a good consultation after a series of poor consultations may lead to a substantial leap in scores assigned due to the contrast between them.

Feedback from raters showed that the assessment of consultations required significant concentration. Average consultation length was around 15 min: viewing each

consultation and completing the rating scale means each evaluation can take around 20 min.

## CONCLUSIONS

GCRS has good reliability (>0.8) for rating consultations if three raters are used. Systematic differences were observed between raters: adjusting for these further improves reliability of the scale. We are currently developing the scale further by assessing a large sample of consultations in a real-world setting. This will enable a more detailed examination of the ability of the scale to assess performance between consultations of the same doctor. Once further psychometric evaluation is completed, we envisage that GCRS has the capacity to provide a robust yet practical assessment tool for the evaluation of communication skills in everyday practice, linked to the Calgary-Cambridge training approach to target identified areas for improvement.

## Author affiliations

<sup>1</sup>Cambridge Centre for Health Services Research, University of Cambridge, Cambridge, UK

<sup>2</sup>University of Exeter Medical School, University of Exeter, Exeter, UK

<sup>3</sup>Primary Care Unit, University of Cambridge, Cambridge, UK

<sup>4</sup>School of Clinical Medicine, University of Cambridge, Cambridge, UK

**Acknowledgements** The authors would like to thank all participating general practitioners (GPs) and GP evaluators for their assistance with this work. The authors also thank the two reviewers whose thoughtful feedback greatly improved this article.

**Contributors** JBu designed the study, contributed to the analysis and interpretation of data and drafted the article. GA designed the study, undertook the analysis and contributed to the interpretation of data and drafting of the final version of the article. NE undertook data collection, and contributed to the analysis, the interpretation of data and drafting of the final version of the article. JC and MR designed the study, contributed to the interpretation of data and critically revised the article. JBe designed the study, supervised data collection and contributed to the interpretation of data and drafting of the final version of the article. JS designed the study, contributed to the interpretation of data, critically revised the article and devised the Global Consultation Rating Scale. All authors conceived the study and approved the final version of the article.

**Funding** This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

**Competing interests** None.

**Ethics approval** Bromley Research Ethics Committee (REC ref: 12/LO/0421).

**Provenance and peer review** Not commissioned; externally peer reviewed.

**Data sharing statement** No additional data are available.

**Open Access** This is an Open Access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 3.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/3.0/>

## REFERENCES

1. Silverman J, Kurtz S, Draper J. *Skills for communicating with patients*. 3rd edn. Oxford: Radcliffe, 2013.
2. Makoul G. The interplay between education and research about patient-provider communication. *Patient Educ Couns* 2003;50:79–84.
3. Simpson M, Buckman R, Stewart M, et al. Doctor-patient communication: the Toronto consensus statement. *BMJ* 1991;303:1385–7.

## Open Access



4. Stewart M, Brown JB, Boon H, *et al.* Evidence on patient-doctor communication. *Cancer Prev Control* 1999;3:25–30.
5. Suchman AL. Research on patient-clinician relationships: celebrating success and identifying the next scope of work. *J Gen Intern Med* 2003;18:677–8.
6. von Fragstein M, Silverman J, Cushing A, *et al.* UK consensus statement on the content of communication curricula in undergraduate medical education. *Med Educ* 2008;42:1100–7.
7. Association of American Medical Colleges. Report 3: Contemporary Issues in Medicine: Communication in Medicine. Washington, DC: AAMC, 1999.
8. BMA. *Communication skills education for doctors: a discussion document*. London: BMJ, 2003.
9. Cowan DH, Laidlaw JC. Improvement of teaching and assessment of doctor-patient communication in Canadian medical schools. *J Cancer Educ* 1993;8:109–17.
10. Department of Health. *Medical schools: delivering the doctors of the future*. London: Department of Health, 2004.
11. General Medical Council. *Tomorrow's doctors: recommendations on undergraduate medical education*. London: GMC, 2009.
12. The Royal College of Physicians and Surgeons of Canada. Canadian Medical Education Directions for Specialists 2000 Project: Skills for the New Millennium: Report of the Societal Needs Working Group, 1996.
13. Workshop Planning Committee. Consensus statement from the workshop on teaching and assessment of communication in Canadian medical schools. *CMAJ* 1992;147:1149–50.
14. Boon H, Stewart M. Patient-physician communication assessment instruments: 1986 to 1996 in review. *Patient Educ Couns* 1998;35:161–76.
15. Schirmer JM, Mauksch L, Lang F, *et al.* Assessing communication competence: a review of current tools. *Fam Med* 2005;37:184–92.
16. NICE. Patient experience in adult NHS services. Quality Standards, QS15. 2012.
17. van der Vleuten CP, Schuwirth LW, Scheele F, *et al.* The assessment of professional competence: building blocks for theory development. *Best Pract Res Clin Obstet Gynaecol* 2010;24:703–19.
18. van Thiel J, Ram P, van Dalen J. *MAAS-Global manual*. Maastricht, Netherlands: University of Maastricht, 2000.
19. Kurtz S, Silverman J, Benson J, *et al.* Marrying content and process in clinical method teaching: enhancing the Calgary-Cambridge guides. *Acad Med* 2003;78:802–9.
20. Kurtz SM, Silverman J, Draper J. *Teaching and learning communication skills in medicine*. 2nd edn. Oxford, San Francisco: Radcliffe Medical, 2005.
21. Kurtz SM, Silverman JD. The Calgary-Cambridge referenced observation guides: an aid to defining the curriculum and organizing the teaching in communication training programmes. *Med Educ* 1996;30:83–9.
22. Gillard S, Benson J, Silverman J. Teaching and assessment of explanation and planning in medical schools in the United Kingdom: cross sectional questionnaire survey. *Med Teach* 2009;31:328–31.
23. Howells RJ, Davies HA, Silverman JD, *et al.* Assessment of doctors' consultation skills in the paediatric setting: the Paediatric Consultation Assessment Tool. *Arch Dis Child* 2010;95:323–9.
24. Radford A, Stockley P, Silverman J, *et al.* Development, teaching, and evaluation of a consultation structure model for use in veterinary education. *J Vet Med Educ* 2006;33:38–44.
25. Silverman J, Archer J, Gillard S, *et al.* Initial evaluation of EPSCALE, a rating scale that assesses the process of explanation and planning in the medical interview. *Patient Educ Couns* 2011;82:89–93.
26. Mussweiler T. Comparison processes in social judgments: mechanisms and consequences. *Psychol Rev* 2003;110:472–89.
27. Page L, Page K. Last shall be first: a field study of biases in sequential performance evaluation on the idol series. *J Econ Behav Organ* 2010;73:186.
28. Rothoff KW. (Not Finding a) Sequential Order Bias in Elite Level Gymnastics, 2013.
29. Williams RG, Klamen DA, McGaghie WC. Cognitive, social and environmental sources of bias in clinical performance ratings. *Teach Learn Med* 2003;15:270–92.
30. Postgraduate Medical Education and Training Board. *Developing and maintaining an assessment system—a PMETB guide to good practice*. London: GMC, 2007.
31. Brennan RL. Generalizability Theory. *Educ Meas Issues Pract* 1992;11:27–34.
32. Rethans JJ, Sturmans F, Drop R, *et al.* Assessment of the performance of general practitioners by the use of standardized (simulated) patients. *Br J Gen Pract* 1991;41:97–9.
33. Rethans JJ, Norcini JJ, Baron-Maldonado M, *et al.* The relationship between competence and performance: implications for assessing practice performance. *Med Educ* 2002;36:901–9.
34. Krupat E, Frankel R, Stein T, *et al.* The Four Habits Coding Scheme: validation of an instrument to assess clinicians' communication behavior. *Patient Educ Couns* 2006;62:38–45.
35. Enzer I, Robinson J, Pearson M, *et al.* A reliability study of an instrument for measuring general practitioner consultation skills: the LIV-MAAS scale. *Int J Qual Health Care* 2003;15:407–12.
36. Harasym PH, Woloschuk W, Cuning L. Undesired variance due to examiner stringency/leniency effect in communication skill scores assessed in OSCEs. *Adv Health Sci Educ Theory Pract* 2008;13:617–32.
37. Bartman I, Roy M, Smee S. *Catching the hawks and doves: a method for identifying extreme examiners on objective structured clinical examinations*. Ottawa: Medical Council of Canada, 2011.

BMJ Open

## Assessing communication quality of consultations in primary care: initial reliability of the Global Consultation Rating Scale, based on the Calgary-Cambridge Guide to the Medical Interview

Jenni Burt, Gary Abel, Natasha Elmore, John Campbell, Martin Roland, John Benson and Jonathan Silverman

BMJ Open 2014 4:

doi: 10.1136/bmjopen-2013-004339

Updated information and services can be found at:  
<http://bmjopen.bmj.com/content/4/3/e004339>

*These include:*

### Supplementary Material

Supplementary material can be found at:  
<http://bmjopen.bmj.com/content/suppl/2014/03/07/bmjopen-2013-004339.DC1>

### References

This article cites 25 articles, 3 of which you can access for free at:  
<http://bmjopen.bmj.com/content/4/3/e004339#ref-list-1>

### Open Access

This is an Open Access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 3.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/3.0/>

### Email alerting service

Receive free email alerts when new articles cite this article. Sign up in the box at the top right corner of the online article.

### Topic Collections

Articles on similar topics can be found in the following collections

[General practice / Family practice](#) (728)  
[Health services research](#) (1618)  
[Medical education and training](#) (313)  
[Patient-centred medicine](#) (493)

### Notes

To request permissions go to:  
<http://group.bmj.com/group/rights-licensing/permissions>

To order reprints go to:  
<http://journals.bmj.com/cgi/reprintform>

To subscribe to BMJ go to:  
<http://group.bmj.com/subscribe/>

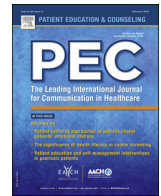






Contents lists available at ScienceDirect

## Patient Education and Counseling

journal homepage: [www.elsevier.com/locate/pateducou](http://www.elsevier.com/locate/pateducou)

# The reliability of a modified Kalamazoo Consensus Statement Checklist for assessing the communication skills of multidisciplinary clinicians in the simulated environment

Eleanor B. Peterson<sup>a,\*</sup>, Aaron W. Calhoun<sup>a</sup>, Elizabeth A. Rider<sup>b,c</sup>

<sup>a</sup> Department of Pediatrics, University of Louisville School of Medicine, Louisville, USA

<sup>b</sup> Department of Pediatrics, Harvard Medical School, Boston, USA

<sup>c</sup> Institute for Professionalism and Ethical Practice, Boston Children's Hospital, Boston, USA

## ARTICLE INFO

## Keywords:

Communication skills  
Communication assessment  
Communication competency  
Assessment tools  
Communication education  
Simulation  
Kalamazoo Consensus Statement

## ABSTRACT

**Objective:** With increased recognition of the importance of sound communication skills and communication skills education, reliable assessment tools are essential. This study reports on the psychometric properties of an assessment tool based on the Kalamazoo Consensus Statement Essential Elements Communication Checklist.

**Methods:** The Gap-Kalamazoo Communication Skills Assessment Form (GKCSAF), a modified version of an existing communication skills assessment tool, the Kalamazoo Essential Elements Communication Checklist-Adapted, was used to assess learners in a multidisciplinary, simulation-based communication skills educational program using multiple raters. 118 simulated conversations were available for analysis. Internal consistency and inter-rater reliability were determined by calculating a Cronbach's alpha score and intra-class correlation coefficients (ICC), respectively.

**Results:** The GKCSAF demonstrated high internal consistency with a Cronbach's alpha score of 0.844 (faculty raters) and 0.880 (peer observer raters), and high inter-rater reliability with an ICC of 0.830 (faculty raters) and 0.89 (peer observer raters).

**Conclusion:** The Gap-Kalamazoo Communication Skills Assessment Form is a reliable method of assessing the communication skills of multidisciplinary learners using multi-rater methods within the learning environment.

**Practice implications:** The Gap-Kalamazoo Communication Skills Assessment Form can be used by educational programs that wish to implement a reliable assessment and feedback system for a variety of learners.

© 2014 Elsevier Ireland Ltd. All rights reserved.

## 1. Introduction

Sound interpersonal and communication skills are critical to the provision of quality healthcare. Effective communication with patients, families and physicians has been shown to enhance coping, mitigate grief, improve adherence to treatment, alter perceptions of care and reduce medical errors and litigation [1–6]. The National Board of Medical Examiners (NBME), Association of American Medical Colleges (AAMC), Institute of Medicine, and Accreditation Council on Graduate Medical Education (ACGME)

have suitably placed a priority on the teaching and assessment of interpersonal and communication skills in undergraduate and graduate medical education [6–10]. Consequently, in the United States, achieving competency in communication has become a factor for promotion, graduation and licensure [7–9]. Teaching and assessing communication skills remains a complex and historically under-represented component of medical education [10,11]. Fortunately, increased awareness of the importance of communication and relationships in healthcare, and more emphasis on the importance of communication skills training in medical education, has led to an ever growing body of literature regarding the teaching and assessing of communication skills available to educators [10,12–17]. This article reports on the psychometric properties of an assessment tool which was derived from The Kalamazoo Consensus Statement [18], an exemplar in the field of medical communication research, education and assessment.

\* Corresponding author at: Division of Pediatric Critical Care Medicine, University of Louisville, 571 S. Floyd St. STE 332, Louisville, KY 40202, USA.  
Tel.: +1 502 852 3720; fax: +1 502 852 3998.

E-mail address: [ebpete01@louisville.edu](mailto:ebpete01@louisville.edu) (E.B. Peterson).

<http://dx.doi.org/10.1016/j.pec.2014.07.013>

0738-3991/© 2014 Elsevier Ireland Ltd. All rights reserved.

The Kalamazoo Consensus Statement was developed in 1999 by 21 North American leaders from the fields of medical education and communication [18]. Their intent was to delineate a list of elements essential to physician–patient communication for the purpose of facilitating the development, implementation and evaluation of communication curricula [18]. The result was a list of seven “essential elements,” or communication tasks, that define effective physician–patient communication. This consensus statement has since served as a framework for the development of numerous educational programs [10,15,19–23].

In subsequent years the same group met to create the Kalamazoo Essential Elements Communication Checklist (KEECC), an assessment tool for the purpose of rating learners’ competency across the seven essential elements of the Kalamazoo Consensus Statement [10]. The essential elements, or competencies (Builds the Relationship, Opens the Discussion, Gathers Information, Understands the Patient’s and Family’s Perspective, Shares Information, Reaches Agreement, and Provides Closure), are rated using a categorical 4-option scale across 24 sub-competencies. This tool has applicability to all levels of training and various settings [10]. Two additional iterations of the KEECC, the Kalamazoo Essential Elements Communication Checklist-Adapted (KEECC-A) [10] and the Gap-Kalamazoo Communication Skills Assessment Form (GKCSAF) [10] have been published. The GKCSAF has been adapted for multi-rater use, a powerful method for assessing communication skills that enhances self-insight [11,24]. In combination, these three tools have been used in undergraduate and graduate medical education and healthcare education programs nationally and internationally [10,11,24,25].

Simulation, either through the use of role-play or standardized patients, is an increasingly common and effective educational modality for use in communication skills education [3,13,15]. With the growth of simulation-based training comes the need for reliable assessment tools for use in the simulated environment. While psychometric data exists regarding the KEECC [9], KEECC-A [25] and GKCSAF [11], to our knowledge no study has evaluated inter-rater reliability among the communication elements of the Kalamazoo Tools, nor has there been a psychometric analysis for a multidisciplinary field of learners in the simulated environment.

The objective of this paper, therefore, is to build on the work of previous studies, by reporting the internal consistency and inter-rater reliability of the GKCSAF when used for multi-rater assessment of multi-disciplinary learners in a simulation-based communication skills education program.

## 2. Methods

### 2.1. Tool development

Three assessment tools based on the Kalamazoo Consensus Statement have been published [10]. The original tool, the KEECC, rated learners categorically (i.e., done well, needs improvement, not done, not applicable) on seven competencies and 24 sub-competencies [10,18]. Rider and colleagues at Harvard Medical School adapted the KEECC by adding a 5-point Likert scale (1 = poor to 5 = excellent) [10]. This adapted version, the KEECC-A, allows for evaluation of the seven Kalamazoo Essential Elements on a global ratings scale and the 24 sub-competencies function as a rubric for this checklist [10]. The Likert scale can also be used to rate each competency and sub-competency. Calhoun, Rider and colleagues modified the KEECC-A to include two more communications elements, Demonstrates Empathy and Communications Accurate Information, creating the GKCSAF [10,24]. This latest Kalamazoo Consensus Statement instrument was also modified for use by multiple raters (modeled after 360° assessment tools) and includes a section for gap analysis [24]. Gap analysis is a novel application of multi-rater feedback that consists of comparing rating scores from different groups of raters, for example faculty or peer observers, with self-score of the participant or participant team [11]. This comparison of scores has been shown to enhance learner self-insight [11]. The GKCSAF contains Likert-scale, forced-choice, and free-text fields, enabling it to provide absolute and relative scores for each aspect of communication and specific comments regarding strengths and areas needing improvement. A similar version of the instrument was created for simulated patients/families using language that was assessed by Microsoft Word as suitable for a reader at the United States 6th grade reading level, which roughly translates to a reading level appropriate for a 10–12 year old (Table 1).

**Table 1**  
Description of the Kalamazoo Consensus Statement assessment instruments.<sup>a</sup>

Kalamazoo instrument	Data type	Instrument description	Psychometric studies
Kalamazoo Essential Elements Communication Checklist	Categorical ratings: Done well Needs improvement Not done Not applicable	Includes the Kalamazoo Consensus Statement 7 core communication competencies and 24 sub-competencies	Schirmer JM, Mauksch L, Lang F, Marvel MK, Zoppi K, Epstein RM, Brock D, Pryzbylski M. Assessing communication competence: a review of current tools. <i>Fam Med</i> 2005;37:184–92
Kalamazoo Essential Elements Communication Checklist-Adapted <sup>b</sup>	5-point Likert scale: 1 = poor to 5 = excellent	Global ratings on the 7 core competencies Second version with ratings on 7 core and 24 sub-competencies	Joyce BL, Steenbergh T, Scher E. Use of the Kalamazoo Essential Elements Communication Checklist (Adapted) in an institutional interpersonal and communication skills curriculum. <i>J Grad Med Educ</i> 2010;2:165–9
Gap-Kalamazoo Communication Skills Assessment Form	Likert-scales, forced-choice and free-text fields to provide absolute and relative scores for each competency; and specific comments on strengths and areas needing improvement	Global ratings on the 7 core competencies and 2 additional competencies: Demonstrates Empathy, and Communicates Accurate Information Versions: • Clinician/Faculty (also used by Peer Facilitators) • Self-assessment • Patient/Family (6th grade reading level)	Calhoun AW, Rider EA, Meyer EC, Lamiani G, Truog RD. Assessment of communication skills and self-appraisal in the simulated environment: feasibility of multi-rater feedback with gap analysis. <i>Simul Healthc</i> 2009;4:22–9

<sup>a</sup> The instruments are published in: Rider EA, Nawotniak RH. A practical guide to teaching and assessing the ACGME core competencies, 2nd ed. Marblehead, MA: HCPro Inc.; 2010.

<sup>b</sup> To preserve research integrity, we recommend using the authentic versions of the Kalamazoo instruments. The version of the GKCSAF used in this study is included as an Appendix with this article.

2.2. Tool implementation

The GKCSAF has been used for four years to assess communication competencies of participants in the Program for the Approach the Complex Encounters (PACE). PACE is a simulation-based curriculum at the University of Louisville School of Medicine developed to enhance the skills of multidisciplinary healthcare professionals in navigating challenging communication situations [15]. PACE relies on the Kalamazoo Consensus Statement competencies as a framework for communication skills education. During a PACE session, after a brief discussion of communication strategies, resident/nurse (or rarely resident/chaplain) clinician teams embark on a simulated conversation with a patient family portrayed by standardized patients (SP). Clinician teams always consist of one physician and one allied health professional, however, determination of which participants simulate which conversation are left up to the participants themselves. PACE sessions are typically attended by two to three faculty members who help guide post-simulation feedback and discussion. Each simulated conversation is rated by PACE faculty members, standardized patients, peer-observers and the participants themselves in a 360° fashion using the GKCSAF.

2.3. Tool training

Faculty, peer observers, standardized patients and participants were not trained specifically on the use of the GKCSAF prior to this study. This was done intentionally as many assessment tools have been validated by studies in which raters were formally trained on the use of the tool in question. Extensive training, however, is not always possible given the issues of lack of free time that chronically plague busy clinical faculty, residents with duty-hours restrictions and hospital staff carrying full-time work schedules. Thus, we wanted to assess the psychometric properties of the GKCSAF in an environment that most closely reflects how we anticipate this tool will be used.

2.4. Scoring

The GKCSAF is composed of nine essential communication elements rated on a 5-point Likert Scale (1 = Poor, 2 = Fair, 3 = Good, 4 = Very good, 5 = Excellent). In the PACE sessions, four versions of the Gap-Kalamazoo Tool are generated for each simulated conversation, generated by the four groups of raters: a self-assessment, faculty assessment, peer observer assessment and standardized patient (SP) assessment. Competency-specific overall scores are calculated by averaging individual scores for each competency. Learners are provided a written feedback form following their PACE session, detailing cumulative assessment scores from all raters across all communication elements.

2.5. Statistical analysis

For the purpose of statistical analysis, faculty and peer observer ratings were used. The unit of analysis was the clinician team. To assess internal consistency, a Cronbach's alpha score was calculated for simulated conversations to provide an overall alpha for faculty and peer ratings, respectively. These groups were chosen due to the relatively consistent number of raters across all sessions, allowing for more consistent statistical assessment. In addition to this, we calculated a separate Cronbach's alpha for each faculty rater across all sessions and averaged these values to generate an additional Cronbach's alpha. This was done to assure the accuracy of the initial score, given the possibility of intra-session correlations in rating that could artificially elevate the statistic. As the same peer observers did not rate every conversation within a PACE session, we were unable to perform a separate Cronbach's alpha for peer observers in the same manner. Inter-rater reliability was analyzed using intra-class correlation coefficients (two-way random, consistency measures) (ICC). This statistic was calculated for all simulated conversations in which 3 faculty members or peer observers provided ratings. ICC's were calculated for each communication element and for the overall average score of each tool. Cronbach's alpha scores and ICCs are reported for faculty and peer observers separately. Statistics were calculated using SPSS ver 21.

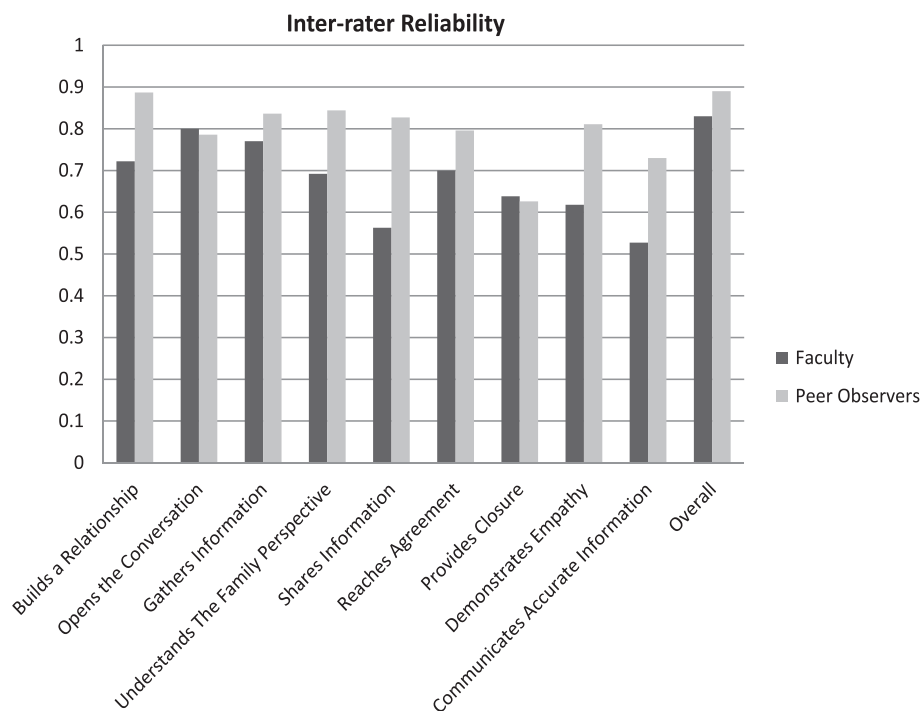


Fig. 1. Intra-class correlation coefficient scores for communication domains of the Gap-Kalamazoo Communication Skills Assessment Form for faculty and peer observers.

### 3. Results

#### 3.1. Subjects

Since its inception in 2009, PACE has simulated 118 conversations for 173 participants. Participants include medical residents 2–4 years after receiving their M.D. degree (categorical pediatric and combined pediatric/internal medicine residents,  $N = 108$ ), nurses (bedside nurses, nursing administrators and nursing students,  $N = 63$ ) and hospital chaplains ( $N = 2$ ). Of those conversations, 44 had 3 faculty raters and 25 had 3 peer observer raters rendering them eligible for analysis by ICC.

#### 3.2. Tool internal consistency

There were 118 faculty rated conversations and 72 peer observer rated conversations from which to calculate a Cronbach's alpha. The Gap-Kalamazoo tool demonstrated an overall Cronbach's Alpha of 0.844 for faculty and 0.880 for peer observers. Faculty rater-specific Cronbach's alpha scores were 0.837 ( $N = 106$  of conversations used for calculation), 0.818 ( $N = 104$ ) and 0.90 ( $N = 61$ ), respectively. The overall average of the faculty rater sub-alpha scores was 0.852.

#### 3.3. Tool reliability

Faculty ICC scores ranged from 0.527 to 0.800 for each domain of communication. Among faculty, the lowest ICC's were noted for the elements of Communicates Accurate Information and Shares Information (0.527 and 0.563, respectively), while the elements with the highest ICC's were Opens the Discussion and Gathers Information (0.800 and 0.770, respectively). The overall ICC was 0.830.

Peer observer ICC scores ranged from 0.626 to 0.887 for each domain of communication. Among peer observers, only one communication domain, Provides Closure, scored  $< 0.7$ . Five elements had ICCs  $> 0.8$ . The overall ICC was 0.890 (Fig. 1).

### 4. Discussion and conclusion

#### 4.1. Discussion

The three published assessment tools developed from the Kalamazoo Consensus Statement (Table 1) are valuable resources for communication skills education [10]. Psychometric analysis of these tools strengthens their applicability across a variety of learning environments. A 2005 analysis of the psychometric properties of the KEECC demonstrated a Cronbach's alpha of 0.88 [9]. Previously published psychometric data analysis of the KEECC-A reported good internal consistency for a cumulative communications rating when used to evaluate medical residents during a simulated clinical encounter [25]. Good internal consistency has been previously published for the Gap-Kalamazoo Tool but was based on a small sample size of only seven learners [11]. To our knowledge no study has evaluated inter-rater reliability among the communication elements of the Kalamazoo Tools, nor has there been a psychometric analysis for a multidisciplinary field of learners. This latter consideration is of special importance, as the GKCSAF was specifically designed for a multidisciplinary, multi-rater assessment.

We designed the PACE curriculum to include multi-rater feedback. Multi-rater feedback is a holistic approach to evaluation particularly suited to communication skills training that places the learner at the center of multiple relationships including peers, patients/families and faculty [11]. By encompassing the feedback of patients/families and multi-disciplinary clinician peers, real-world validity is enhanced and more comprehensive feedback can be generated for learners [11]. Likewise, the GKCSAF is designed for multi-rater use, therefore, we chose to assess the internal

consistency and inter-rater reliability for both faculty and peer observer ratings. However, we did not generate a combined ICC value that included both groups. This was done deliberately because we expect that perceptions of skill will differ among the groups of raters. This is due to the nature of multi-rater feedback, which postulates differences in the perspective and hence content of feedback provided between disciplinary groups. If this were not case, multi-rater feedback would be unnecessary as all perceptions of skill will be the same. In support of this view, participants receive written feedback that encompasses the ratings and comments from all groups of raters, and a global general score is not provided.

#### 4.1.1. Internal consistency

The GKCSAF demonstrates good internal consistency with a Cronbach's alpha of 0.844 and 0.880 for faculty ( $N = 118$ ) and peer observer ratings ( $N = 75$ ), respectively. These scores are consistent with previously published data for earlier versions of the tool. Knowing that calculating an overall alpha carried the risk of bias, due to a potential of clustering scores for a given conversation, we calculated a sub-alpha score per randomly assigned faculty rater to ensure the overall alpha was not falsely elevated. Finding an average sub-alpha similar to the overall alpha lends credibility to the internal consistency and lessens the concern about potential bias within a conversation. As mentioned above, we were unable to perform such a sub-analysis for peer observer ratings, as peer observers changed with every given conversation and hence could not be separated in the same manner as faculty. The strength of this study is the number of conversations analyzed, at 118 for faculty, and 72 for peer observer, which is much higher than previously reported psychometric data regarding the Gap-Kalamazoo tool.

#### 4.1.2. Inter-rater reliability

For the purposes of assessing inter-rater reliability, we chose to use conversations that had 3 raters for statistical reasons. This limited our data set to 44 faculty-rated conversations and 25 peer-observer-rated conversations.

The ICC scores for faculty ratings across the nine communication elements assessed in the Gap-Kalamazoo tool ranged from 0.527 to 0.800 but demonstrated high inter-rater reliability with an overall ICC 0.830. Specifically, Communicates Accurate Information and Shares Information had relatively low ICCs of 0.527 and 0.563, respectively, Demonstrates Empathy, Provides Closure and Understands the Patient's and Family's Perspective had acceptable ICCs between 0.6 and 0.7, while the remaining four elements of Builds a Relationship, Opens the Discussion, Gathers Information, Reaches Agreement had good ICCs of  $\geq 0.7$ . It was of interest to us that certain elements of the communication checklist demonstrated higher inter-rater reliability than others. Particularly, the elements of Communicates Accurate Information and Shares Information showed the poorest inter-rater reliability. While we feel that Communicates Accurate Information and Shares Information are two important and distinct communication tasks, the lower ICCs for these two elements could represent a higher subjectivity for these elements or even a perceived redundancy or confusion regarding the essence of these tasks. This could represent a need for clarifying language within the evaluative rubric as to the true conceptual content of these elements. Of note, the overall average scores of the communication encounter demonstrated higher reliability among raters than any individual domain, conceivably indicating that general impressions of overall communication skill are preserved with the Gap-Kalamazoo tool. Hence, even if individual elements lacked agreement, there was consensus regarding the clinician teams' overall performance during the simulated conversation.

The ICC scores for peer observer ratings across the nine communication elements assessed in the Gap-Kalamazoo tool ranged from 0.626 to 0.887 with an overall inter-rater reliability of



0.890. One communication domain, Provides Closure, demonstrated acceptable inter-rater reliability with an ICC of 0.626. Three domains, Opens the Discussion, Reaches Agreement and Communicates Accurate Information had ICCs in the good range of  $\geq 0.7$  while the remaining five elements displayed excellent inter-rater reliability with ICCs  $\geq 0.8$ . Parallel to faculty ratings, the overall rating of the communication encounter demonstrated a higher ICC than any individual domain at 0.890, again suggesting that overall ratings of skill may be preserved among raters even if perceptions of skill for individual communication tasks differ.

In general, higher inter-rater reliability was demonstrated among peer observers than faculty raters. We can think of several reasons why this might exist. First, it is possible that peer observer scores tend to cluster in one direction. We also wondered whether peer observers might cluster scores in a more generally favorable manner. To test the theory that peer observers might perceive overall communication skills as better than faculty raters, we compared the average ratings among peer observer and faculty raters and found they did not differ significantly (3.93 vs. 3.98,  $p$ -value 0.54 by Mann-Whitney  $U$ ). Second, peer observers were unfamiliar with the GKCSAF prior to completing the assessment tool and this could have led to differences in perception of the communication elements, as opposed to faculty who had prior experience with the tool.

#### 4.1.3. Limitations

While we feel this study shows the Gap-Kalamazoo tool a useful and reliable instrument for assessing learners participating in a communication skills curriculum, there are several limitations that bear discussion.

First, generalizability theory is an alternative method to assess the reliability of assessment tools and is felt to be superior to more traditional means of determining reliability as it can detect multiple sources of error [26]. A generalizability study, had we been able to perform it, would have yielded more information than our current approach. Unfortunately, the structure of our dataset rendered a generalizability study impossible.

Second, this tool is intended for use by multiple raters but we were unable to analyze reliability within all groups of raters. Although ratings are obtained from the four groups, faculty, peer observers, standardized patients and self/participants, we only had sufficient data to analyze faculty and peer observers. To calculate inter-rater reliability we chose to use conversations that were rated by three individuals. Unfortunately, we had no conversations in which more than two standardized patients or participants ("self-scores") rated a conversation so we were unable to assess the psychometric properties within these groups of raters. A study in which reliability was analyzed with all groups of raters would certainly be a stronger study but we did not have the data to perform such an analysis. We still feel the tool demonstrated reasonable reliability within the two groups of raters mentioned above, supporting its' use in a multi-rater fashion. Additionally, due to the variability in peer observer ratings, we were unable to perform a "sub-alpha" to confirm the accuracy of the Cronbach's alpha score for peer observer ratings as we were for faculty ratings. It is possible, then, that the reported Cronbach's alpha score of 0.880 for peer observer ratings is falsely elevated.

Third, other than the theories briefly mentioned above, it is unclear to us why some elements of the tool performed well while others showed generally poor inter-rater reliability, particularly among faculty members. Unfortunately, we have not had the opportunity to discuss the use of the tool among faculty raters, as doing so may elucidate why it was easier to reach agreement among certain elements than others.

Last, and most important, a significant limitation of this study is the fact that post-simulation debriefing occurred prior to

completion of the assessment tool. Results regarding inter-rater reliability should be viewed with caution, knowing that post-simulation discussion likely led to some normalization of the data. The order of debriefing in relation to tool completion was a conscious decision from the outset of curriculum development in an effort to create and preserve a learning atmosphere. Simulating complicated, emotionally charged conversations while being viewed by others is a vulnerable position for learners, and we strive to create a safe learning environment that promotes an atmosphere of self-discovery. The GKCSAF takes approximately 10–15 min to complete, time we felt, would create a disruption of the learning environment and place an unwanted emphasis on evaluation and assessment for our learners. We designed the curriculum not for the purposes of validating the assessment tool but with the goal of creating an effective communication skills curriculum. In doing so, we placed a higher priority on the learning environment than the rigors of the study presented here. We realize that this was a judgment call and whether or not completing the assessment prior to debriefing would affect learners as we purport remains to be seen. We do contend that holding the debriefing prior to completing the assessment tool led to less normalization of the data than one might think due to the nature of the debriefing session. The debriefing component of this curriculum relies heavily on participant self-directed learning and discovery using recorded simulations for playback and review. Feedback and discussion is directed using frame-by-frame analysis of the conversations, led by the self-insights of the participants and observers. Participants and peer observers are unfamiliar with the GKCSAF. There is no mention of, or reference to, the Kalamazoo Essential Elements framework during the discussion. To summarize, the possibility exists that influence on raters from the debriefing session led to inflation of the inter-rater reliability of the GKCSAF. Given that the environment in which we use the tool is similar to how it will likely be used in practice, we still feel the GKCSAF is a useful tool, viewed within the constraints mentioned above.

#### 4.2. Conclusions

The importance of developing sound communication skills among healthcare professionals and the greater emphasis on communication skills education in undergraduate and graduate medical education makes reliable assessment methodologies essential. The Gap-Kalamazoo Communication Skills Assessment Form is linked to an accepted theoretic framework, builds on studies utilizing earlier versions of the Kalamazoo assessment tools, and has been demonstrated to have good psychometric reliability, and therefore begins to meet this important need. Further research exploring the inter-rater reliability among all groups of raters, completion of the assessment tool prior to debriefing, and use of generalizability theory would further define the usefulness of this tool.

#### 4.3. Practice implications

Despite the limitations mentioned above, the Gap-Kalamazoo Communication Skills Assessment Form can be used by educational programs that wish to implement a reliable assessment and feedback system for a variety of multidisciplinary learners.

*Note:* A different tool with different contents, but also titled the Kalamazoo Essential Elements Communication Checklist-Adapted, is found on the Internet. To preserve research integrity, we recommend using the authentic, copyrighted, validated version. Questions regarding use of the GKCSAF tool can be directed to [aaron.calhoun@louisville.edu](mailto:aaron.calhoun@louisville.edu) or [elizabeth\\_rider@hms.harvard.edu](mailto:elizabeth_rider@hms.harvard.edu) (member, Kalamazoo Consensus Statement group).

Appendix

Gap-Kalamazoo Communication Skills Assessment Form\* – Faculty/Peer Assessment

Date:	Your Name:	Your Title:
-------	------------	-------------

Title of Case:	Title of Conversation:
----------------	------------------------

How well did the participant(s) do the following (please select one):

	1 Poor	2 Fair	3 Good	4 Very Good	5 Excellent
<b>A: Builds a relationship (includes the following):</b> <ul style="list-style-type: none"> <li>• Greets and shows interest in the patient’s family</li> <li>• Uses words that show care and concern throughout the interview</li> <li>• Uses tone, pace, eye contact, and posture that show care and concern</li> <li>• Responds explicitly to patient and family statements about ideas and feelings</li> </ul>					
<b>B: Opens the discussion (includes the following):</b> <ul style="list-style-type: none"> <li>• Allows patient and family to complete opening statement without interruption</li> <li>• Asks “is there anything else?” to elicit full set of concerns</li> <li>• Explains and/or negotiates an agenda for the visit</li> </ul>					
<b>C: Gathers information (includes the following):</b> <ul style="list-style-type: none"> <li>• Addresses patient and family statements using open-ended questions</li> <li>• Clarifies details as necessary with more specific or “yes/no” questions</li> <li>• Summarizes and gives family opportunity to correct or add information</li> <li>• Transitions effectively to additional questions</li> </ul>					
<b>D: Understands the patient’s and families perspective (includes the following):</b> <ul style="list-style-type: none"> <li>• Asks about life events, circumstances, other people that might affect health</li> <li>• Elicits patient’s and family’s beliefs, concerns, and expectations about illness and treatment</li> </ul>					
<b>E: Shares information (includes the following):</b> <ul style="list-style-type: none"> <li>• Assesses patient’s/family’s understanding of problems and desire for more info</li> <li>• Explains using words that family can understand</li> <li>• Asks if family has any more questions</li> </ul>					
<b>F: Reaches agreement (includes the following):</b> <ul style="list-style-type: none"> <li>• Includes family in choices and decisions to the extent they desire</li> <li>• Checks for mutual understanding of diagnostic and/or treatment plans</li> <li>• Asks about acceptability of diagnostic and/or treatment plans</li> <li>• Identifies additional resources as appropriate</li> </ul>					
<b>G: Provides closure (includes the following):</b> <ul style="list-style-type: none"> <li>• Asks if patient and family have questions, concerns or other issues</li> <li>• Summarizes</li> <li>• Clarifies future time when progress will again be discussed</li> <li>• Provides appropriate contact information if interim questions arise</li> <li>• Acknowledges patient and family, and closes interview</li> </ul>					
<b>H. Demonstrates Empathy (includes the following):</b> <ul style="list-style-type: none"> <li>• Clinician’s demeanor is appropriate to the nature of the conversations</li> <li>• Shows compassion and concerns</li> <li>• Identifies/labels/validates patient’s and family’s emotional responses</li> <li>• Responds appropriately to patients and family’s emotional cues</li> </ul>					
<b>I: Communicates accurate information (includes the following):</b> <ul style="list-style-type: none"> <li>• Accurately conveys the relative seriousness of the patient’s condition</li> <li>• Takes other participating clinician’s input into account</li> <li>• Clearly conveys expected disease course</li> <li>• Clearly presents and explains options for future care</li> <li>• Gives enough clear information to empower decision making</li> </ul>					

\*Adapted from: Essential Elements: The Communication Checklist, © 2001 Kalamazoo Consensus Statement Group, and from: Rider EA. Interpersonal and Communication Skills. In: Rider EA, Nawotniak RH. *A Practical Guide to Teaching and Assessing the ACGME Core Competencies, 2nd edition*. Marblehead, MA: HCPPro, Inc., 2010. © 2010 HCPPro, Inc. Used with permission. Contacts: Elizabeth Rider, MSW, MD - elizabeth\_rider@hms.harvard.edu (member, Kalamazoo Consensus Statement Group) and Aaron Calhoun, MD - aaron.calhoun@louisville.edu (PERCS Program)

**What did the participant(s) do best? (Please pick three choices)**

- 
- Builds a Relationship
  - Opens the Discussion
  - Gathers Information
  - Understands the Patient's and Family's Perspective
  - Shares Information
  - Reaches Agreement
  - Provides Closure
  - Demonstrates Empathy
  - Communicates Accurate Information
- 

**Why did you choose those particular answers?****In which domains could the participant(s) improve? (Please pick three choices)**

- 
- Builds a Relationship
  - Opens the Discussion
  - Gathers Information
  - Understands the Patient's and Family's Perspective
  - Shares Information
  - Reaches Agreement
  - Provides Closure
  - Demonstrates Empathy
  - Communicates Accurate Information
- 

**What could have been done better?**

**\*Adapted from:** Essential Elements: The Communication Checklist, © 2001 Kalamazoo Consensus Statement Group, and from: Rider EA. Interpersonal and Communication Skills. In: Rider EA, Nawotniak RH. *A Practical Guide to Teaching and Assessing the ACGME Core Competencies, 2nd edition*. Marblehead, MA: HCPro, Inc., 2010. © 2010 HCPro, Inc. Used with permission. Contacts: Elizabeth Rider, MSW, MD - elizabeth\_rider@hms.harvard.edu (member, Kalamazoo Consensus Statement Group) and Aaron Calhoun, MD - aaron.calhoun@louisville.edu (PERCS Program)

**Conflict of interest**

None.

**Acknowledgments**

Funding for the PACE program was initially provided by the Kosair Charities Community Trust Fund Fellows Research Grant. Ongoing maintenance of funding is provided by the University of Louisville and the Norton Healthcare System as part of the annual budget for ongoing simulation-based skills training.

**References**

- [1] Fallowfield L, Jenkins V. Communicating sad, bad, and difficult news in medicine. *Lancet* 2004;363:312–9.
- [2] Fallowfield L. Giving sad and bad news. *Lancet* 1993;342:476–8.
- [3] Rosenbaum M, Kreiter C. Teaching delivery of bad news using experiential sessions with standardized patients. *Teach Learn Med* 2002;14:144–9.
- [4] Meert K, Thurston C, Thomas R. Parental coping and bereavement outcome after the death of a child in the pediatric intensive care unit. *Pediatr Crit Care Med* 2001;2:324–8.
- [5] Jurkovich G, Pierce B, Pananen L, Rivara F. Giving bad news: the family perspective. *J Trauma* 2000;48:865–70.
- [6] Zick A, Granieri M, Makoul G. First-year medical students' assessment of their own communication skills: a video-based, open-ended approach. *Patient Educ Couns* 2007;68:161–6.
- [7] ACGME core competency [ACGME Outcome Project Website]. Accreditation council for graduate medical education. <http://acgme.org/outcome/comp/compMin.asp> [accessed September 2010].
- [8] USMLE. USMLE 2013 Step 2 clinical skills content description and general information. Available at: <http://www.usmle.org/pdf/step-2-cs-info-manual.pdf> [accessed December 2013].
- [9] Schirmer JM, Mauksch L, Lang F, Marvel MK, Zoppi K, Epstein RM, et al. Assessing communication competence: a review of current tools. *Fam Med* 2005;37:184–92.
- [10] Rider EA. Interpersonal and communication skills. In: Rider EA, Nawotniak RH, editors. *A practical guide to teaching and assessing the ACGME core competencies*. 2nd ed., Marblehead, MA: HCPPro Inc.; 2010.
- [11] Calhoun AW, Rider EA, Peterson E, Meyer EC. Multi-rater feedback with gap analysis: an innovative means to assess communication skill and self-insight. *Patient Educ Couns* 2010;80:321–6.
- [12] Deveugele M, Derese A, De Maesschalck S, Willems S, De Maesenner J. Teaching communication skills to medical students, a challenge in the curriculum? *Patient Educ Couns* 2005;58:265–70.
- [13] Lane C, Rollnick S. The use of simulated patients and role-play in communication skills training: a review of the literature to August 2005. *Patient Educ Couns* 2007;67:13–20.
- [14] Brown J. Transferring clinical communication skills from the classroom to the clinical environment: perceptions of a group of medical students in the United Kingdom. *Acad Med* 2010;85:1052–9.
- [15] Peterson EB, Porter MB, Calhoun AC. A Simulation-based curriculum to address relational crises in medicine. *J Grad Med Educ* 2012;4:351–6.
- [16] Yudkowsky R, Alseidi A, Cintron J. Beyond fulfilling the core competencies: an objective structured clinical examination to assess communication and interpersonal skills in a surgical residency. *Curr Surg* 2004;61:499–503.
- [17] Symons A, Swanson A, McGuigan D, Orrange S, Akl E. A tool for self-assessment of communication skills and professionalism for residents. *BMC Med Educ* 2009;9:1.
- [18] Participants in the Bayer–Fetzer Conference on Physician–Patient Communication in Medical Education. Essential elements of communication in medical encounters: the Kalamazoo Consensus Statement. *Acad Med* 2001;76:390–3.
- [19] Rider EA, Hinrichs MM, Lown BA. A model for communication skills assessment across the undergraduate curriculum. *Med Teach* 2006;28:e127–34.
- [20] Baribeau DA, Mukovozov I, Sabljic T, Eva KW, deLottinville CB. Using an objective structured video exam to identify differential understanding of aspects of communication skills. *Med Teach* 2012;34:e242–50.
- [21] Wong RY, Saber SS, Ma I, Roberts JM. Using television shows to teach communication skills in internal medicine residency. *BMC Med Educ* 2009;9:9.
- [22] Razack S, Meterisian S, Morin L, Snell L, Steinert Y, Tabatabai D, et al. Coming of age as communicators: differences in the implementation of common communications skills training in four residency programmes. *Med Educ* 2007;41:441–9.
- [23] Joyce BL, Scher E, Steenbergh T, Voutt-Goos MJ. Development of an institutional resident curriculum in communication skills. *J Grad Med Educ* 2011;4:524–8.
- [24] Calhoun AW, Rider EA, Meyer EC, Lamiani G, Troug RD. Assessment of communication skills and self-appraisal in the simulated environment: feasibility of multirater feedback with gap analysis. *Simul Healthc* 2009;4:22–9.
- [25] Joyce BL, Steenbergh T, Scher E. Use of the Kalamazoo Essential Elements Communication Checklist (Adapted) in an institutional interpersonal and communication skills curriculum. *J Grad Med Educ* 2010;2:165–9.
- [26] Bloch R, Norman G. Generalizability theory for the perplexed: a practical introduction and guides: AMEE Guide No. 68. *Med Teach* 2012;34:960–92.



Adv in Health Sci Educ (2009) 14:575–594  
DOI 10.1007/s10459-008-9142-2

ORIGINAL PAPER

## Evaluating the effectiveness of rating instruments for a communication skills assessment of medical residents

Cherdsak Iramaneerat · Carol M. Myford · Rachel Yudkowsky · Tali Lowenstein

Received: 28 June 2008 / Accepted: 17 October 2008 / Published online: 5 November 2008  
© Springer Science+Business Media B.V. 2008

**Abstract** The investigators used evidence based on response processes to evaluate and improve the validity of scores on the Patient-Centered Communication and Interpersonal Skills (CIS) Scale for the assessment of residents' communication competence. The investigators retrospectively analyzed the communication skills ratings of 68 residents at the University of Illinois at Chicago (UIC). Each resident encountered six standardized patients (SPs) portraying six cases. SPs rated the performance of each resident using the CIS Scale—an 18-item rating instrument asking for level of agreement on a 5-category scale. A many-faceted Rasch measurement model was used to determine how effectively each item and scale on the rating instrument performed. The analyses revealed that items were too easy for the residents. The SPs underutilized the lowest rating category, making the scale function as a 4-category rating scale. Some SPs were inconsistent when assigning ratings in the middle categories. The investigators modified the rating instrument based on the findings, creating the Revised UIC Communication and Interpersonal Skills (RUCIS) Scale—a 13-item rating instrument that employs a 4-category behaviorally anchored rating scale for each item. The investigators implemented the RUCIS Scale in a subsequent communication skills OSCE for 85 residents. The analyses revealed that the RUCIS Scale functioned more effectively than the CIS Scale in several respects (e.g., a more uniform distribution of ratings across categories, and better fit of the items to the measurement model). However, SPs still rarely assigned ratings in the lowest rating category of each scale.

---

C. Iramaneerat (✉)

Department of Surgery, Faculty of Medicine, Siriraj Hospital, Mahidol University,  
Siamindra Building 12th fl., 2 Prannok Rd. Bangkoknoi, Bangkok 10700, Thailand  
e-mail: sicir@mahidol.ac.th

C. M. Myford

Department of Educational Psychology, College of Education, University of Illinois at Chicago,  
Chicago, IL, USA

R. Yudkowsky · T. Lowenstein

Department of Medical Education, College of Medicine, University of Illinois at Chicago, Chicago, IL,  
USA

 Springer

**Keywords** Validity · Rating scale · Communication skills · Many-faceted Rasch measurement · OSCE

## Introduction

Communication and interpersonal skills are one of the six core competencies for which residency programs have to demonstrate training outcomes (Accreditation Council for Graduate Medical Education 1999). An assessment of residents' communication skills that can provide valid inferences about their ability to exchange information and ally with patients requires an observed interaction with patients. The Accreditation Council for Graduate Medical Education (ACGME) and the American Board of Medical Specialties (ABMS) recommend using an assessment format that asks residents to interact with standardized patients (SPs) in an Objective Structured Clinical Examination (OSCE) as the most desirable approach for communication skills assessment (Bashook and Swing 2000).

The rating instrument that a standardized patient uses to record his/her observations of a resident's performance during a communication skills OSCE plays a critical role in providing valid inferences from an assessment. A rating instrument not only guides the observation but also dictates the scoring of the performance of individual residents. Several rating instruments for the assessment of medical communication skills by SPs in OSCE settings have been developed and validated, including the Interpersonal and Communication Skills Checklist (Cohen et al. 1996), the Interpersonal Skills Rating Form (Schnabl et al. 1991), the Arizona Clinical Interview Rating Scale (Stillman et al. 1976, 1986), the Brown University Interpersonal Skill Evaluation (Burchard and Rowland-Morin 1990), the SEGUE Framework (Makoul 2001), the Liverpool Communication Skills Assessment Scale (LCSAS) (Humphris 2002; Humphris and Kaney 2001), and the Patient-Centered Communication and Interpersonal Skills (CIS) Scale (Yudkowsky et al. 2004, 2006).

Despite the many available rating instruments for communication skills assessment in OSCE settings, choosing an appropriate instrument to score residents' performance in a communication skills OSCE is not an easy task. Validity evidence that supports the use of scores obtained from these rating instruments is quite limited. Researchers conducting validity studies of these instruments have focused mainly on reporting internal consistency reliability, inter-rater reliability, and correlations of scores with measures of other variables. According to the *Standards for Educational and Psychological Testing* (American Educational Research Association et al. 1999), such validity research only provides evidence based on internal structure and relations to other variables, leaving out evidence based on test content, response processes, and consequences.

In this study, we evaluated validity evidence related to the use of one of the existing communication skills OSCE rating instruments—the Patient-Centered Communication and Interpersonal Skills (CIS) Scale. We focused on evidence based on response processes, a source of validity evidence that test score users often overlook. In the context of a communication skills OSCE, the validity evidence based on response processes refers to the evaluation of the extent to which the SPs apply rating criteria to rate the residents' performance in a manner that is consistent with the intended interpretation and uses of scores (American Educational Research Association et al. 1999).

There are several approaches that researchers can use to gather validity evidence based on response processes. Researchers can collect some pieces of evidence before the OSCE

administration (e.g., documenting the rating criteria and the processes for selecting, training, and qualifying SPs). Researchers can collect other pieces of evidence at the time a SP rates the performance (e.g., engaging SPs in verbal think-aloud during the rating process, thus allowing researchers to know what SPs are thinking while deciding what rating they will assign (Heller et al. 1998)). The focus of this study was on gathering validity evidence related to response processes after an OSCE administration (i.e., when all the ratings were available to us). That is, we carried out a psychometric analysis of ratings to investigate to what extent the OSCE ratings were consistent with the intended uses of the scores. OSCE ratings are the result of the interaction between residents, cases, items (and their rating scales), and SPs. A comprehensive validity study of response processes for an OSCE would require close examination of responses related to all these components of an OSCE. In this study, we limited the scope of our analyses to response processes related to items and scales on the rating instrument. That is, we investigated the extent to which SPs used the rating instrument to rate the residents' performance in a way that was consistent with the intended uses of the scores.

This study looked at the use of the CIS Scale in the scoring of internal medicine residents' performance in communication skills OSCEs carried out at the University of Illinois at Chicago (UIC). The purposes of our study were (1) to evaluate the effectiveness of the CIS Scale in scoring the residents' performance in the communication skills OSCE, (2) to use the findings obtained from the analysis to determine whether the rating instrument needed to be revised to improve its effectiveness, (3) to use the results from the analysis to guide the instrument revision process, and (4) to compare the original CIS Scale to the modified rating instrument to determine whether the modifications helped improve the scale's functioning, thus in effect enhancing the validity of the inferences made from scores on the communication skills OSCE. In the course of evaluating the effectiveness of these two rating instruments, we demonstrate how researchers can analyze OSCE rating data to provide validity evidence related to response processes.

## Method

### Research design

We carried out the study in two phases. The first phase was a retrospective analysis of the communication skills OSCE ratings for internal medicine residents obtained in 2003, in which SPs used the CIS scale to rate the performance of residents. We identified certain items and scales on that rating instrument that did not function effectively and revised the rating instrument to address those weaknesses. We piloted the revised instrument with a small group of SP trainers and medical faculty members and then further revised the instrument based on the comments obtained from the pilot study. This led to a development of a revised rating instrument for communication skills assessment called the Revised UIC Communication and Interpersonal Skills (RUCIS) scale.

In the second phase of the study, we implemented the RUCIS scale in the 2007 communication skills OSCE for internal medicine residents. We carried out an analysis to evaluate the effectiveness of the revised rating instrument in order to determine whether the instrument modifications helped improve the effectiveness of the instrument. Both the 2003 and 2007 communication skills OSCEs were mandatory formative assessments conducted as part of the standard curriculum of the residency program.

## Participants

Participants in the 2003 communication skills OSCE included 68 internal medicine residents (51% PGY-2 and 49% PGY-3; 66% male and 34% female) and 8 SPs (38% male and 62% female). Participants in the 2007 communication skills OSCE included 85 internal medicine residents (54% PGY-1 and 46% PGY-2; 47% male and 53% female) and 17 SPs (29% male and 71% female).

## Rating instruments

The CIS Scale, which SPs used to rate the performance of residents in the 2003 communication skills OSCE, is an 18-item rating instrument. Each item asks SPs to provide an agreement rating using a 5-category rating scale, in which 1 corresponds to “strongly disagree” and 5 corresponds to “strongly agree.” Since all items are statements of desirable communication behaviors, higher ratings indicate higher level of communication competence (See Appendix A).

The RUCIS Scale, which SPs used to rate the performance of residents in the 2007 communication skills OSCE, is a 13-item rating instrument. Each item contains a short description of the particular aspect of communication under consideration and four behaviorally anchored rating categories unique to each item. For each item, the lowest rating category always describes the least appropriate behavior for that aspect of communication, while the highest rating category always describes the most appropriate behavior for that aspect. In addition to the four rating categories for each item, six items also have a “not applicable” option that SPs could use when they did not observe the behavior related to that aspect of communication (See Appendix B).

## SP training

In the 2003 communication skills OSCE, all the SPs took part in an intensive training program to learn how to portray the cases and how to rate resident performance before participating in the OSCE. The training program included a review and discussion of the case script and repeatedly practicing the appropriate portrayal of the cases under the supervision of a trainer. Training on the CIS scale included a review and discussion of the scale and practice using it to rate a videotaped or observed performance. There was no attempt to reach agreement between the SP and trainer in the ratings they assigned, but divergent ratings were noted and discussed. The trainer ensured that each SP could portray the case consistently and rate the performance of residents to the trainer’s satisfaction before the SP was allowed to participate in the communication skills OSCE.

In the 2007 communication skills OSCE, all the SPs also took part in an intensive SP training program similar to the training for the 2003 communication skills OSCE to ensure an accurate portrayal of the cases before participating in the OSCE. However, this time we employed a frame-of-reference (FOR) approach in training the SPs to provide ratings (Bernadin and Buckley 1981). Prior to training, a group of SP trainers reviewed selected videotaped OSCE sessions and provided a consensus “gold standard” rating for each item in each encounter. During the training sessions SPs rated the selected videotaped OSCE sessions using the RUCIS scale, compared their ratings to the trainers’ “gold standard” ratings, and discussed the rationale for the gold standard. By practicing and receiving feedback from several videotaped OSCE sessions, the SPs developed a common rating standard (i.e., frame) by which to evaluate residents’ performances.

### OSCE administrations

Both OSCEs employed the same cases and the same administration format. Six residents were assessed in each half-day session. In each session, each resident encountered six different SPs in six different clinical scenarios (cases). In each case, residents spent 10 min in the encounter with the SP, 5 min reviewing task-related educational materials while the SP rated the performance, and another 5 min receiving verbal feedback from the SP. The task-related educational materials consisted of printed documents describing effective ways to interact with a patient in the situation they just encountered. The verbal feedback session provided SPs and residents with the opportunity to discuss effective and ineffective behaviors observed during the encounter, and to practice techniques that the SP suggested. The SP did not inform the resident of his/her specific ratings. The six communication tasks that residents encountered were: (1) providing patient education, (2) obtaining informed consent, (3) dealing with a patient who refuses treatment, (4) counseling an elderly patient who has been abused, (5) giving bad news to a patient, and (6) conducting a physical examination. We repeated the OSCE sessions once or twice a week until all residents had the opportunity to participate in the OSCE, which took 2 and 4 months, for the 2003 and 2007 communication skills OSCE, respectively.

### Analyses

Because the OSCE is a multi-faceted assessment method where the rating of a resident's performance depends upon many factors, including the communication competence of the resident, the difficulty of the item on the rating instrument, the severity of the SP, and the difficulty of the case, we used a many-faceted Rasch measurement (i.e., Facets) model (Linacre 1989) to analyze the data. The Facets model uses a logarithmic function of the odds of receiving a rating in a given category as compared to the odds of receiving a rating in the next lower category to define the communication competence of residents, the difficulty of items, the severity of SPs, and the difficulty of cases. All measures of these four facets are reported on the logit scale, which is a linear, equal interval scale. Higher logit measures indicate more competent residents, more difficult items, more severe SPs, and more difficult cases. Because there were multiple rating categories for each item, the Facets model also calculated a set of *step thresholds* for each item. (A step threshold is the transition point between two adjacent categories, where the probabilities of receiving a rating in the two categories are equal.) We used the Facets computer program (Linacre 2005) to conduct the analyses.

To ensure that the analyses to obtain validity evidence based on response processes would be based on reliable data, we first examined the degree of reproducibility of residents' communication competence measures—validity evidence related to the internal structure of test scores. We calculated a measure of internal consistency reliability, the resident separation reliability, which is an index analogous to KR-20 or Cronbach's Alpha. Because ratings of multiple items on the same case by the same SP can be dependent on one another, which could lead to overestimation of reliability (Sireci et al. 1991; Thissen et al. 1989), we used cases (rather than items) as scoring units. That is, we averaged the ratings a SP gave to all items in a given case to produce a case score, which we considered as one rating in the Facets analysis.

An effective rating instrument for an OSCE should produce ratings that satisfy two tests related to response processes. The first one involves determining whether each rating scale functioned appropriately (i.e., were the categories on the scales that the SPs used

well-defined, mutually exclusive, and exhaustive). The second one involves determining whether each item on the rating instrument functioned properly (i.e., when evaluating each resident's performance, did SPs assign ratings for each item in a consistent fashion).

We used the following six criteria (Linacre 2004) as guidelines for determining whether each rating scale category for each item functioned effectively (i.e., to determine whether the rating categories of each item were well-defined, mutually exclusive, and exhaustive):

- (1) There should be at least 10 ratings in each rating category to allow accurate calibration of step thresholds.
- (2) The frequency distribution of ratings across categories should have a uniform or unimodal pattern. If SPs use only a few of the rating categories and rarely use other rating categories, the resulting irregular distribution of ratings indicates a poorly functioning scale that cannot effectively differentiate residents according to their levels of communication competence.
- (3) The average measures of residents' communication competence should increase as the rating categories increase. In other words, residents who receive ratings in higher categories should have higher overall communication competence measures than those who receive ratings in lower categories.
- (4) The step thresholds should increase as the rating categories increase. This criterion mirrors the third criterion. Failure of step thresholds to increase as the rating categories increase is called *step disordering*, which suggests that SPs may have difficulty differentiating the performance of residents in those categories. One or more of the rating categories for a particular item may not be clearly defined.
- (5) The step thresholds should advance at least 1 logit, but not more than 5 logits. The finding that two step thresholds advance by less than 1 logit would suggest that those two rating categories are practically inseparable. That is, SPs may not be able to reliably differentiate between them. On the other hand, step thresholds that are too far apart are an indication of a possible dead zone on the scale where measurement loses its precision.
- (6) The outfit mean-square value for each rating category should be less than 2.0. An outfit mean-square value is a statistical index that indicates how well the ratings in each category fit the measurement model. Its value can range from 0 to infinity, with an expected value of 1. A high outfit mean-square value for a rating category is an indicator that some SPs used that rating category in an unexpected or surprising manner that was inconsistent with the way that other SPs used that category.

In addition to evaluating the functioning of the scale categories, we evaluated fit statistics for each item on the instrument to determine whether SPs provided aberrant ratings on any items, which might indicate problematic response processes. These fit statistics are indices that indicate how well the rating data for each item fit the measurement model. In this study, we examined both outfit and infit mean-square statistics for each item. We calculated an outfit mean-square value for each item by dividing the sum of the squared standardized residuals for the item by its degree of freedom. (A residual is the difference between the rating a SP assigned a resident on an item and the rating the measurement model predicted the SP would assign.) This calculation produces a value that can range from 0 to infinity, with an expectation of 1.0. Values larger than 1.0 indicate the presence of unmodeled noise in the ratings for that item (i.e., unexpected ratings that SPs assigned when evaluating residents, given how SPs assigned ratings for other items). By contrast, values less than 1.0 indicate that there was too little variation in the ratings SPs assigned for that item (Linacre and Wright 1994; Wright and Masters 1982). However, outfit



mean-square values are very sensitive to outlier ratings. To reduce the influence of outlier ratings, we weighted each squared standardized residual by its information function before we summed them. (This involved applying differential weights to standardized residuals. That is, residuals that resulted from SP ratings of items and cases that were far too easy or too difficult for residents received less weight than those that resulted from SP ratings of items and cases that were at the appropriate difficulty level for residents.) This calculation produced an infit mean-square statistic that has the same distribution and interpretation as an outfit mean-square statistic, but is more immune to the influence of the ratings for residents on items or cases that are far too easy or difficult for them. Wright and Linacre (1994) recommended that an appropriate mean-square fit statistic for judge-mediated ratings should be in the range of 0.4–1.2.

From the analysis of the 2003 communication skills OSCE ratings, we identified the items and rating categories on the CIS Scale that did not function effectively according to one or more of these criteria. We then used these findings to guide the development of a modified rating instrument—the RUCIS Scale. We implemented the RUCIS Scale in the 2007 communication skills OSCE and evaluated the effectiveness of the revised instrument using the same criteria to determine whether the modifications helped improve the effectiveness of the instrument, thus in effect enhancing the validity of the score interpretation.

## Results

### Evaluating the effectiveness of the CIS scale

The analysis of the 2003 communication skills OSCE revealed that this group of residents was highly competent relative to the items and cases on the CIS Scale (Table 1). The average resident communication competence measure was higher than the average item and case difficulty measures, and there were few items or cases appropriate for measuring the communication competence of residents who were in the upper range of the communication competency continuum (i.e., in the 0.75–2.5 logits range). These findings suggest that these items and cases were not very well suited to measuring the communication competence of this group of residents (i.e., it was a relatively easy assessment for them). Using cases as scoring units, our analysis yielded a resident separation reliability of 0.74.

**Table 1** Summary of measures obtained from the analysis of the communication skills OSCEs

Measurement facets	Minimum (logits)	Maximum (logits)	Mean (logits)	SD (logits)
<b>A. 2003 Communication skills OSCE</b>				
Resident competence	−0.40	2.44	0.78	0.61
Item difficulty	−0.71	0.83	0	0.44
Case difficulty	−0.45	0.30	0	0.25
<b>B. 2007 Communication skills OSCE</b>				
Resident competence	−1.85	1.68	−0.17	0.68
Item difficulty	−0.91	0.98	0	0.60
Case difficulty	−0.99	0.65	0	0.49





**Table 2** Comparing the functioning of the CIS scale (2003) and RUCIS scale (2007) using Linacre's (2004) guidelines

	CIS scale 5-category scales 18 items	RUCIS scale 4-category scales 13 items
Resident separation reliability	0.74	0.71
Criteria for evaluating the functioning of the categories on each rating scale		
At least 10 ratings in each category	5 items (28%)	6 items (46%)
Uniform/unimodal distribution of ratings across categories	1 item (6%)	12 items (92%)
Residents with higher category ratings have higher overall communication competence measures	7 items (39%)	12 items (92%)
No step disordering	9 items (50%)	11 items (85%)
Step thresholds advance by at least 1 logit, but not more than 5 logits	1 item (6%)	10 items (77%)
An outfit mean-square value <2.0 for each rating category	11 items (61%)	13 items (100%)
Criteria for evaluating the functioning of the items on the instrument		
Outfit mean-square values <1.2	14 items (78%)	12 items (92%)
Infit mean-square values <1.2	16 items (89%)	13 items (100%)

appeared to be relatively easy for these residents, resulting in an unbalanced distribution of ratings across the five rating categories: about 70–80% of all ratings were 4 s or 5 s. The only item that exhibited an acceptable rating distribution was item 18, which showed a unimodal distribution that peaked in the middle categories.

The analysis also revealed that some SPs experienced difficulty in differentiating between the middle categories of the 5-category agreement scale, as demonstrated by the failure of the average measures and step thresholds to increase properly along with the rating categories. Only seven items (items 4, 5, 6, 9, 11, 16, and 18) exhibited proper advancement of average resident communication competence measures as the rating categories increased. Nine items (items 2, 4, 6, 10, 11, 13, 14, 15, and 17) showed disordered step thresholds. Seven items (items 7, 8, 9, 10, 13, 14, and 15) had one or more rating categories with outfit mean-square values equal to or greater than 2, reflecting inconsistent use of the categories. Only one item (item 12) had all adjacent step thresholds separated by at least one logit. The other 17 items had one or more step thresholds that were too close to adjacent thresholds, especially for step thresholds in the middle of the scale. However, none of the 18 items had step thresholds that advanced by more than five logits, suggesting that there were no significant gaps between the categories.

We summarized item fit statistics in Table 3. Four items (items 3, 5, 10, and 15) had outfit mean-square values higher than 1.2, indicating that some SPs assigned ratings for those items that were unexpectedly high or low, given the other ratings that the SPs assigned. Items 10 and 15 had infit mean-square values higher than 1.2. A close examination of the unexpected ratings for items 10 (I felt you encouraged me to ask questions) and 15 (I felt you were careful to use plain language) revealed that six out of eight SPs were inconsistent in rating item 10, and seven out of eight SPs were inconsistent in rating item 15. Apparently, the SPs did not have a shared understanding of what they were evaluating when using these two items. This finding suggested that we needed to revise these items to make them clearer to SPs.

**Table 3** Summary of item fit statistics

Item fit statistics	Minimum	Maximum	Mean	SD
A. 2003 Communication skills OSCE				
Outfit mean-square values	0.71	2.35	1.08	0.37
Infit mean-square values	0.76	1.72	1.00	0.23
B. 2007 Communication skills OSCE				
Outfit mean-square values	0.86	1.22	1.00	0.08
Infit mean-square values	0.86	1.16	1.00	0.07

### Modifying the rating instrument

The findings from our validity study revealed that there were several aspects of the CIS Scale that did not function well. Using these findings as our guide, we worked with medical faculty and SPs to revise the CIS Scale in several ways. Instead of using a single Likert-style agreement rating scale that was applicable to all items on the instrument, we devised a behaviorally anchored rating scale (BARS) (Bernardin and Smith 1981; Smith and Kendall 1963) that provided a detailed description of the specific communication behavior characteristic of each rating category for each item. Our expectation was that the change in the scale format would make each rating scale more specific to the context of a particular item and less open to SPs' idiosyncratic interpretations.

Because our analysis revealed that the lowest rating category on the CIS Scale was a non-functioning category, we decided to change the scale format from 5-category scales to 4-category scales. To address the problem of an unbalanced rating distribution in which 70–80% of ratings were positive ratings, while only 20–30% of ratings were neutral or negative ratings, we developed 4-category scales that were saturated on the positive side. In other words, we created a separate rating scale for each item with only one category describing inadequate performance and three categories describing satisfactory communication behaviors that exemplified progressively higher levels of performance.

We also provided a “not applicable” option for six items. Our goal was to eliminate some unexpected ratings that SPs assigned in the neutral category of the agreement scale when they found themselves unable to rate a certain aspect of communication during the encounter because they did not observe any evidence that the resident engaged in that aspect.

Although we did not change the content coverage of the rating instrument, we revised the items to eliminate redundancy and improve their clarity. We combined into one item the redundant items that addressed the same aspect of communication. Specifically, we combined items 1 and 2 into an item on friendly communication; combined items 7, 8, and 9 into an item on discussion of options; combined items 10, 11, and 12 into an item on encouraging questions; and combined items 13 and 14 into an item on providing a clear explanation. We created a new item on physical examination to allow SPs to separate the act of providing an explanation of a physical examination from the act of providing an explanation about medical conditions.

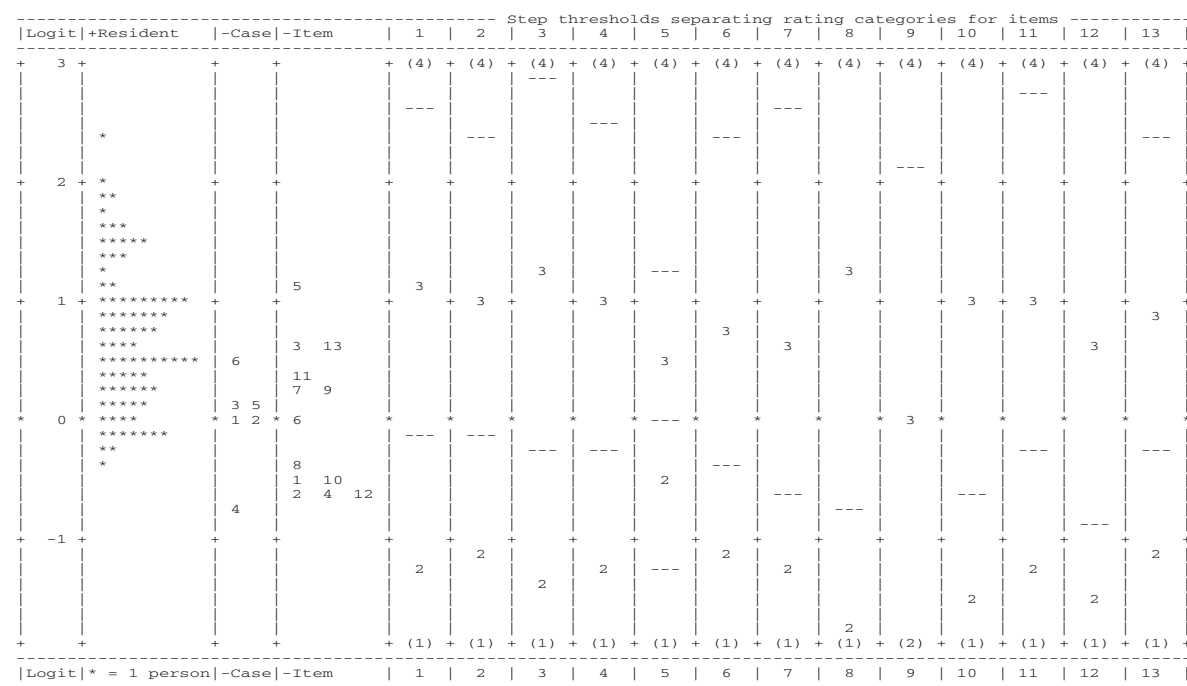
Finally, we attempted to make several items more difficult by requiring that residents demonstrate communication behaviors that are more sophisticated and/or difficult to perform to qualify for a rating in the highest category.

These modifications led to the development of a revised rating instrument, called the RUCIS Scale (Appendix B), which we later used in the scoring of residents' performance in the 2007 communication skills OSCE.

Evaluating the effectiveness of the RUCIS scale

The analysis of the 2007 communication skills OSCE revealed that this set of items was better targeted for measuring the communication competence of the residents (See Table 1 and columns 2–4 of Fig. 2). The distribution of resident communication competence measures was better aligned with the distributions of item and case difficulty measures than was the case for the 2003 communication skills OSCE. Using cases as scoring units, our analysis yielded a resident separation reliability of 0.71. Despite fewer numbers of items on the RUCIS Scale, the ratings on this revised instrument could achieve the same level of internal consistency reliability as the level obtained from the CIS Scale.

Table 2 provides a point-by-point comparison of the findings from our analyses of the functioning of the CIS Scale and the RUCIS Scale. We found that seven items on the revised instrument still had fewer than 10 ratings assigned in the lowest category. Beyond this, nearly all the items and rating scales appearing on the RUCIS Scale satisfied Linacre’s criteria. All items but one had a uniform distribution of ratings that peaked in the middle or at the high end. Item 5 (interest in me as a person) was the only item that had a rating distribution that peaked in rating category 1. Item 2 (respectful treatment) was the only item that did not show increasing average measures as rating categories increased. The rating categories for all items fit the measurement model (i.e., all outfit mean-square values for the rating categories were less than 2). Items 7 and 12 were the only two items with disordered step thresholds. Some of the distances between step thresholds for Items 5, 6, and 10 were too narrow (i.e., less than one logit apart). However, all the step thresholds for



**Fig. 2** A construct map showing the better alignment between the resident communication competence measures and the item and case difficulty measures for the 13 items on the RUCIS Scale



the other items were appropriately ordered and advanced by more than one logit but less than five logits.

We summarized item fit statistics obtained from the analysis of the 2007 communication skills OSCE in Table 3. All items showed good fit to the measurement model according to their infit mean-square values. Item 5 (interest in me as a person) was the only item with an outfit mean-square value higher than 1.2, indicating too much unexplained variance in the ratings that SPs assigned for this item. Thus, it was the only item that needed close examination to try to determine what made it difficult for SPs to use the item's behaviorally anchored rating scale to assign consistent ratings.

## Discussion

This study demonstrated the process of using validity evidence obtained from a Facets analysis to help revise an assessment instrument such as an OSCE rating scale. Validation is a continuing process of gathering and evaluating various sources of evidence to determine whether that evidence supports (or refutes) the proposed score interpretation. The two phases of this study correspond to the two stages of validation that Kane (2006) described. In the first phase of the study, we focused on finding ways to build a measurement instrument that possessed appropriate psychometric properties that would support the intended uses of OSCE scores. This phase corresponds to the *development stage* of validation. In the second phase, we critically evaluated whether the newly developed rating instrument actually functioned as predicted. This phase corresponds to the *appraisal stage* of validation.

In the first phase of our study, validity evidence based on response processes helped us identify several aspects of the CIS Scale that did not function as intended. The validity evidence suggested that the 5-category Likert-style agreement scale functioned as an unbalanced 4-category rating scale (i.e., most of the ratings were positive ratings, while only a few ratings were neutral or negative). This finding indicated that the items on the CIS Scale were too easy for this sample of residents. Results from our analyses also suggested that some SPs were unable to differentiate performance in the middle categories of the scale. Additionally, we found that some SPs assigned a number of surprising or unexpected ratings for item 10 (I felt you encouraged me to ask questions) and for item 15 (I felt you were careful to use plain language), suggesting that these SPs were not able to consistently apply the rating criteria for these two items to rate some residents' performances. All these pieces of validity evidence provided useful information to guide the development of a revised rating instrument in our attempt to address these weaknesses of the CIS Scale.

In the second phase of our validity study, we implemented the revised instrument in a later administration of the communication skills OSCE and carried out the same types of analyses that had revealed the inadequacies of the CIS Scale. We considered this as a test of whether the revised instrument could withstand the same validity challenges as its predecessor. We found that in many aspects the RUCIS Scale helped improve score interpretability. The SPs more consistently applied the rating criteria to rate residents' performances. The items on the RUCIS Scale fit the measurement model quite well. Providing a clear description of communication behavior that was appropriate for each rating category for the two misfitting items on the CIS Scale (items 10 and 15) helped eliminate confusion among SPs in rating these two aspects of communication (as demonstrated by good item fit statistics for items 7 and 10 on the RUCIS Scale).

However, the modifications we made to the rating instrument did not address all the validity issues we identified in the CIS Scale. There was one area in which the revised instrument did not show significant improvement over its predecessor. When using the behaviorally anchored rating scales, SPs still assigned only a few ratings in the lowest rating category of many items. This could be due to a restricted range of communication competence among the particular sample of residents assessed. We developed the RUCIS Scale with a broad range of communication competence in mind—from very incompetent physicians to very competent physicians. The subjects included in the 2007 communication skills OSCE were a single group of residents in one residency program. This limited the range of observable communication skills that SPs were likely to see. If we were to assess a broader range of subjects, ranging from medical students in their early years of training to experienced physicians practicing in various specialties from geographically diverse medical settings, the SPs would be more likely to observe a broader range of communication behaviors and would be more likely to employ the full range of rating categories appearing on each behaviorally anchored rating scale. Testing this hypothesis would require that researchers conduct additional studies to evaluate validity generalization (American Educational Research Association et al. 1999). That is, we are suggesting that researchers carry out studies to determine the extent to which variations in situational facets (e.g., residents from different residency programs, different SPs, etc.) may affect the assignment of ratings. The studies would help us determine how generalizable the results we obtained are across subjects that differ in education and experience, and across SPs.

Another possible explanation for non-uniform distributions of ratings is that SPs may have been uncomfortable assigning very low ratings to residents. If this were the case, then SP trainers could address this issue during the training, helping SPs understand that it is appropriate (and expected) that they will assign low ratings if they see evidence of physician behaviors that warrant those ratings. However, we would be a bit cautious in following this criterion too strictly. For a formative assessment or in a summative assessment where residents had not been properly trained, a uniform distribution of ratings is to be expected. However, in a summative assessment where the majority of residents have practiced the skills so that they are well prepared for the communication tasks, a skew distribution of ratings where only few residents would have ratings in lower categories can be obtained, which might not suggest a problem with the rating instrument.

The evaluation of item fit statistics for the RUCIS scale revealed that item 5 (interest in me as a person) was the only item with too much unexplained variance in its ratings. Interestingly, two of the SPs were responsible for 65% of the statistically significantly unexpected ratings (i.e., ratings with an absolute value of their standardized residuals larger than 2.0) for this item. This finding suggests that the source of error in the ratings of item 5 might be due to the inconsistency of only two SPs, and that the fit of the item might be improved through additional training of these two SPs to clear up any confusion they might have experienced when rating this item.

Although we carried out the study in two phases that addressed both the development and appraisal stages of validation (Kane 2006), this study by no means presents a complete validation effort. Validation is an ongoing process of gathering various sources of evidence to support proposed score interpretations. One could consider the findings from the second phase of this study as input to further modify the rating instrument to craft an even more psychometrically sound assessment, thus cycling back to the development stage of validation once again. For example, our results suggest that item 5 on the RUCIS Scale is still problematic, since it continues to show inadequate fit to the measurement model. Additional modification on this item is a potential area for further instrument improvement.

There are some limitations related to the interpretation and application of the findings from this study. The first limitation is the instrument's limited focus on patient-centered medical communication skills. The ACGME's (1999) definition of communication skills emphasizes the importance of the ability to communicate not only with patients but also with other members of a healthcare team. The RUCIS Scale does not address the skills needed to communicate effectively with other members of a healthcare team. The psychometric properties of the RUCIS Scale demonstrated in this study might only apply to its use in an OSCE setting where SPs are trained properly on how to use the rating instrument. Another limitation of this study is the homogeneity of the resident samples we examined. Since our participants were internal medicine residents from a single training program, they were relatively homogeneous in their medical communication experience. Communication behaviors that were not observed in these residents might be evident when other groups of physicians are assessed. A multi-center trial of the rating instrument that involves medical schools from various geographical regions could study how the RUCIS Scale functions with a more heterogeneous group of physicians.

We hope that the findings from our study will benefit the medical education community in several ways. First, the product of this validation effort—the RUCIS Scale, along with validity evidence that supports its uses in the communication skills OSCE, should serve the needs of many residency programs, especially given the increasing interest in communication skills assessment that the ACGME Outcome Project has generated. Second, our study provides a concrete example of how to use a many-faceted Rasch measurement approach to improve the quality of SP rating instruments and to provide validity evidence based on response processes as outlined in the 1999 *Standards for Educational and Psychological Testing* (American Educational Research Association et al. 1999). Finally, this study generated many interesting ideas for future research.

## Appendix A

### Items on the Patient-Centered Communication and Interpersonal Skills (CIS) scale

1. I felt you greeted me warmly upon entering the room.
2. I felt you were friendly throughout the encounter. You were never crabby or rude to me.
3. I felt that you treated me like we were on the same level. You never “talked down” to me or treated me like a child.
4. I felt you let me tell my story and were careful to not interrupt me while I was speaking.
5. I felt you were telling me everything; being truthful, up front and frank; not keeping things from me.
6. I felt you showed interest in me as a “person.” You never acted bored or ignored what I had to say.
7. I felt that you discussed options with me.
8. I felt you made sure that I understood those options.
9. I felt you asked my opinion, allowing me to make my own decision.
10. I felt you encouraged me to ask questions.
11. I felt you displayed patience when I asked questions.
12. I felt you answered my questions, never avoiding them.



13. I felt you clearly explained what I needed to know about my problem; how and why it occurred.
14. I felt you clearly explained what I should expect next.
15. I felt you were careful to use plain language and not medical jargon when speaking to me.
16. I felt you approached sensitive/difficult subject matters, such as religion, sexual history, tobacco/drug/alcohol history, sexual orientation, giving bad news, etc., with sensitivity and without being judgmental.
17. I felt the resident displayed a positive attitude during the verbal feedback session.
18. If given the choice in the future, I would choose this resident as my personal physician.

Note: All items are rated on a 5-point Likert scale, ranging from 1 (strongly disagree) to 5 (strongly agree).

## Appendix B

### Revised UIC Communication and Interpersonal Skills scale

#### *Instruction*

Please choose the option that best describes how you feel toward the resident's communication skills. Some items also have a "not applicable" option. Select this option when the context of the case does not allow you to observe that aspect of the resident's performance.

#### (1) **Friendly communication**

- You did not greet me, or greeted me perfunctorily, or communicated with me rudely during the encounter.
- Your greeting and/or behavior during the encounter was generally polite but impersonal or distant.
- You greeted me warmly and communicated with me in a friendly, personal manner throughout the encounter.
- Your greeting and overall communication were friendly and compassionate. Your tone of voice was appropriate for the situation. Overall, you created an exceptionally warm and friendly environment that made me feel comfortable to tell you all of my problems.

#### (2) **Respectful treatment**

- You showed an obvious sign of disrespect during the encounter. You treated me as an inferior.
- You did not show disrespect to me. However, I observed some signs of condescending behavior. Although I believe it was unintentional, it made me feel that I was not at the same level with you.
- You gave several indications of respecting me. If there was a physical exam, this includes draping me appropriately.
- You were exceptionally respectful throughout the encounter. Your verbal and non-verbal communication showed respect for my privacy, my opinions, my rights, and my socioeconomic status.



**(3) Listening to my story**

- You rarely gave me any opportunity to tell my story or frequently interrupted me while I was talking, not allowing me to finish what I said. Sometimes I felt you were not paying attention (for example, you asked for information that I already provided).
- You let me tell my story without interruption, or only interrupted appropriately and respectfully. You seemed to pay attention to my story and responded to what I said appropriately.
- You allowed me to tell my story without interruption, responded appropriately to what I said, and asked thoughtful questions to encourage me to tell more of my story.
- You were an exceptional listener. You encouraged me to tell my story and checked your understanding by restating important points.

**(4) Honest communication**

- You did not seem truthful and frank. I felt that there might be something that you were trying to hide from me.
- You did not seem to hide any critical information from me.
- You explained the facts of the situation without trivializing negative information or possibilities (e.g., side effects, complications, failure rates).
- You were exceptionally frank and honest. You fully explained the positive and negative aspects of my condition. You openly acknowledged your own lack of knowledge or uncertainty, and things you would have to consult with others. When appropriate, you also suggested I seek a second opinion.
- Not applicable. There was no information for the resident to provide.

**(5) Interest in me as a person**

- You never showed interest in me as a person. You only focused on the disease or medical issue.
- In addition to talking about my medical issue, you spent some time getting to know me as a person.
- You spent some time exploring how my medical issue affects my personal or social life.
- You were exceptionally interested in me as a person. You not only explored how my medical problem affects my personal and social life, but also showed your willingness to help me address those challenges.

**(6) Discussion of options/plans**

- You did not explain any options or plans; you just told me what you would do without asking for my opinion.
- You explained options to me, but did not involve me in decision making. If you solicited my opinion, you just ignored it. You made all the decisions for me based on your medical opinion.
- You discussed options with me, made recommendations, solicited my opinion regarding the options/plans, and incorporated my opinion into your medical planning.

- You not only solicited my input, but also explored the reasons for my choice and showed your understanding and respect for my decisions by negotiating a mutually agreeable plan.
- Not applicable. There were no decisions to be made in this case.

**(7) Encouraging my questions**

- You did not solicit questions, or frequently avoided my questions, or did not provide helpful answers.
- You sometimes asked if I had questions, but seldom waited at least 5 seconds to allow me to formulate questions. You addressed my questions briefly without avoiding them.
- You actively encouraged me to ask questions, paused to allow me to formulate them, and provided clear and sufficient answers to all of my questions.
- You actively encouraged me to ask questions several times during the encounter, with sufficient wait time. You spent significant time and effort to answer my questions clearly and confirmed that I understood the answer and that my concerns were addressed.

**(8) Providing clear explanation**

- You rarely explained things to me; you did not help me better understand my situation.
- You gave me only brief explanations of my situation; you did not help me understand what would happen next.
- You gave me a full and understandable explanation of my situation, pertinent findings, and important next steps.
- You gave me a full explanation of my situation, your thinking about it and your recommendation, and probed my understanding by letting me summarize pertinent information.
- Not applicable. There was nothing to be explained in this case.

**(9) Physical examination**

- You never or rarely warned me about what you were going to do with my body. You also never or rarely explained what you found from the physical examination.
- You did not warn me about what you were going to do with my body, OR did not explain to me pertinent findings (both negative and positive) from your physical examination.
- You told me what you were going to do to my body AND described what you found.
- You helped me understand clearly what you were going to do to my body. You also provided clear explanation of what you found from the physical examination and the implications of your findings for my situation.
- Not applicable. There was no physical examination in this case.

**(10) Appropriate vocabulary**

- You used vocabulary that was too simple or too complex for me, or frequently used medical terms without explaining them to me. Sometimes I could not understand what you told me without asking for explanations of terms you used.

- Your vocabulary was generally appropriate but you sometimes inadvertently used medical terms without explaining them to me.
  - Your vocabulary was appropriate and if needed you provided brief explanations of any medical terms you used without need for prompting.
  - Your vocabulary was appropriate and you always provide clear and full explanation of relevant medical terms you used. In addition, you helped me better my understanding of my condition with the medical terms you explained to me.
- (11) **Sensitive subject matters (e.g., sexual history, tobacco/alcohol/drug use, religious/cultural issues, giving bad news, or difficult emotional states)**
- You never warned me before approaching sensitive subject matters. You seemed judgmental and clearly expressed your disapproval of my positions or feelings, making me feel uncomfortable about discussing these subjects or feelings with you.
  - You were careful and non-judgmental in discussing sensitive subject matters. However, you did not express understanding of my feelings and did not provide much emotional support.
  - You were sensitive about discussing difficult subjects and were respectful of my feelings. I never sensed that you were judgmental or disapproving of my positions or feelings on these subjects. You showed empathic understanding of my position or feelings and provided appropriate emotional support.
  - You were unusually empathic, sensitive and respectful of me and of my feelings and provided exceptional emotional support. In addition, you verbally reflected these back to me (e.g., “You sound sad”) to show your understanding.
  - Not applicable. There were no sensitive subject matters in this case.
- (12) **Receptiveness to feedback**
- You did not seem open to my feedback about your performance. You responded defensively or dismissively too many of my comments.
  - You listened to my feedback agreeably but passively. You did not actively participate during the feedback session.
  - You were able to describe some of your own effective and ineffective behaviors, were attentive to my comments, and had an open discussion with me about alternative behaviors.
  - You actively solicited additional feedback and showed signs of integrating my feedback into your behavioral repertoire. For example, you tried to role-play the communication techniques I suggested.
  - Not applicable. I provided no feedback.
- (13) **Do I want to see you again as my personal physician?**
- I did not feel comfortable in communicating with you at all. I would rather see a different physician.
  - I think you were okay in general and might come see you again.
  - I was impressed by the way you communicated with me. I would like to see you again.

- I was very impressed with you. I think you are one of the best physicians I have ever seen. I would feel very comfortable discussing any medical problems with you, and would recommend you to my friends.

## References

- Accreditation Council for Graduate Medical Education (1999). *The ACGME outcome project*. Retrieved August 2007, from <http://www.acgme.org/outcome/>.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Bashook, P. G., & Swing, S. (2000). *Toolbox of assessment methods*. Retrieved August 2007, from <http://www.acgme.org/outcome/assess/assHome.asp>.
- Bernadin, H. J., & Buckley, M. R. (1981). Strategies in rater training. *Academy of Management Review*, *6*, 205–212. doi:10.2307/257876.
- Bernardin, H. J., & Smith, P. C. (1981). A clarification of some issues regarding the development and use of behaviorally anchored rating scales (BARS). *The Journal of Applied Psychology*, *66*, 458–463. doi:10.1037/0021-9010.66.4.458.
- Burchard, K. W., & Rowland-Morin, P. A. (1990). A new method of assessing the interpersonal skills of surgeons. *Academic Medicine*, *65*, 274–276. doi:10.1097/00001888-199004000-00012.
- Cohen, D. S., Colliver, J. A., Marcy, M. S., Fried, E. D., & Schwartz, M. H. (1996). Psychometric properties of a standardized-patient checklist and rating-scale form used to assess interpersonal and communication skills. *Academic Medicine*, *71*(1(Suppl)), S87–S89.
- Heller, J. I., Sheingold, K., & Myford, C. M. (1998). Reasoning about evidence in portfolios: Cognitive foundations for valid and reliable assessment. *Educational Assessment*, *5*, 5–40. doi:10.1207/s15326977ea0501\_1.
- Humphris, G. M. (2002). Communication skills knowledge, understanding and OSCE performance in medical trainees: A multivariate prospective study using structural equation modeling. *Medical Education*, *36*, 842–852. doi:10.1046/j.1365-2923.2002.01295.x.
- Humphris, G. M., & Kaney, S. (2001). The Liverpool Brief Assessment System for communication skills in the making of doctors. *Advances in Health Sciences Education*, *6*, 69–80. doi:10.1023/A:1009879220949.
- Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: Praeger.
- Linacre, J. M. (1989). *Many-faceted Rasch measurement*. Chicago, IL: MESA Press.
- Linacre, J. M. (2004). Optimizing rating scale category effectiveness. In E. V. Smith Jr. & R. M. Smith (Eds.), *Introduction to Rasch measurement: Theory, models and applications* (pp. 258–278). Maple Grove, MN: JAM Press.
- Linacre, J. M. (2005). *Facets (Version 3.57) [computer program]*. Chicago, IL: Winsteps.
- Linacre, J. M., & Wright, B. D. (1994). Chi-square fit statistics. *Rasch Measurement Transactions*, *8*, 350.
- Makoul, G. (2001). The SEGUE Framework for teaching and assessing communication skills. *Patient Education and Counseling*, *45*, 23–34. doi:10.1016/S0738-3991(01)00136-7.
- Schnabl, G. K., Hassard, T. H., & Kopelow, M. L. (1991). The assessment of interpersonal skills using standardized patients. *Academic Medicine*, *66*(9 (Suppl)), S34–S36.
- Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement*, *28*, 237–247. doi:10.1111/j.1745-3984.1991.tb00356.x.
- Smith, P. C., & Kendall, L. M. (1963). Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales. *The Journal of Applied Psychology*, *47*, 149–155. doi:10.1037/h0047060.
- Stillman, P. L., Sabers, D. L., & Redfield, D. L. (1976). The use of paraprofessionals to teach interviewing skills. *Pediatrics*, *57*, 769–774.
- Stillman, P. L., Swanson, D. B., Smee, S., Stillman, A. E., Ebert, T. H., Emmel, V. S., et al. (1986). Assessing clinical skills of residents with standardized patients. *Annals of Internal Medicine*, *105*, 762–771.
- Thissen, D., Steinberg, L., & Mooney, J. (1989). Trace lines for testlets: A use of multiple-categorical response models. *Journal of Educational Measurement*, *26*(3), 247–260. doi:10.1111/j.1745-3984.1989.tb00331.x.

- Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, 8(3), 370. Available from: URL: <http://www.rasch.org/rmt/rmt383b.htm>.
- Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Chicago: MESA Press.
- Yudkowsky, R., Alseidi, A., & Cintron, J. (2004). Beyond fulfilling the core competencies: An objective structured clinical examination to assess communication and interpersonal skills in a surgical residency. *Current Surgery*, 61, 499–503. doi:10.1016/j.cursur.2004.05.009.
- Yudkowsky, R., Downing, S. M., & Sandlow, L. J. (2006). Developing an institution-based assessment of resident communication and interpersonal skills. *Academic Medicine*, 81(12), 1115–1122. doi:10.1097/01.ACM.0000246752.00689.bf.

## Standardized Patients and Scenario Preparation

เชิดศักดิ์ ไอรณณรัตน์

ภาควิชาศัลยศาสตร์ คณะแพทยศาสตร์ศิริราชพยาบาล

มหาวิทยาลัย มหิดล

### 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 L.C. Standardized patients. In: Diellehorst L.H, Dunnington G.L, Foise J.R. Teaching and learning in medical and surgical education: Lessons learned for the 21<sup>st</sup> 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.)

### Key Success for SP

- ฐานข้อมูลผู้ป่วยมาตรฐาน
- การมีบทที่ดี
- การซักซ้อม
- การจัดสภาวะแวดล้อมในขณะแสดง
- การประเมินคุณภาพ

### The Set Up

- Volunteer
- Equipments
- Fact sheet
- Time
- Recording
- Role assignment => Derole





25 Apr 2018

หัวข้อ : Assessing leadership and team management

## การประเมินการทำงานเป็นทีมทางการแพทย์

ผศ.พญ. ธัชวราภรณ์ จิระติวานนท์

การทำงานเป็นทีมมีความสำคัญอย่างยิ่ง ในการทำงานทางการแพทย์ และเป็นส่วนหนึ่งของทักษะ non-technical หรือ “ทักษะทางปัญญา และทักษะทางสังคม ที่ส่งเสริมกับความรู้ และทักษะหัตถการที่ถูกต้องเหมาะสม เพื่อเพิ่มศักยภาพและความปลอดภัยในการดูแลผู้ป่วย”<sup>(1)</sup> การขาดซึ่งทักษะการทำงานเป็นทีมที่ดี ส่งผลกระทบต่อการทำงานที่ไม่ราบรื่น เกิดความขัดแย้งกัน เกิดความผิดพลาด และที่สำคัญที่สุดคือ ส่งผลต่อความปลอดภัยของผู้ป่วย จึงจำเป็นอย่างยิ่งที่แพทย์และบุคลากรทางการแพทย์ทุกระดับ ควรได้รับการเรียนรู้ทักษะนี้และฝึกฝนจนเป็นกิจวัตร การประเมินการทำงานเป็นทีม นอกจากจะช่วยให้ผู้ประเมินรับรู้ระดับความสามารถของผู้เรียนแล้ว ยังช่วยให้ผู้เรียนได้ตระหนักถึงระดับความสามารถของตนเอง และผลักดันให้มีการพัฒนาตนเองต่อไป หากแต่การประเมินการทำงานเป็นทีม นั้น มีปัจจัยหลายประการเข้ามาเกี่ยวข้อง ที่อาจส่งผลให้การประเมินขาดประสิทธิภาพ ผู้ประเมิน จึงควรมีความรู้ความเข้าใจในหลักการและข้อจำกัดดังกล่าว เพื่อให้การประเมินมีประสิทธิภาพสูงสุด

### ทีมทางการแพทย์<sup>2</sup>

ในการทำงานทางการแพทย์ จำเป็นต้องทำงานร่วมกันกับบุคลากรอื่น ทั้งในหน่วยงานเดียวกันและต่างสาขาวิชาชีพ การทำงานเป็นทีม (teamwork) หมายถึง การทำงานร่วมกันของบุคคลตั้งแต่ 2 คนขึ้นไป เพื่อจุดมุ่งหมายร่วมกัน โดยแต่ละบุคคลต้องมีบทบาทชัดเจน มีการใช้ทรัพยากรร่วมกัน และสื่อสารกันอย่างมีประสิทธิภาพ เน้นที่กระบวนการ ไม่ได้เน้นที่ผลลัพธ์ (taskwork) ทีมทางการแพทย์มีความแตกต่างจากองค์กรอื่นๆ คือ บุคลากรในทีมมีความหลากหลายในวิชาชีพ (multidisciplinary team) ที่ทำงานภายใต้งานที่ซับซ้อนและมีการเปลี่ยนแปลงอยู่ตลอดเวลา มีความกดดันจากเวลาและสภาวะของผู้ป่วยเข้ามาเกี่ยวข้อง ทีมมักเป็นลักษณะของทีมเฉพาะกิจ ที่ต้องสามารถทำงานร่วมกันได้<sup>3</sup>

### หลักการของการประเมินการทำงานเป็นทีม<sup>2</sup>

การประเมินการทำงานเป็นทีมมักพิจารณาจาก การปฏิบัติงานของทีม (team performance) ซึ่งเป็น dynamic process เป็นผลรวมของทั้ง teamwork และ taskwork หรืออาจพิจารณาจากประสิทธิภาพของการทำงานเป็นทีม (team effectiveness) ซึ่งเป็นการประเมิน team performance ให้เข้ากับเกณฑ์การพิจารณาบางอย่าง เช่น ความพึงพอใจของบุคลากรภายในทีม การใช้ทรัพยากรอย่างเหมาะสม หรือ จำนวนความผิดพลาดที่เกิดขึ้นกับผู้ป่วย เป็นต้น ในที่นี้จะเน้นที่การประเมิน team performance ซึ่งมักมีวัตถุประสงค์หลักเพื่อการให้ feedback พฤติกรรมเพื่อการพัฒนา และการประเมินเพื่อดูประสิทธิภาพของการทำงานเป็นทีมหลังจากที่มีการเรียนรู้ฝึกฝนการทำงานร่วมกัน

องค์ประกอบของการประเมิน team performance

1. สามารถตรวจจับและแยกแยะลักษณะของพฤติกรรมที่เกิดขึ้น ทั้งในลักษณะของตัวบุคคลและในระดับทีมได้ (teamwork and taskwork, multilevel of performance)
2. เป็นการประเมินในสถานการณ์ที่มีความจำเพาะและมีบริบทที่ชัดเจน
3. ควรเป็น competency based ที่มุ่งเน้นให้เกิดการเปลี่ยนแปลงพฤติกรรม
4. มีรายละเอียดของพฤติกรรมนั้น
5. ควรมาจากหลายๆแหล่งประเมิน

6. สามารถสังเกตการณ์ทำงานของทีมที่มีการเปลี่ยนแปลงแบบ dynamic และเทียบกับพฤติกรรมก่อนหน้าได้

### เราควรประเมินอะไรบ้าง

การประเมินที่ดีควรทำควบคู่ไปกับการเรียนการสอนที่มีประสิทธิภาพ จึงควรเริ่มต้นที่การกำหนดเป้าหมายและวัตถุประสงค์ของการเรียนการสอน และพฤติกรรมที่อยากให้องค์กรเกิดการเปลี่ยนแปลง แล้วจึงมาพัฒนาแบบประเมินที่เน้นการประเมินพฤติกรรมที่ฝึกฝนได้และสังเกตได้ ผู้ประเมินจำเป็นต้องมีความรู้ในสิ่งที่จะทำการประเมิน ร่วมกับมีการฝึกการสังเกตพฤติกรรมและทำการประเมินจนมีความเชี่ยวชาญ โดยอาจพิจารณาสิ่งที่จะควรประเมินตาม teamwork competency ซึ่งมีองค์ประกอบตาม KSAs (knowledge, skill และ attitude) ดังนี้

1. Team knowledge competencies ได้แก่ การที่ทีมมีการ shared mental models ระหว่างบุคคลในทีม โดยทุกคนในทีม รับรู้บทบาทหน้าที่ความรับผิดชอบของกันและกัน
2. Team skill competencies ได้แก่ ทักษะต่างๆ ที่ควรได้รับการฝึกฝน เพื่อการทำงานเป็นทีมที่มีประสิทธิภาพ เช่น

#### a. ทักษะการเป็นผู้นำ (team leadership)

หน้าที่หลักของหัวหน้าทีม คือ การกำหนดเป้าหมายที่ชัดเจนในการดูแลผู้ช่วย และวางแผนจัดการและประสานงานต่างๆ ที่เกิดขึ้นให้กับลูกทีม โดย

- จัดลำดับความสำคัญของงาน และวางแผนการทำงาน
- ให้การตัดสินใจในกิจกรรมต่างๆ ที่เกิดขึ้น
- แบ่งงานภายในทีมตามความสามารถของแต่ละบุคคลอย่างเท่าเทียม
- ใช้ทรัพยากรที่มีอยู่ทั้งสิ่งของและมนุษย์อย่างคุ้มค่า
- ติดตามการทำงานที่เกิดขึ้น และพร้อมปรับเปลี่ยนแผนงานเพื่อให้บรรลุเป้าหมาย
- แก้ปัญหาความขัดแย้งที่อาจเกิดขึ้นภายในทีม

#### b. ทักษะการสื่อสาร (communication skill)

เป็นทักษะที่เกิดขึ้นตลอดการทำงานในทีม การประเมินลักษณะการสื่อสารอาจทำได้โดย

- ประเมินลักษณะของข้อความที่ใช้สื่อสาร ว่ามีความกระชับ ชัดเจน และถูกต้องหรือไม่
- ประเมินทักษะเฉพาะของการสื่อสารแล้วแต่สถานการณ์ เช่น การส่งต่อข้อมูลโดยใช้ทักษะ ISBAR (I=introduction, S=situation, B=background, A=assessment, R=recommendation) หรือ การสื่อสารแบบ closed-loop communication เป็นต้น

#### c. การช่วยกันสังเกตการณ์การทำงานซึ่งกันและกัน (mutual performance monitoring) โดยหากมีงานส่วนใดในทีมที่มีปัญหา บุคลากรในทีมสามารถสังเกตปัญหานั้นและช่วยกันแก้ปัญหานั้นได้

#### d. การปรับตัว (adaptability) ทีมควรมีความสามารถในการปรับตัวให้เข้ากับบทบาทการทำงานที่มีการเปลี่ยนแปลงตลอดเวลา ตามการดำเนินโรคของผู้ป่วย และตามสภาพแวดล้อมและทรัพยากรที่มี

#### e. การแก้ปัญหาความขัดแย้ง (conflict management)

3. Team attitude competencies ได้แก่ การไว้ใจกันและกันภายในทีม (mutual trust) มีเจตคติที่ดีต่อการทำงานร่วมกัน เป็นต้น

จะเห็นได้ว่า ควรมีการเลือกประเด็นมาใช้ในการประเมิน จากจุดประสงค์ของการประเมินเป็นสำคัญ เป็นการยากที่จะประเมินทุกองค์ประกอบได้ในสถานการณ์เดียว และในขณะเดียวกัน ในสถานการณ์เดียวกันมีจุดที่ประเมินใน

หัวข้อหนึ่งๆได้มากกว่า 1 ครั้ง จึงจำเป็นอย่างยิ่งที่ผู้ทำการประเมินจะต้องเข้าใจบริบทของแบบประเมิน และสถานการณ์ที่จะประเมินเป็นอย่างดี

### เราสามารถประเมินการทำงานเป็นทีมได้ทีเดียวบ้าง

การทำงานเป็นทีมสามารถประเมินได้ทั้งในระหว่างการทำงาน และในห้องเรียนสถานการณ์จำลอง หากแต่การประเมินในห้องเรียนสถานการณ์จำลอง เป็นการประเมินที่ได้รับความนิยมมากกว่าเนื่องจาก ผู้ประเมินสามารถกำหนดสถานการณ์และวางแผนสถานการณ์ให้เกิดพฤติกรรมที่ต้องการได้ สามารถประเมินได้ซ้ำๆในสถานการณ์เดียวกัน โดยไม่มีความเสี่ยงต่อผู้ป่วย สามารถบันทึกสถานการณ์เพื่อนำเอามาประเมินซ้ำ เป็นการประเมินพฤติกรรมที่คาดหวัง maximal performance ของทีม ในขณะที่การประเมินในสถานการณ์จริง เป็นการประเมินพฤติกรรมในบริบทที่เกิดขึ้นจริง ซึ่งมีประโยชน์อย่างยิ่งในการนำไปพัฒนาการทำงานโดยเฉพาะในเชิงการพัฒนาคุณภาพของหน่วยงาน หากแต่การประเมินทำได้ค่อนข้างยาก ทั้งจากลักษณะงานทางการแพทย์ที่มีความซับซ้อน ภาวะวิกฤติที่อาจเกิดขึ้นกับผู้ป่วย หรืออัตรากำลังคนเมื่อเทียบกับจำนวนผู้ป่วยในวันที่จะทำการประเมิน เป็นต้น อาจพิจารณาการประเมินให้เป็นไปในลักษณะ hybrid approach กล่าวคือ การทำสถานการณ์จำลองในที่ทำงานจริง ซึ่งใช้ความสมจริงของสถานที่ และใช้หุ่นจำลองมาแทนผู้ป่วย

### เราจะประเมินการทำงานเป็นทีมได้อย่างไร

การประเมินการทำงานเป็นทีมมักทำใน 2 ลักษณะ คือ การประเมินด้วยตัวเอง และการประเมินจากการสังเกตพฤติกรรม

#### การประเมินด้วยตัวเอง

สามารถทำได้โดยการใช้แบบสอบถาม และบุคคลในทีมให้ประเมินตนเองตามหัวข้อต่างๆทั้งที่เป็นตัวเอง ทีมโดยรวม หรือระบบทั้งหมด มีประโยชน์ในแง่การประเมิน team competency ในส่วนที่เป็น knowledge และ attitude การประเมินโดยวิธีนี้มีข้อจำกัดอยู่มาก ทั้งจากการประเมินที่มีลักษณะ subjective ค่อนข้างมาก ผู้ประเมินที่ไม่เข้าใจหลักการจริงๆของการทำงานเป็นทีม และมีแนวโน้มที่ผู้ฝึกปฏิบัติในระดับต้นจะให้คะแนนตัวเองสูง ปัจจุบันมีแบบประเมินที่ได้รับความนิยมใช้กันแพร่หลาย เช่น TeamSTEPPS Teamwork Attitudes Questionnaire<sup>3</sup>, the Mayo High Performance Team-work Scale<sup>4</sup> เป็นต้น

#### การประเมินโดยการสังเกตพฤติกรรม

เป็นรูปแบบหลักที่ใช้ในการประเมินการทำงานเป็นทีม โดยอาศัยผู้ประเมินที่ได้รับการฝึกฝน และแบบประเมินที่มีมาตรฐาน หรือมีการหาค่าความเที่ยงตรงและความเชื่อมั่นของเครื่องมือ โดยพิจารณาว่าจะประเมินในลักษณะรายบุคคล หรือเป็นทีม มีการประเมินที่ทั้งส่วนที่เป็น teamwork และ taskwork การประเมินโดยการสังเกตสามารถทำได้หลายวิธี ดังนี้

1. ประเมินเป็น rating scale โดยการสังเกตพฤติกรรมนั้นๆ แล้วประเมินเป็นลำดับคะแนน ซึ่งมักมีลักษณะเป็น rubric scale ที่มีการอธิบายความกำกับในแต่ละลำดับคะแนนชัดเจน เช่นใน Ottawa Global Rating Scale<sup>5</sup>, Observational Teamwork Assessment for Surgery (OTAS)<sup>6</sup>, Anesthetists' Non-Technical Skill (ANTS)<sup>7</sup> เป็นต้น

2. ประเมินด้วย checklist เป็นการประเมินเพื่อดูว่าเกิดพฤติกรรมนั้นๆขึ้นในสถานการณ์หรือไม่ เป็นลักษณะการประเมินที่ได้รับความนิยมน้อยกว่า เนื่องจากการทำงานเป็นทีม มีคุณภาพของพฤติกรรมเข้ามาเกี่ยวข้อง ตัวอย่างแบบประเมินแบบ checklist เช่น Ottawa Checklist Crisis Resource Management<sup>5</sup>
3. ประเมินด้วยแบบประเมินที่จำเพาะกับเหตุการณ์ (Event-based tools)<sup>6</sup> เป็นการสร้างแบบประเมินที่คาดเดาว่า เหตุการณ์ลักษณะนี้จะเกิดพฤติกรรมใดขึ้นบ้าง มีข้อดีคือ สามารถสังเกตพฤติกรรมนั้นๆได้ง่ายและชัดเจน ลดปัญหาความคลาดเคลื่อนจากผู้ประเมิน หากแต่แบบประเมินลักษณะนี้มีความจำเพาะต่อสถานการณ์นั้นๆค่อนข้างมาก ต้องสร้างแบบประเมินใหม่เมื่อเปลี่ยนสถานการณ์

กล่าวโดยสรุป การจะประเมินการทำงานเป็นทีมให้มีประสิทธิภาพ ควรทำควบคู่ไปกับการฝึกอบรมทักษะการทำงานเป็นทีม ที่มีวัตถุประสงค์ที่ชัดเจนและมีความจำเพาะกับบริบทของตนเอง การประเมินนี้เป็นการประเมินพฤติกรรม สามารถทำได้ทั้งในสถานการณ์จริงและในสถานการณ์จำลอง ประเมินเป็นรายบุคคลหรือเป็นทีม โดยต้องมีเครื่องมือประเมินที่มีความเที่ยงตรงและแม่นยำเพียงพอ ผู้ประเมินต้องได้รับการฝึกฝนให้เข้าใจหลักการของการทำงานเป็นทีม สามารถสังเกตพฤติกรรมในสถานการณ์ และเข้าใจแบบประเมินที่จะนำมาใช้ให้มากที่สุด

#### เอกสารอ้างอิง

1. Rhona Flin POC, Margaret Crichton. Safety at the sharp end: a guide to non-technical skills. Hamshire, England: Ashgate Publishing Limited; 2008.
2. Salas E, Frush K. Improving patient safety through teamwork and team training: Oxford University Press; 2012.
3. Baker DP, Amodeo AM, Krokos KJ, Slonim A, Herrera H. Assessing teamwork attitudes in healthcare: development of the TeamSTEPPS teamwork attitudes questionnaire. Qual Saf Health Care. 2010;19(6):e49.
4. Malec JF, Torsher LC, Dunn WF, Wiegmann DA, Arnold JJ, Brown DA, et al. The mayo high performance teamwork scale: reliability and validity for evaluating key crew resource management skills. Simul Healthc. 2007;2(1):4-10.
5. Kim J, Neilipovitz D, Cardinal P, Chiu M. A comparison of global rating scale and checklist scores in the validation of an evaluation tool to assess performance in the resuscitation of critically ill patients during simulated emergencies (abbreviated as "CRM simulator study IB"). Simul Healthc. 2009;4(1):6-16.

6. Sevdalis N, Lyons M, Healey AN, Undre S, Darzi A, Vincent CA. Observational teamwork assessment for surgery: construct validation with expert versus novice raters. *Ann Surg.* 2009;249(6):1047-51.
7. Fletcher G, Flin R, McGeorge P, Glavin R, Maran N, Patey R. Anaesthetists' Non-Technical Skills (ANTS): evaluation of a behavioural marker system. *Br J Anaesth.* 2003;90(5):580-8.
8. Rosen MA, Salas E, Wu TS, Silvestri S, Lazzara EH, Lyons R, et al. Promoting teamwork: an event-based approach to simulation-based teamwork training for emergency medicine residents. *Acad Emerg Med.* 2008;15(11):1190-8.



## Ottawa Crisis Resource Management Checklist

ACTION	YES (2 points)	With Prompting (1 point)	NO (0 points)
<b>PROBLEM SOLVING</b>			
Prompt ABC assessment			
Implements concurrent management approach (4 points)			
<b>SITUATIONAL AWARENESS</b>			
Avoids fixation error (4 points)			
Re-assesses and re-evaluates situation (4 points)			
<b>RESOURCE UTILIZATION</b>			
Calls for help when indicated			
Delegates and directs appropriately			
<b>LEADERSHIP</b>			
Maintains calm demeanor			
Acts decisively and maintains control of crisis			
Maintains global perspective			
<b>COMMUNICATION</b>			
Communicates clearly and concisely			
Closes the loop and uses names			
Listens to team input			
<b>TOTAL SCORE (30 points)</b>			

Resident #:

Scenario #:

Staff #:

Date:



# TeamSTEPPS



## Team Performance Observation Tool

Date: \_\_\_\_\_  
 Unit: \_\_\_\_\_  
 Team: \_\_\_\_\_  
 Shift: \_\_\_\_\_

**Rating Scale**  
 (circle 1)  
 Please comment  
 if 1 or 2

1 = Very Poor  
 2 = Poor  
 3 = Acceptable  
 4 = Good  
 5 = Excellent

<b>1. Team Structure</b>	<b>Rating</b>
a. Assembles a team	
b. Establishes a leader	
c. Identifies team goals and vision	
d. Assigns roles and responsibilities	
e. Holds team members accountable	
f. Actively shares information among team members	
Comments:	
<b>Overall Rating – Team Structure</b>	
<b>2. Leadership</b>	<b>Rating</b>
a. Utilizes resources efficiently to maximize team performance	
b. Balances workload within the team	
c. Delegates tasks or assignments, as appropriate	
d. Conducts briefs, huddles, and debriefs	
e. Empowers team members to speak freely and ask questions	
Comments:	
<b>Overall Rating – Leadership</b>	
<b>3. Situation Monitoring</b>	<b>Rating</b>
a. Includes patient/family in communication	
b. Cross monitors fellow team members	
c. Applies the STEP process when monitoring the situation	
d. Fosters communication to ensure team members have a shared mental model	
Comments:	
<b>Overall Rating – Situation Monitoring</b>	
<b>4. Mutual Support</b>	<b>Rating</b>
a. Provides task-related support	
b. Provides timely and constructive feedback to team members	
c. Effectively advocates for the patient	
d. Uses the Two-Challenge rule, CUS, and DESC script to resolve conflict	
e. Collaborates with team members	
Comments:	
<b>Overall Rating – Mutual Support</b>	
<b>5. Communication</b>	<b>Rating</b>
a. Coaching feedback routinely provided to team members, when appropriate	
b. Provides brief, clear, specific and timely information to team members	
c. Seeks information from all available sources	
d. Verifies information that is communicated	
e. Uses SBAR, call-outs, check-backs and handoff techniques to communicate effectively with team members	
Comments:	
<b>Overall Rating – Communication</b>	
<b>TEAM PERFORMANCE RATING</b>	

## ระดับการประเมินในระบบ non-technical สำหรับวิสัญญีแพทย์

ตารางด้านล่างนี้ใช้สำหรับประเมินทักษะ non-technical โดยการสังเกตพฤติกรรม หากไม่พบพฤติกรรมของหัวข้อประเมินในสถานการณ์นั้นๆ ให้ใส่คำว่า “ไม่สามารถประเมินได้”

### ระดับการประเมินในระบบ ANTS

ระดับคะแนน	คำอธิบาย
4 - ดี	การกระทำแสดงถึงการทำงานที่มีมาตรฐานสูง ส่งเสริมให้เกิดความปลอดภัยแก่ผู้ป่วย และเป็นแบบอย่างที่ดีต่อบุคคลอื่น
3 - พอใช้	การกระทำอยู่ในระดับมาตรฐานที่น่าพอใจ แต่ยังสามารถพัฒนาได้อีก
2 - ต้องปรับปรุง	การกระทำอยู่ในเกณฑ์น่าเป็นห่วง ต้องการการพัฒนา
1 - ต้องปรับปรุงอย่างมาก	การกระทำเป็นอันตรายหรือมีผลต่อความปลอดภัยของผู้ป่วย จำเป็นอย่างยิ่งที่จะต้องได้รับการแก้ไข
N - ไม่สามารถประเมินได้	สถานการณ์นี้ไม่สามารถสังเกตทักษะนี้ได้

หมวดหมู่	องค์ประกอบ	คะแนน*	พฤติกรรมที่สังเกตได้	คะแนนในหมวดหมู่ และบันทึกสำหรับการอภิปราย
การจัดการกับงาน	การวางแผนและการเตรียมตัว			
	การลำดับความสำคัญก่อนหลัง			
	การคงไว้ซึ่งมาตรฐานวิชาชีพ			
	การจัดทำและใช้ทรัพยากร			
การทำงานเป็นทีม	การประสานงานภายในทีม			
	การแลกเปลี่ยนข้อมูลภายในทีม			
	การใช้อำนาจรับผิดชอบและ กล้าแสดงความคิดเห็นอย่างสร้างสรรค์			
	การประเมินความสามารถของผู้ร่วมงาน			
	การให้ความช่วยเหลือผู้อื่น			
	การรวบรวมข้อมูล			
การตระหนักในสถานการณ์	การรับรู้และเข้าใจสถานการณ์			
	การวางแผนล่วงหน้าสำหรับภาวะฉุกเฉินที่ อาจเกิดขึ้น			
	การพิจารณาตัวเลือกในการตัดสินใจ			
การตัดสินใจ	การประเมินความเสี่ยงและวิเคราะห์ทางเลือก			
	การประเมินสถานการณ์ซ้ำ			

\* 4 ตี, 3 พอใช้, 2 ต้องปรับปรุงอย่างมาก, 1 ไม่สามารถประเมินได้

**แบบประเมินการจัดการในภาวะวิกฤติ Ottawa Global Rating Scale**

**เกณฑ์ในการประเมิน**

การประเมินนี้ เน้นการประเมินศักยภาพในการจัดการกับสถานการณ์ฉุกเฉิน และการดูแลผู้ป่วยที่อยู่ในภาวะวิกฤติ โดยผู้ที่ได้รับการประเมินควรมีมาตรฐานพื้นฐานในการดูแลผู้ป่วยเป็นอย่างดีมาก่อน เช่น เป็นแพทย์ประจำบ้านอาวุโสที่ผ่านการดูแลผู้ป่วยในหออภิบาล รวมไปถึงแพทย์เฉพาะทางอาวุโสที่มีประสบการณ์การจัดการสถานการณ์วิกฤติ นอกจากนี้ ความรู้ที่ใช้ในการดูแลผู้ป่วยจะได้รับการประเมินร่วมด้วย หากแต่การประเมินนี้จะมุ่งเน้นที่ทักษะการจัดการกับสถานการณ์ฉุกเฉิน ทักษะด้านกลังนี้เป็นหัวใจสำคัญในการดูแลผู้ป่วยในภาวะวิกฤติ ซึ่งจะได้รับการประเมินในระหว่างที่ฝึกปฏิบัติในสถานการณ์จำลอง

**ทักษะความเป็นผู้นำ**

- อยู่ในความสงบและควบคุมสถานการณ์ได้
- ตัดสินใจได้รวดเร็ว ฉับไว
- มองเห็นภาพรวมของสถานการณ์

**การตระหนักรู้สถานการณ์**

- หลีกเลี่ยงความผิดพลาดจากการยึดมั่นในความคิด (fixation error)
- ประเมิน และทบทวนสถานการณ์ซ้ำ
- คาดหมายเหตุการณ์ที่อาจเกิดขึ้น

**ทักษะการสื่อสาร**

- สื่อสารได้กระชับและชัดเจน
- ใช้คำพูดและการกระทำอื่นใดเพื่อการสื่อสารได้ตรงประเด็น
- รับฟังความเห็นของบุคคลอื่นในทีม

**ทักษะการแก้ปัญหา**

- ความรอบคอบและรู้ขั้นตอนก่อนหลัง (ABC)
- ความรวดเร็วในการปฏิบัติงาน (การทำงานหลายอย่างพร้อมกัน)
- มีแผนการรองรับในภาวะวิกฤติ

**การใช้ทรัพยากร**

- ขอความช่วยเหลือได้เหมาะสมทันเวลา
- เรียกใช้อุปกรณ์ได้เหมาะสม
- จัดลำดับงานก่อนหลังได้เหมาะสม

**สรุปภาพรวม**

สรุปภาพรวมการปฏิบัติงาน

1	2	3	4	5	6	7
• มือใหม่ ที่ทักษะทุกด้านต้องการพัฒนาอย่างมาก	• มือใหม่ ที่มีประสบการณ์ ทักษะหลายอย่างต้องการพัฒนา	• มือใหม่ ที่มีประสบการณ์ ทักษะหลายอย่างต้องการพัฒนา	• มือใหม่ ที่มีประสบการณ์ ทักษะหลายอย่างต้องการพัฒนา	• มือใหม่ ที่มีประสบการณ์ ทักษะส่วนใหญ่ต้องการพัฒนาเล็กน้อย	• มือใหม่ ที่มีประสบการณ์ ทักษะส่วนใหญ่ต้องการพัฒนาเล็กน้อย	• มือใหม่ ที่มีประสบการณ์ ทักษะส่วนใหญ่ต้องการพัฒนาเพียงเล็กน้อยในบางทักษะ

1. ทักษะความเป็นผู้นำ

1	2	3	4	5	6	7
• เกิดความตระหนัก ควบคุมตัวเองไม่ได้ ไม่สามารถตัดสินใจ ไม่สามารถมองเห็นภาพรวมของสถานการณ์	• ควบคุมตัวเองไม่ได้บ่อยครั้ง ตัดสินใจช้า มองภาพรวมได้ไม่ชัดเจน	• ควบคุมตัวเองได้บ่อยครั้ง ตัดสินใจช้า มองภาพรวมของสถานการณ์ได้เป็นส่วนใหญ	• ควบคุมสติได้เป็นส่วนใหญ่ ตัดสินใจได้ถูกต้องแม้จะเข้าไปบ้าง มองภาพรวมของสถานการณ์ได้เป็นส่วนใหญ	• ควบคุมสติได้ตลอดเวลา ตัดสินใจได้ถูกต้องแม้จะรวดเร็วบ้าง มองภาพรวมของสถานการณ์ได้โดยตลอด	• ควบคุมสติได้ตลอดเวลา ตัดสินใจได้ถูกต้องแม้จะรวดเร็วบ้าง มองภาพรวมของสถานการณ์ได้โดยตลอด	• ควบคุมสติได้ตลอดเวลา ตัดสินใจได้ถูกต้องแม้จะรวดเร็วบ้าง มองภาพรวมของสถานการณ์ได้โดยตลอด

2. ทักษะการแก้ปัญหา

1	2	3	4	5	6	7
• ในการแก้ปัญหาไม่สามารถ ประเมินสถานการณ์ตามขั้น ตอน ABC ไม่แก้ปัญหาหลายอย่างพร้อม กันได้ แม้จะมีสิ่งชี้ นำ ไม่มีแผนสำรองเมื่อเกิดติดขัด	• ประเมินสถานการณ์ตามขั้น ตอนได้ช้าหรือไม่ครบถ้วน ยังไม่สามารถแก้ปัญหาหลาย อย่างพร้อมกันได้ทันที นึกถึงแผนสำรองบ้างเมื่อเกิด การติดขัด	• ประเมินสถานการณ์ตามขั้น ตอนได้ช้าหรือไม่ครบถ้วน ยังไม่สามารถแก้ปัญหาหลาย อย่างพร้อมกันได้ มีแผนสำรองเมื่อเกิดการติดขัด	• ประเมินสถานการณ์ตามขั้น ตอนได้ดี สามารถแก้ไขปัญหาหลาย อย่างพร้อมกันได้ มีแผนสำรองเมื่อเกิดการติดขัด	• ประเมินสถานการณ์ตามขั้น ตอนได้ดี สามารถแก้ไขปัญหาหลาย อย่างพร้อมกันได้ และคำนึงถึงแผนสำรองทันทีที่เกิด การติดขัด	• ประเมินสถานการณ์ตามขั้น ตอนได้ดี สามารถแก้ไขปัญหาหลาย อย่างพร้อมกันได้ และคำนึงถึงแผนสำรองทันทีที่เกิด การติดขัด	• ประเมินสถานการณ์ตามขั้น ตอนได้ดี สามารถแก้ไขปัญหาหลาย อย่างพร้อมกันได้ และคำนึงถึงแผนสำรองทันทีที่เกิด การติดขัด

3. การตระหนักรู้ในสถานการณ์/ความรู้เท่าทันสถานการณ์

1	2	3	4	5	6	7
<ul style="list-style-type: none"> <li>ติดขัดในการวิเคราะห์สถานการณ์ แม้จะมีสิ่งกระตุ้นชี้แนะ ไม่สามารถระบุการใด ๆ ลงในหน้าในสถานการณ์ได้</li> </ul>	<ul style="list-style-type: none"> <li>หลีกเลี่ยง fixation error เฉพาะเมื่อมีสิ่งกระตุ้นชี้แนะ ไม่สามารถประเมินสถานการณ์ซ้ำ หากไม่มีตัวชี้แนะ การกระทำที่การลงหน้าต่าง ๆ เป็นไปได้อย่าง</li> </ul>	<ul style="list-style-type: none"> <li>หลีกเลี่ยง fixation error ได้ดีพอควร สามารถประเมินสถานการณ์ซ้ำได้ด้วยสิ่งกระตุ้นชี้แนะเพียงเล็กน้อย</li> </ul>	<ul style="list-style-type: none"> <li>หลีกเลี่ยง fixation error สามารถระบุการต่าง ๆ ลงหน้าได้ เป็นส่วนใหญ่</li> </ul>	<ul style="list-style-type: none"> <li>หลีกเลี่ยง fixation error ได้ดี มีการประเมินสถานการณ์ซ้ำได้อย่างต่อเนื่อง โดยปราศจากสิ่งกระตุ้น</li> </ul>	<ul style="list-style-type: none"> <li>สามารถระบุการต่าง ๆ ลงหน้าได้</li> </ul>	<ul style="list-style-type: none"> <li>หลีกเลี่ยง fixation error ได้ดี มีการประเมินสถานการณ์ซ้ำได้อย่างต่อเนื่อง โดยปราศจากสิ่งกระตุ้น สามารถระบุการต่าง ๆ ลงหน้าได้เป็นอย่างดี</li> </ul>

4. ทักษะการใช้ทรัพยากร

1	2	3	4	5	6	7
<ul style="list-style-type: none"> <li>ไม่สามารถบริหารบุคลากรและทรัพยากรได้</li> <li>ไม่มีการจัดลำดับการทำงานก่อนหลังหรือ ขอความช่วยเหลือได้ทันที</li> </ul>	<ul style="list-style-type: none"> <li>สามารถบริหารบุคลากรและทรัพยากรได้ แต่ยังไม่ประสิทธิภาพ จัดลำดับการทำงานหรือขอความช่วยเหลือได้เฉพาะเมื่อมีสิ่งชี้แนะ</li> </ul>	<ul style="list-style-type: none"> <li>สามารถบริหารบุคลากรและทรัพยากรได้ดีปานกลาง สามารถจัดลำดับการทำงานและ/หรือขอความช่วยเหลือได้เมื่อมีสิ่งกระตุ้นชี้แนะเล็กน้อย</li> </ul>	<ul style="list-style-type: none"> <li>สามารถบริหารบุคลากรและทรัพยากรได้ดีปานกลาง สามารถจัดลำดับการทำงานและ/หรือขอความช่วยเหลือได้เมื่อมีสิ่งกระตุ้นชี้แนะเล็กน้อย</li> </ul>	<ul style="list-style-type: none"> <li>สามารถบริหารบุคลากรและทรัพยากรได้อย่างมีประสิทธิภาพ สามารถจัดลำดับการทำงานและขอความช่วยเหลือได้เมื่อมีสิ่งกระตุ้นชี้แนะ</li> </ul>	<ul style="list-style-type: none"> <li>สามารถบริหารบุคลากรและทรัพยากรได้อย่างมีประสิทธิภาพ สามารถจัดลำดับการทำงานและขอความช่วยเหลือได้เมื่อมีสิ่งกระตุ้นชี้แนะ</li> </ul>	<ul style="list-style-type: none"> <li>สามารถบริหารบุคลากรและทรัพยากรได้อย่างมีประสิทธิภาพ สามารถจัดลำดับการทำงานและขอความช่วยเหลือได้เมื่อมีสิ่งกระตุ้นชี้แนะ</li> </ul>

5. ทักษะการสื่อสาร

1	2	3	4	5	6	7
<ul style="list-style-type: none"> <li>ไม่สื่อสารกับบุคลากรอื่น รวมถึงไม่รู้ว่ามีบุคลากรอื่นสื่อสารด้วย</li> <li>ไม่ใช้การสื่อสารที่ทั้งภาษาพูดโดยตรงและภาษาภายใน</li> </ul>	<ul style="list-style-type: none"> <li>สื่อสารกับบุคลากรอื่นบ้าง แต่ข้อมูลที่ใช้ไม่ชัดเจน พังความรอบข้างแต่ยังไม่โต้ตอบไม่ได้</li> </ul>	<ul style="list-style-type: none"> <li>สื่อสารกับบุคลากรอื่นบ้าง แต่ข้อมูลที่ใช้ไม่ชัดเจน พังความรอบข้างแต่ยังไม่โต้ตอบไม่ได้</li> </ul>	<ul style="list-style-type: none"> <li>สื่อสารกับบุคลากรอื่นได้ดีเป็นส่วนใหญ่ รับฟังความรอบข้าง สามารถใช้ภาษาพูดและภาษาภายในได้เป็นส่วนใหญ</li> </ul>	<ul style="list-style-type: none"> <li>สื่อสารกับบุคลากรอื่นได้ดีเป็นส่วนใหญ่ รับฟังความรอบข้าง สามารถใช้ภาษาพูดและภาษาภายในได้เป็นส่วนใหญ</li> </ul>	<ul style="list-style-type: none"> <li>สื่อสารกับบุคลากรอื่นได้ดีเป็นส่วนใหญ่ รับฟังความรอบข้าง สามารถใช้ภาษาพูดและภาษาภายในได้เป็นส่วนใหญ</li> </ul>	<ul style="list-style-type: none"> <li>สื่อสารกับบุคลากรอื่นได้ชัดเจนและถูกต้องเวลา กระตุ้นการสื่อสารให้เกิดขึ้นภายในทีมใช้ภาษาพูดและภาษาภายในได้ดีตลอดเวลา</li> </ul>

Table 2  
Observational checklist for example scenario. 'Hits' are a dichotomous scoring of whether or not the targeted behavior was observed

Simulation Scenario: ED Resuscitation Teamwork Competency: Leadership			
Event	Critical Response	Hits	IG*
55-year-old male is visiting another ED patient and suddenly collapses. The resident is called into the room, but the RN begins yelling out orders.	The resident identifies him-/herself as the team leader. Establishes unresponsiveness. Opens the airway (jaw thrust and chin lift). Inspects for chest rise and fall. Listens for air movement from the mouth. Feels for a pulse. Resident calls for/activates a "CODE" to recruit more team members to help.		
The patient remains apneic and pulseless. The confederates just stand at the patient's bedside awaiting direction/orders.	Asks for the RN to put the patient on a cardiac monitor and for an IV to be secured. Asks for the ED tech to begin chest compressions.		
The ED RN states "Do you want me to put the patient on a monitor, put in an IV, or bag the patient? There's only one of me!"	Asks the RN to place the patient on a monitor first and then to work on securing an IV. Begins bagging the patient him-/herself.		
ED tech performs very shallow chest compressions at 40 per minute.	Educates the ED tech on how to perform appropriate chest compressions (identify the xiphoid process and lower third of sternum, interlock hands, perform chest compressions at a depth of approximately 1.5 to 2 inches at a rate of 100/minute).		
ED tech suddenly stops performing chest compressions because he or she is "too tired to continue." An ED tech arrives simultaneously. RN states that she is having a hard time maintaining a seal on the BVM.	Directs another team member to take over chest compressions. Directs the RN to take over BVM ventilation and asks the ED tech to prepare the intubation equipment (or vice versa).		
The patient is successfully intubated and the RN starts yelling at the ED tech and accusing him/her of dislodging the ETT in the process of securing it.	Inquires about the ETCO <sub>2</sub> reading. Identifies the alveolar waveform demonstrated on capnograph. Auscultates the chest and epigastric region. Reassesses the pulse oximetry reading and waveform.		
Once the patient is stabilized, an ECG is performed showing an STEMI in the anterior and lateral leads.	Resident summarizes the case, management rendered, and recommended course of action/catheterization to the cardiologist on-call (SBAR format).		

BVM = bag/valve/mask; ECG = electrocardiogram; ED = emergency department; ETCO<sub>2</sub> = end-tidal carbon dioxide; ETT = endotracheal tube; IV = intravenous line; RN = nurse; SBAR = situation-background-assessment-recommendation; STEMI = ST-segment elevation myocardial infarction.  
\*IG = instructor-guided, used when the behavior is observed, but was coached by a trainer.

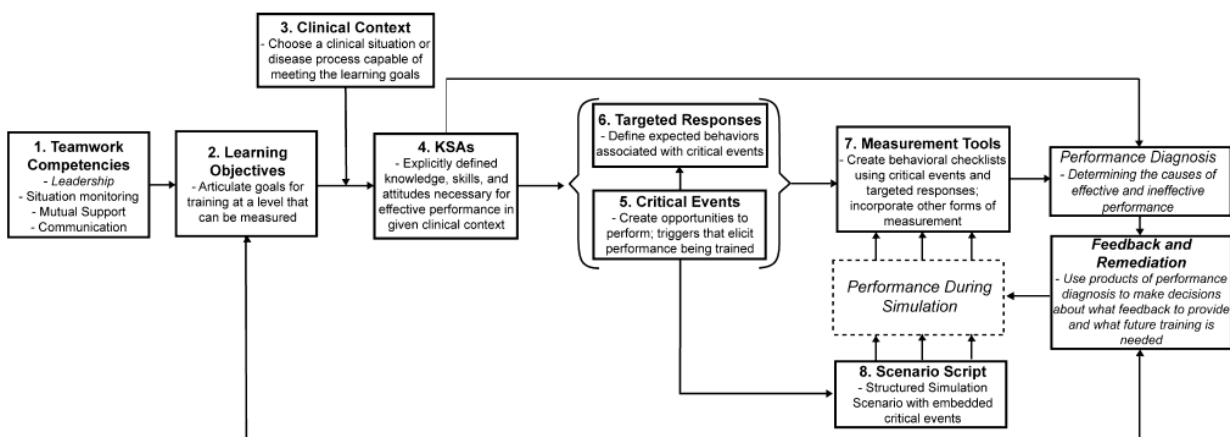


Figure 1. Overview of event-based approach to training (EBAT) process for teamwork training in emergency medicine (EM). KSA = knowledge, skills, and attitudes.



25 Apr 2018

หัวข้อ : Rubric scale development

**Rubric**

Lertbunnaphong T, M.D.  
Department of OBGYN  
Faculty of Medicine, Siriraj Hospital  
Mahidol University

---

---

---

---

---

---

---

---

**Scope**

**What is Rubric?**

**Why Rubric?**

**How to create Rubric?**

**Workshop for Rubric**

---

---

---

---

---

---

---

---

**Rubric = RED**  
An authoritative rule

---

---

---

---

---

---

---

---

**Coherent sets of criteria**  
**Description of levels of performance**

---

---

---

---

---

---

---

---

**Observation without judgment**

How appropriately?  
 How completely?  
 How well?

---

---

---

---

---

---

---

---


---

---

---

---

---

---

---

---

	NO	NO, but	Yes, but	YES

---

---

---

---

---

---

---

---

	Novice	Basic	Proficient	Expert

Target performance

---

---

---

---

---

---

---

---

	Novice	Basic	Proficient	Expert

Target performance

---

---

---

---

---

---

---

---

**Inference of description**  
**Low or High**

---

---

---

---

---

---

---

---

	 <b>Level 1</b> หัวเราะเบา ๆ	 <b>Level 2</b> หัวเราะคิกคัก	 <b>Level 3</b> หัวเราะปกติ	 <b>Level 4</b> หัวเราะลั่น
ความดัง	เสียงดัง ได้ยินเฉพาะคนที่นั่ง/ยืนใกล้ ๆ	เสียงดังปานกลาง ได้ยินเฉพาะคนในบริเวณใกล้เคียง ดูสุภาพ	เสียงดังมากจนทุกคนในห้องได้ยิน และ ดูไม่สุภาพ	ดังมากจนน่ารำคาญหรือทำให้คนอื่นต้องหันมอง
ระยะเวลา	สั้นมาก	หัวเราะเข้าไปมาได้เพียงรอบเดียว	หัวเราะเข้าไปไม่ได้นาน มีจังหวะขึ้น ๆ ลง ๆ สลับกัน	นานและคงที่จนคนอื่นต้องขอหรืออยากให้หยุด
การเคลื่อนไหวร่างกาย	ริมฝีปากอาจเปิดหรือปิด	ริมฝีปากเปิดออกหน้ายิ้ม	มีการขยับอวัยวะอื่น ๆ ที่นอกเหนือจากใบหน้า อย่างน้อยเพียง 1 ส่วน เช่น หมุนไหล่ หรือ พายศีรษะ	เคลื่อนไหวร่างกายทั้งตัว ดังนี้ หัวค้อม ร่างกายสั่นไปทั่ว หรือล้มลงกับพื้น

---

---

---

---

---

---

---

---

**“Lowest inference descriptors**  
**that you can use and still accomplish**  
**your purpose of assessing important qualities”**

---

---

---

---

---

---

---

---

**“Leaving descriptions open to professional judgment (some inferences) is better than locking things down with overly rigid descriptions”**

---

---

---

---

---

---

---

**“You match the performance to the description rather than “Judge” it”**

---

---

---

---

---

---

---

**Type of rubric**  
**Analytic VS holistic**  
**General VS task-specific**

---

---

---

---

---

---

---

**Analytic rubric**

	Poor	Average	Good	Excellent
Outcome A	description	description	description	description
Outcome B	description	description	description	description
Outcome C	description	description	description	description
Outcome D	description	description	description	description
Outcome E	description	description	description	description
Outcome F	description	description	description	description

---

---

---

---

---

---

---

---

**Analytic rubric**

**Feedback**  
**Formative**  
**Motivation**  
**Take more time**  
**(score, inter-rater reliability)**

---

---

---

---

---

---

---

---

**Holistic rubric**

	Performance
Outcome A	description
Outcome B	description
Outcome C	description
Outcome D	description
Outcome E	description
Outcome F	description

---

---

---

---

---

---

---

---

**Holistic rubric**  
Fast scoring  
Inter-rater reliability  
Summative evaluation  
Cannot feedback... how to improve!

---

---

---

---

---

---

---

**General rubric**  
General knowledge & skills  
different tasks with same learning outcomes

---

---

---

---

---

---

---

**General rubric**  
Focus on what learning outcomes  
Reusable with several tasks  
learners' self evaluation  
  
Lower reliability  
Require practice to apply well

---

---

---

---

---

---

---



**General rubric**  
Essay or reports  
General surgical skills  
Problem solving skills  
Teamwork & Communication

---

---

---

---

---

---

---

**Task specific rubric**  
Specific contents of knowledge & skills

---

---

---

---

---

---

---

**Task specific rubric**  
Scoring direction with lower inference  
High reliability  
  
Cannot share with students  
Not useful for formative assessment/feedback  
Need to write new rubrics for each task

---

---

---

---

---

---

---

**The most powerful aspect of Rubrics**  
**“to conceptualize their learning outcomes**  
**and to monitor their own progress”**

---

---

---

---

---

---

---

**Analytic - General rubric**

---

---

---

---

---

---

---

**Score weighing**

		Poor	Average	Good	Excellent
		30%	60%	80%	100%
10%	Outcome A	description	description	description	description
20%	Outcome B	description	description	description	description
20%	Outcome C	description	description	description	description
30%	Outcome D	description	description	description	description
10%	Outcome E	description	description	description	description
10%	Outcome F	description	description	description	description

---

---

---

---

---

---

---

**Scope**  
**Why Rubric?**

---

---

---

---

---

---

---

**Rubrics help teacher teach**

---

---

---

---

---

---

---

**Rubrics help teacher teach**  
**1.Focus on learning outcomes, not tasks/activities**  
**2.Coordinate instruction and assessment**  
**(doing task/feedback/practice/revise task/practice/grading)**  
**3.Honest on grading (even on bad day)**

---

---

---

---

---

---

---

**Rubrics help students learn**

---

---

---

---

---

---

---

**Rubrics help students learn**

- 1. Guidance for learning and self assessment  
(better quality work)
- 2. Improve student learning & success
- 3. Peer to peer feedback

---

---

---

---

---

---

---

**Misconceptions about Rubrics**

---

---

---

---

---

---

---

**Confusing learning outcomes  
with tasks/activities**  
  
neatness, color, handwriting

---

---

---

---

---

---

---

**Confusing Rubrics  
with requirements or quantities**  
  
Focus on elements or direction  
Use “checklists” before assignments

---

---

---

---

---

---

---

**Confusing Rubrics  
with evaluating rating scales**

---

---

---

---

---

---

---

**Rating scale and checklists**  
**Family of rubric**

---

---

---

---

---

---

---

---

**Rating scale**

	Poor (D)	Average (C)	Good (B)	Excellent (A)
Process/ outcome A	√			
Process/ outcome B		√		
Process/ outcome C		√		
Process/ outcome D			√	
Process/ outcome E		√		
Process/ outcome F				√

---

---

---

---

---

---

---

---

**Checklists**

Process A	√
Process B	√
Process C	√
Process D	√
Process E	
Process F	

---

---

---

---

---

---

---

---

**Scope**  
**How to create Rubric?**

---

---

---

---

---

---

---

---

**Steps to create Rubrics**


---

---

---


---

---

---

---

---

	<b>Good</b>	<b>Basic</b>	<b>Need improvement</b>
<b>Taste</b>			
<b>Creativity</b>			
<b>Appearance</b>			

---

---

---

---

---

---

---

---



	<b>Good</b>	<b>Basic</b>	<b>Need improvement</b>
<b>Taste</b>	1	2	1
<b>Creativity</b>	1	2	1
<b>Appearance</b>	1	2	1

---

---

---


---

---

---

---

---

	<b>Good</b>	<b>Basic</b>	<b>Need improvement</b>
<b>Taste</b>	เนื้อเค้กนุ่ม รสชาติดีกว่าที่เคยกินมา	เนื้อเค้กนุ่ม แต่รสชาติธรรมดา	เนื้อเค้กหยาบแห้ง รสชาติแย่กว่าที่เคยกินมา
<b>Creativity</b>	โดดเด่น ใส่ใจรายละเอียด ดูแปลกใหม่	ดูธรรมดา เหมือนเค้กทั่วไป	ไม่น่าสนใจ ไม่ได้ใส่ใจรายละเอียด
<b>Appearance</b>	น่ากิน มีสีสัน ตกแต่งสวยงาม	น่ากิน มีสีสัน แต่ยังไม่สวยงาม	สีไม่น่ากิน และการตกแต่งไม่สวยงาม

---

---

---

---

---

---

---

---

### How about my rubric? Obstetrics patient report

1. ความรู้ (20 คะแนน)	ดีมากเยี่ยม				พอใช้				จำเป็นต้องปรับปรุง			
	1	2	3	4	5	6	7	8	9	10	11	12
2. ความละเอียดของ รายละเอียด (10 คะแนน)	ดีเยี่ยม	ดี	พอใช้	ไม่พอใช้	ดีเยี่ยม	ดี	พอใช้	ไม่พอใช้	ดีเยี่ยม	ดี	พอใช้	ไม่พอใช้
3. ความละเอียดของ ข้อมูล (20 คะแนน)	ดีมากเยี่ยม	ดี	พอใช้	ไม่พอใช้	ดีมากเยี่ยม	ดี	พอใช้	ไม่พอใช้	ดีมากเยี่ยม	ดี	พอใช้	ไม่พอใช้
4. ความละเอียดของ การวิเคราะห์ (10 คะแนน)	ดีมากเยี่ยม	ดี	พอใช้	ไม่พอใช้	ดีมากเยี่ยม	ดี	พอใช้	ไม่พอใช้	ดีมากเยี่ยม	ดี	พอใช้	ไม่พอใช้
5. ความละเอียดของ การนำเสนอ (10 คะแนน)	ดีมากเยี่ยม	ดี	พอใช้	ไม่พอใช้	ดีมากเยี่ยม	ดี	พอใช้	ไม่พอใช้	ดีมากเยี่ยม	ดี	พอใช้	ไม่พอใช้

---

---

---

---

---

---

---

---

**Scope**  
**Workshop for Rubric**

---

---

---

---

---

---

---

**Quick Rubric**  
**[www.quickrubric.com](http://www.quickrubric.com)**

---

---

---

---

---

---

---

**iRubric**  
**[www.rcampus.com](http://www.rcampus.com)**

---

---

---

---

---

---

---

## การพัฒนาการประเมินด้วยรูบรีค (Rubrics)

ตรีภพ เลิศบรรณพงษ์

### คำจำกัดความ

Rubrics มีคำจำกัดความที่หลากหลาย เช่น

1. การประเมินที่เชื่อมโยงเป้าหมายการเรียนรู้กับงานที่ผู้เรียนได้รับมอบหมายด้วยการกำหนดเกณฑ์ การประเมินหรือบรรทัดฐาน และคำอธิบายตามระดับของสมรรถนะที่พึงประสงค์จากระดับที่ดีที่สุดจนถึงระดับที่แย่ที่สุด
2. ชุดของเกณฑ์การประเมินสำหรับงานที่ผู้เรียนได้รับมอบหมายซึ่งมีความสอดคล้องกับเป้าหมายการเรียนรู้ และมีความสอดคล้องตามระดับของสมรรถนะสอดคล้องกับเป้าหมายการเรียนรู้ที่กำหนด

### องค์ประกอบสำคัญของ rubrics

1. **เกณฑ์การประเมิน (criteria)** หมายถึง แง่มุมหรือประเด็นของเป้าหมายการเรียนรู้ที่ต้องการประเมิน ซึ่งต้องกำหนดให้มีความชัดเจนและเหมาะสม สามารถสังเกตได้ แต่ละเกณฑ์สะท้อนแง่มุมหรือประเด็นของเป้าหมายการเรียนรู้ที่แตกต่างกันและโดยภาพรวมสามารถครอบคลุมเป้าหมายการเรียนรู้ทั้งหมด รวมทั้งสามารถนำไปสู่คำอธิบายเพื่อสะท้อนสมรรถนะแต่ละระดับได้อย่างครอบคลุมและต่อเนื่อง
2. **ระดับของสมรรถนะ (level of performance)** หมายถึง ระดับการจำแนกความสามารถของผู้เรียนตามสมรรถนะที่แสดงออกหรือสามารถสังเกตได้ แนะนำให้จำแนกเป็น 4 ระดับ ดังนี้
  - 2.1. YES ผู้เรียนสามารถทำได้สำเร็จตามเป้าหมายการเรียนรู้อย่างสมบูรณ์
  - 2.2. YES, but ผู้เรียนสามารถทำได้สำเร็จตามเป้าหมายการเรียนรู้ แต่ยังมีจุดอ่อนที่ต้องพัฒนาบางอย่าง
  - 2.3. NO, but ผู้เรียนไม่สามารถทำได้สำเร็จตามเป้าหมายการเรียนรู้ แต่มีทักษะหรือทำกระบวนการบางอย่างได้
  - 2.4. NO ผู้เรียนไม่สามารถทำได้สำเร็จตามเป้าหมายการเรียนรู้ และขาดทักษะหรือกระบวนการที่เหมาะสม

อย่างไรก็ตาม สามารถจำแนกระดับของสมรรถนะเป็น 3 ระดับได้เช่นเดียวกัน เช่น ดี/ผ่าน/ไม่ผ่าน เป็นต้น แต่จะไม่มีการจำแนกเพียง 2 ระดับ เพราะจะกลายเป็น Checklists แทนที่จะเป็นรูบรีค และหากต้องการประเมินมากกว่า 4 ระดับ เช่น 5 หรือ 6 ระดับ ก็จะเป็นการยากที่จะจำแนกสมรรถนะแต่ละระดับได้อย่างชัดเจน ดังนั้นในทางปฏิบัติ การจำแนกระดับของสมรรถนะเป็น 3 หรือ 4 ระดับจึงพบเห็นได้ในการประเมินด้วยรูบรีคเป็นส่วนใหญ่

3. **คำอธิบายสมรรถนะตามระดับความสามารถ (Description of performance)** ต้องเขียนให้เข้าใจอย่างชัดเจนจนสามารถสังเกตได้ ตามลำดับจากมากไปหาน้อย หรือ น้อยไปหามาก โดยผู้ประเมินต้องสามารถแยกแยะความแตกต่างระหว่างระดับของสมรรถนะในแต่ละเกณฑ์ รวมทั้งการกำหนดมาตรฐานของสมรรถนะที่ยอมรับได้ของผู้เรียนไว้อย่างเหมาะสม ในทางปฏิบัติ การกำหนดคำอธิบายมีอาจทำให้เกิดความกระจ่างได้อย่างสมบูรณ์ แต่ต้องพยายามให้มีการอนุมานหรือตีความน้อยที่สุด และมีความกระจ่างมากพอที่จะสะท้อนเป้าหมายการประเมิน อย่างไรก็ตามคำอธิบายที่ดีต้องยอมให้มีการใช้วิจารณญาณของผู้ประเมินบ้าง เท่าที่จำเป็น

### รูปแบบของสมรรถนะที่สามารถประเมินได้ด้วย rubrics

1. **กระบวนการเรียน (processes)** เช่น การนำเสนอหน้าชั้นเรียน พฤติกรรมขณะปฏิบัติงาน ภาวะผู้นำ การทำงานเป็นทีม ความสามารถในการตรวจหรือประเมินผู้ป่วย ความสามารถในการใช้เครื่องมือต่าง ๆ เป็นต้น
2. **ผลผลิตของการเรียน (products)** เช่น รายงานผู้ป่วย รายงานวิจัย แฟ้มสะสมผลงาน เป็นต้น

## ประเภทของ rubrics

### 1. Holistic and Analytic rubrics

- 1.1. **Holistic rubrics** คือ การประเมินไม่มีการแยกระดับของสมรรถนะในเกณฑ์การประเมินเป้าหมายการเรียนรู้ แต่ใช้คำอธิบายภาพรวมของแต่ละเกณฑ์เพียงอย่างเดียว
- 1.2. **Analytic rubrics** คือ การประเมินที่แยกระดับของสมรรถนะในเกณฑ์เป้าหมายการเรียนรู้จากดีที่สุดจนถึงแย่มากที่สุดด้วยคำอธิบายสมรรถนะตามระดับความสามารถได้อย่างต่อเนื่อง

### 2. General and task-specific rubrics

- 2.1. **General rubrics** คือ การประเมินที่เน้นความรู้และทักษะทั่วไป สามารถใช้ซ้ำได้ในสมรรถนะที่มีเป้าหมายการเรียนรู้เหมือนกัน เช่น การเขียนรายงาน การทำงานเป็นทีม ทักษะการแก้ปัญหา ทักษะการผ่าตัดทั่วไป เป็นต้น
- 2.2. **Task-specific rubrics** คือ การประเมินที่จำเพาะเจาะจงกับสมรรถนะใดสมรรถนะหนึ่ง โดยไม่สามารถใช้ซ้ำกับสมรรถนะอื่น ๆ ได้

## ประโยชน์ของ rubrics

1. ช่วยครูผู้สอนให้จัดจ้อยอยู่กับเป้าหมายการเรียนรู้ มิใช่มุ่งเน้นกิจกรรม ทำให้การประเมินมีความน่าเชื่อถือและเกิดความสอดคล้องของการประเมินรูปแบบต่าง ๆ เช่น การสะท้อนกลับ (feedback) การประเมินภาคปฏิบัติ หรือ การตัดเกรด
2. ช่วยนักเรียนให้เรียนรู้ผ่านการประเมินตนเองให้ก้าวหน้าตามระดับสมรรถนะจนประสบความสำเร็จ หรือนำมาใช้ในการเรียนรู้ร่วมกันระหว่างหมู่เพื่อนผ่านการสะท้อนกลับซึ่งกันและกัน

## ความเข้าใจผิดบางประการในการประยุกต์ใช้ rubrics

1. มุ่งเน้นกิจกรรมกว่าเป้าหมายการเรียนรู้
2. มุ่งเน้นองค์ประกอบ/ลักษณะ/คุณสมบัติของกิจกรรมมากกว่าเป้าหมายการเรียนรู้
3. ใช้การประเมินแบบ rating scale แทน rubrics

## ขั้นตอนการพัฒนา rubrics

ประกอบด้วย 4 ขั้นตอน ได้แก่

### 1. การกำหนดเกณฑ์การประเมิน

- สอดคล้องกับเป้าหมายการเรียนรู้
- ชัดเจนและมีความเข้าใจที่ตรงกันระหว่างผู้สอนและผู้เรียน
- สามารถสังเกตด้วยการดูหรือการฟัง
- แต่ละเกณฑ์มีความแตกต่างกันตามแง่มุม/ประเด็นของเป้าหมายการเรียนรู้ อย่างชัดเจน
- เกณฑ์ทั้งหมดเมื่อรวมกันสามารถสะท้อนเป้าหมายการเรียนรู้ทั้งหมดที่ต้องการ
- สามารถเขียนคำอธิบายสมรรถนะตามระดับความสามารถได้อย่างต่อเนื่อง

### 2. การกำหนดระดับของสมรรถนะ (แนะนำ 3-4 ระดับ)

- Good/normal/poor
- Good/basic/need improvement
- Good/standard/poor

- Best/better/average/poor
- Excellent/good/average/poor
- Advance/proficient/basic/novice
- Yes/Yes,but/No,but/No
- A/B/C/D

### 3. การกำหนดคำอธิบายสมรรถนะตามระดับ

ทำได้ 2 วิธี คือ

3.1. เริ่มต้นที่ระดับที่ดีที่สุด และ แย่ที่สุดก่อน จากนั้นจึงเขียนระดับที่อยู่ระหว่างกลาง

3.2. เริ่มต้นที่ระดับที่เป็นมาตรฐานหรือยอมรับได้ก่อน จึงเขียนระดับที่สูงและต่ำกว่า

- สามารถสังเกตได้
- เข้าใจง่าย ชัดเจน
- ครอบคลุมทุกระดับของสมรรถนะจากมาก/ดีที่สุด ไปหา น้อย/แย่มากที่สุด
- สามารถแยกแยะระดับของสมรรถนะออกจากกันได้
- กำหนดมาตรฐานที่ยอมรับได้ ไว้ในระดับที่เหมาะสม

### 4. การตรวจสอบคุณภาพของ rubrics

- Feedback จากครูผู้สอน
- Feedback จากผู้เรียน
- การวิเคราะห์จากผลการประเมินในอดีต (bottom up approach)

### แหล่งเรียนรู้เพิ่มเติม

1. Brookhart, Susan M. How to create and use rubrics for formative assessment and grading. 1<sup>st</sup> ed. Virginia: ASCD publication; 2013.
2. Strang, Aimee F. How to create meaningful rubrics for student assessment [Internet]. 2017 [cited 2018 April 11] available from <https://www.youtube.com/watch?v=arfXgcqJDs&t=2352s>
3. Truckee meadows community college (TMCC). Using rubrics for assessment [Internet]. 2015 [cited 2018 April 11] available from [https://www.youtube.com/watch?v=A7D\\_8uO5j-0](https://www.youtube.com/watch?v=A7D_8uO5j-0)

กระดาษบันทึก

กระดาษบันทึก

กระดาษบันทึก



# Question & Comments

ศูนย์ความเป็นเลิศด้านการศึกษาวิทยาศาสตร์สุขภาพ (ศศว)

*Siriraj Health science Education Excellence Center (SHEE)*

ฝ่ายการศึกษาก่อนปริญญา คณะแพทยศาสตร์ศิริราชพยาบาล

สำนักงาน: ตึก อุดมยเดชวิกรม ชั้น 6 ห้อง 656

โทรศัพท์ 02-419-9978/ 02-419-6637 โทรสาร 02-412-3901



[shee.si.mahidol.ac.th](http://shee.si.mahidol.ac.th)



[shee.mahidol@gmail.com](mailto:shee.mahidol@gmail.com)



[mahidol.shee](https://www.facebook.com/mahidol.shee)



SHEE FC

